Introduction

- The goal of this project was to develop a pipeline of multimodal data collection into trained models capable of determining whether or not an individual was telling the truth.
- The sensors used for collection were audio, video, thermal imaging, galvanic skin response, and eye tracking.
- The data was collected from 26 students and faculty over the course of two weeks and is composed of 7.3GB of audio, video and text formats.
- While the intent was to develop a system capable of aiding law enforcement during interrogations, the experimental conditions did not provide for a high stakes dynamic so the data collected should not be assumed to be transferable to that domain. \[^{[1]}\[^{[2]}\]

Methodology

- Participants were first connected to the galvanic skin response sensor to allow for the sensor to achieve equilibrium upon skin contact thus reducing noise.
- They were then centered in front of the thermal camera and calibrated to the SMI eye tracking machine.
- The first phase of the experiment was to answer 10 prerecorded questions and to lie to two of the subject’s choice.
- The next phase was to describe two images, one correctly and the other incorrectly.
- The subjects were then asked to provide a truthful story from their life and subsequently fabricate a story.
- The final phase was to watch two separate humorous video clips. One was a clip from the movie “Bruce Almighty” and the other a combination of clips from the television show “Who’s Line is It Anyway?”. The participant was asked to suppress emotion in one of the clips serving as another form of deceit.
- Between phases, participants were asked to read a neutral text passage from a 1953 weather report so as to reset any sort of physiological response from any previous lies.

Data Augmentation

- Due to the small sample size, additional data was needed in order to appropriately apply machine learning techniques and generate generalized models.
- Audio data was augmented by pitch shifting. A random number between -15 and 0 was added to the pitch of each audio clip.
- Due to the challenges associated with deriving a useful dataset galvanic skin response and eye tracking data was left out from the final classifier.
- Thus there were 5 modalities and 5 separate models that were trained and used to provide input to the final decision tree classifier.
- Data was split into 75 percent training and 25 percent testing and then augmented.

Analysis

- Audio data was collected in two channels. This allowed for two individual spectrograms to be created for use in a convolutional neural network. The x axis represents time, the y axis frequency and the color represents amplitude. All information from the recorded audio is present in this representation.
- Video data was converted into a single image. The process involved using a Cannny edge detector on each frame after being converted to grayscale. The higher threshold was set to 200 and the lower set to 25. The result was a binary image of important edge or not. The resulting images were then summed providing a hierarchy of important locations. The voting space was then normalized and set to a value between 0 and 255 creating a single channel grayscale image that was fed into a convolutional neural network.
- Three video features trained, thermal, full frame, and a video of the face based on a facial detection algorithm run on the full frame video.

Results

- Both a multilayer perceptron and a decision tree were evaluated for use as the final classifier. The decision tree was able to fit better to the data so it was chosen as the final component of the pipeline.