Using Affective Computing to Determine Subject Truthfulness

by

Bryan Passino

A Project Report Submitted

in
Partial Fulfillment of the
Requirements for the Degree of
Master of Science
in
Computer Science

Supervised by

Dr. Reynold Bailey and Dr. Ifeoma Nwogu

Department of Computer Science

B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York

May 2017
Abstract

Using Affective Computing to Determine Subject Truthfulness

Bryan Passino

Supervising Professor: Dr. Reynold Bailey and Dr. Ifeoma Nwogu

We live in a society where lying is all but accepted. It is not because it makes our lives any easier, rather it is so that we can abscond from facing the consequences of the truth. Although the prospect of removing lying from society seems like a scary concept, the long term outlook on such a prospect is a much greater and genuine mutual respect for one another. When the web of lies we all weave comes unraveled, then we can truly live with one another without masks.

More practically speaking however, a great tool for law enforcement would be the ability to discern truth from lie. This would improve efficiency both from a suspects point of view as well as law enforcement agencies. The polygraph test is the current tool used in this domain, however its accuracy is typically called into question and is not reliable enough to be admissible in the court of law.

This work harnessed machine learning techniques to find underlying patterns in deceit. The results are promising, but the setting of the experiments conducted does not match that of law enforcement interrogations. However the pipeline developed could be transferred so long as high stakes contextual data is available to train which is more relatable to those circumstances.
# Contents

Abstract ................................................................. ii

1 Introduction ......................................................... 1

2 Previous Works .................................................... 3

3 Background .......................................................... 5
   3.1 Video ............................................................. 5
   3.2 Face .............................................................. 6
   3.3 Audio ............................................................. 8
   3.4 Eye Tracking ...................................................... 8
   3.5 Galvanic Skin Response ........................................ 10
   3.6 Thermal ........................................................ 11

4 Data Collection .................................................... 13
   4.1 Phase One ....................................................... 14
   4.2 Phase Two ....................................................... 16
   4.3 Phase Three ..................................................... 16
   4.4 Phase Four ....................................................... 17
   4.5 Participant Demographics ..................................... 18

5 Analysis ............................................................. 20
   5.1 Convolutional Neural Network (CNN) ........................ 21
      5.1.1 Convolutional Layer ....................................... 21
      5.1.2 Pooling Layer ............................................... 22
      5.1.3 Rectified Linear Unit (ReLU) ............................... 22
      5.1.4 Fully Connected Layer ..................................... 23
   5.2 Audio Model ..................................................... 23
   5.3 Video Model ..................................................... 25
   5.4 Eye Tracking Data ............................................... 26
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5 GSR Data</td>
<td>27</td>
</tr>
<tr>
<td>5.6 Final Model</td>
<td>27</td>
</tr>
<tr>
<td>6 Results</td>
<td>29</td>
</tr>
<tr>
<td>7 Ethics</td>
<td>32</td>
</tr>
<tr>
<td>8 Future Work</td>
<td>34</td>
</tr>
<tr>
<td>9 Conclusions</td>
<td>36</td>
</tr>
<tr>
<td>Bibliography</td>
<td>37</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Results from the work of Mann et al. [8] .......................................... 4
3.1 Some of the Haar features selected though the work of Viola et al. [12] ... 7
3.2 Internal structure of a dynamic microphone. [4] ................................. 9
3.3 Color spectrum [6] ................................................................. 10
3.4 Equipped GSR device. ............................................................. 11
3.5 A monochrome image taken from a Flir Model A300 ....................... 12
4.1 Subject participating in study. ..................................................... 14
4.2 First image presented in Phase two. ............................................. 16
4.3 Second image presented in Phase two. ........................................... 17
4.4 Gender breakdown of study. ...................................................... 19
4.5 Age range ratio of study participants. ........................................... 19
5.1 Sample channel 1 audio spectrogram from the study. ....................... 24
5.2 Sample channel 2 audio spectrogram from the study. ....................... 25
5.3 Full classification flow of the model. ............................... 28
6.1 Visualization of the final decision tree where x[0] is audio channel 1, x[1]  is audio channel 2, x[2] thermal, x[3] face, and x[4] full frame. .............. 30
6.2 Results of each classifier. ...................................................... 31
Chapter 1

Introduction

Humans are taught not to lie. It is a major component of religion, and even in the absence of a nonsecular upbringing, it is taught in grade schools and homes around the world. Time and time again, ”lying is wrong” is proclaimed to children, and yet a brief glance into adult life illustrates that lying is both ubiquitous and pervasive in the world. Research into the statistics of lying has its own inherent problems, but Serota et al. found that the average rate of lying is 1.65 lies per day, although the results were not normally distributed [11]. It is sufficient to say, lying is not uncommon. Yet it is becoming more difficult to discern fact from fiction, and this becomes especially troubling for those charged with protecting others.

Law enforcement has the thankless job of trying to find criminals so that they can be held accountable for the laws they have chosen to disregard. Criminals, on the other hand, have a vested interest in keeping their nefarious deeds hidden from those who would prosecute them for their transgressions. That dynamic creates a push and pull of the truth in which lines are blurred. There is a need for a reliable beacon of truth and technology may provide a solution to this particular societal woe.

The ability of law enforcement to be able to distinguish a lie from the truth benefits all aspects of society, except perhaps those seeking self preservation. If law enforcement officers were able to interview a suspect and determine if they were guilty of some transgression or, conversely if they were not relevant to the matter at hand, then less time could be spent pursuing the wrong person and more time spent removing criminals from the streets.
The field of affective computing focuses on the research and development of systems capable of detecting and reacting to a user's emotional state. Its foundation is composed of numerous disciplines such as computer science, psychology, social science, biology, and neuroscience. In some respects, affective computing can be seen as an extension of sentiment analysis whereby sentiment is no longer inferred solely from text but instead from other forms of input. Such supplemental signals can range from video or audio recordings to other types of medical sensors such as electrocardiogram (EKG) or electroencephalogram (EEG) data.

Affective measurements allow for insight far beyond what is observable by the human eye and from those observations, a better understanding of a person on a more fundamental level. This form of computing lends itself well to the realm of deceit detection. The discovery of predictable patterns in autonomic body responses may remove the ability to deceive.

A similar detection device, the polygraph test, has long been doubted in its capability to provide definitive indications of veracity. The polygraph uses three indicators, heart rate and blood pressure, respirations, and skin conductivity. A test is conducted by presenting control questions to the examinee to establish a baseline prior to asking questions specific to the truth being sought. While evidence does suggest the polygraph tests can perform better than chance, though with high error rates of both false positives and false negatives [1], they are not widely believed to be reliable and as such are not admissible in the court of law. They are however used in non judicial settings such as employment screening and probation monitoring so despite the shortcomings, there is a need for this type of device.

This work attempts to improve upon the reliability of the polygraph test by incorporating machine learning techniques and supplementary sensor data. The sensors used in this study are video analysis, audio analysis, eye tracking, skin conductivity, and thermal imaging.
Chapter 2

Previous Works

The challenge with lie detection in an experimental setting is that there is typically no reward for successful deception and conversely there is no punishment for failing to get away with the lie. Subjects typically don’t even feel guilty for lying since they were asked to do so for the purposes of the study. This domain in which experiments reside are not capable of providing the appropriate setting for what they are even trying to study.

The work of Mann et al. [8] sought to address some of the deficiencies present given those limitations. Rather than conduct their own deceit study, Mann and her colleagues used prerecorded footage of criminal suspects being interrogated by police. The data was labeled by comparing answers given to other known facts about the case. This allowed for the observations of individuals with a vested interest in getting away with a lie as they were at risk of incarceration otherwise.

The results were broken into functions of truthfulness and deception. As seen in Figure 2.1 there are noticeable differences in actions when someone is lying versus telling the truth. The table does not reflect the individual changes that highlight the divergence in reactions as not all subject’s observations shifted in the same manner. For example, “with 56% showing more gaze aversion and 44% showing less gaze aversion while lying” [8] indicates that there is no common reaction that should be expected by all subjects.

Bhaskaran et al. also worked in realm of high stakes lying. Subjects were chosen who had a strong belief in a cause were chosen for the study. Their task was to chose to “steal” a check written to a group that supported an opposing view [2]. They were then interrogated by retired FBI investigators about the scenario and if caught the opposing
group would receive the funds from the check. If they were able to mislead the interrogator the funds were donated to a group supporting their cause and if they were able to mislead an interrogator into believing they stole the check when they did not actually do so, a supported group would receive the amount as well. In all, there was incentive to lie and get away with it, in an ethical manner while in a controllable experimental setting.

The relevant information to the study that was segmented from the video was eye movement and changes over the course of the interrogation. Based on the data collected and analyzed by the team, Bhaskaran et al. were able to develop a working model that was able to determine a response to a critical question asked of a subject after a baseline and rapport was built during the interrogation [2].

With the absence of a high stakes type of setting, the following work attempts to incorporate and, in some respects, enhance the findings of the previous works. It incorporates supplemental sensors which will provide data on many of the deceitful characteristics noted by Mann et al. It also includes the use of specifically captured and calibrated eye tracking data rather than gaze estimations found ex post facto as in the work of Bhaskaran et al.
Chapter 3

Background

3.1 Video

Colored digital images are composed of pixels which are a single point in an image. The image can be represented by many different color spaces, but for the purposes of this paper, the red, blue, green (RGB) color space will suffice. In the RGB color space, a pixel will contain three numbers ranging from 0 through 255, one for each of the channels, red, blue and green. It is through the combination of the RGB values and their relative strengths that a pixel color is determined.

A video is a series of still images viewed in succession giving the appearance of motion. A time series view of body positioning allows for the analysis of body language in a way which can give indications of dishonesty. For example, the work of Mann et al. found that engaging in deceit causes a subject to change the manner in which they use their hands and arms [8]. Some participants showed a decrease in movement, while others an increase. Despite the opposing changes, video analysis can show a shift from normalcy which, overall, can point towards a person’s truthfulness.

This experiment uses a web camera at a frame rate of 23 frames per second. It uses OpenCV to capture an image and record it to a video file at a resolution of 640 x 480 pixels.
3.2 Face

Facial expression analysis is a subset of the nonverbal communication that is subconsciously transmitted by humans. Micro-expressions in the face are known to hold a multitude of meanings. People are trained to notice and evaluate these expressions and there are even college departments that have a focus in the field such as the Psychology Department at UC Berkeley. In order to obtain an image specific to the face, it must either be done manually, which is preventively time consuming for large projects and contrary to the idea of affective computing, or it must be done automatically by some sort of classifier such as the one developed by Viola et al. [12]. Their work contributed several key methods for quickly and accurately identifying a face in an image and it is their introduction of using Haar like features that allowed for an algorithm fast enough to detect a face in real time. Figure 2.1 shows some of the first several features that were found to be useful.

The way that the classifier works is that it evaluates against these types of features that have been found to be present in the majority of faces and actively works to disprove the notion of detection if a feature is not present. If a feature is found, the algorithm checks the next feature, again seeking to reject rather than to accept that a face is there. Upon algorithm completion, there is no guarantee that there is a face present, rather that all of the known features of a face existed in the evaluated image. Each web camera frame capture is searched using a sliding window across the image as well as through variable scale space. Without a quick rejection facial detection algorithm, the process would require significantly more time than is available in real time applications.

According to Picard et al. neurological studies show that there are separate pathways that control facial expressions. A patient with a partially paralyzed face cannot will their face into a symmetrical smile, however when the smile is genuine and caused by a joke, a two sided smile is possible. Likewise, facial cues of deceit may hide themselves from conscious control and through facial analysis, hints of those facial distortions may become evident and tells of deceit.

Ekman further classifies false expressions into subcategories such as referential and
mock expressions each indicating different subject affect[3]. He states that true emotion requires more muscular control than most people can perform on their own accord. Moreover, he proclaims that if someone is capable of performing that correct muscular contractions the subject may actually invoke the true emotion.

The face is segmented from the full frame video by use of OpenCV’s Haar cascade classifier. Despite the high accuracy of the classifier to determine if there is a face present in the image there are still frames that are dropped. As such, there may be less frames in post analysis. Not all faces found by the classifier are a consistent size so the face image is written to a video file at a standardized resolution of 100 x 100 pixels.
3.3 Audio

When humans speak, there is meaning in the words used, yet there is still more information conveyed in the delivery of those words. Certain inflections can modify the meaning of the same phrase or a pause can dictate the difference between anger and frustration. It is for that reason, that audio analysis can be a crucial input to affective computing. Microphones provide the medium by which computers can hear and they work by converting sound waves into an electrical current. This is accomplished by the vibration of a very thin plastic diaphragm which affects an attached coil. When the coil moves, it does so through a magnetic field which generates an electrical current as seen in Figure 3.2. The current can then be stored for playback and analysis. Researchers have found that qualities of audio analysis such as pitch, intensity, speaking rate and voice quality play vital roles in the recognition of emotions [9].

Mann et al. found that there are two notable changes in subject speech patterns when engaging in deceit. The first is a pause before responding. The second is the use of speech fillers such as the use of the word “um” or “uh”. Both of these patterns can be considered a delaying tactic as an individual is fabricating a lie.

3.4 Eye Tracking

More detailed than the image of the eyes that can be observed in video analysis, eye tracking technology focuses solely on the image of the eye and the direction of focus. An eye tracker typically uses a camera designed to view images in the infrared light spectrum, which as you can see in Figure 3.3 is outside of the visible portion of what humans can see. This allows infrared lights to be shined directly onto an eye without the light causing changes in pupil dilation affecting measurements. Illumination is done in patterns much like structured lighting and deviations from the expected patterns are used to determine the user’s eye movements. This sensor requires user interaction to calibrate. The sensor follows the user’s eye movements in relation to choreographed patterns on the screen to
Figure 3.2: Internal structure of a dynamic microphone. [4]

adjust parameters for accurate tracking. The tracker then updates gaze points based upon the pattern of reflected lighting changes and numerous eye modeling algorithms.

Due to the structured nature of the lighting, the eye tracker must be positioned statically in relation to the eye. This can be accomplished by either a sensor attached to a computer which requires little movement from the user or the use of specially designed glasses that a user can wear which looks both outward and inwards towards the eye. The second option allows for more mobility on the part of the user without sacrificing measurements.

The eye tracker used in these experiments is manufactured by SensoMotoric Instruments (SMI) and is run on a separate computer system. Calibration for the device is done through SMI proprietary software titled Experiment Center using 9 points of measurement. The tracker provides the following information

- time stamp
- pupil diameter
Mann et al. found that one of the most reliable indications of a lie is a decrease in the blinking rate of the subject. The tracker doesn’t explicitly provide that information, however, if the other data points have null values, that is an indication of a blink. Thus a blinking rate can be calculated.

### 3.5 Galvanic Skin Response

Galvanic Skin Response (GSR) is another name for electrodermal activity, or the variation of electrical conductance of the skin. The primary cause of the variation is the productivity of the sweat glands in the skin. If a person experiences stress, then the autonomic nervous system becomes aroused causing an increase in sweat production. The increased production causes the electrical resistance to decrease thus increasing conductivity. This metric allows for the determination of stress of an individual. However, this is an intrusive sensor as can be seen from Figure 3.4. More importantly, the location of the sensor on a user’s hand which is likely to be used for other tasks while computing introduces the possibility...
for noisy data as movement can affect GSR readings.

As mentioned previously, this sensor is one of the three used in a polygraph test and has a long standing record of being useful in determining veracity. The GSR device in this experiment is the Shimmer 3 produced by ConsenSys shown in Figure 3.4. The device is set to a sampling rate of 51.2 hertz and a GSR range of 10 kilo Ohm to 56 kilo Ohm.

Figure 3.4: Equipped GSR device.

3.6 Thermal

Thermal imaging, similar to eye trackers, work in the infrared light spectrum. The sensor inside the camera is able to detect subtle changes in temperature by sensing thermal radiation. From the sensor data, a thermogram is constructed which allows for visual representations of the scene. The representation is monochrome with heat on one end of the range and cold the other, depending on the camera model. A sample image taken from the device used in these experiments, a Flir model A300 can be seen in Figure 3.5. The color bar on the right shows the range of temperatures that the camera is set to detect.
The temperature range 80 degrees Fahrenheit to 98 degrees Fahrenheit was selected due to the value contrast that is relevant to thermal changes in the face relative to the recording environment.

Figure 3.5: A monochrome image taken from a Flir Model A300
Chapter 4

Data Collection

Data was collected in various phases. Each phase presented a different opportunity to collect varied aspects of deceit. Truthfulness is not restricted to a verbal format, and diverse representations provide a more generalized concept for machine learning techniques to detect. The sensors used were described earlier, video, audio, GSR, eye tracking, and thermal imaging.

Prior to the data recording sessions, subjects were connected to the GSR in which there was a sensor placed on the left forefinger and the left middle finger. Due to the sensitivity of the device and its propensity to collect noise along with the real data, subjects were asked to keep that hand still and resting on the table.

The subject was then calibrated to the eye tracking device using the SMI proprietary software for the device. Verbal instructions were given about the first phase and further instructions were given prior to the start of all subsequent phases of the experiment so that they knew what to expect, however there were also written instructions presented on screen between each phase. By the time the subject began the experiment, the GSR sensor was attached for a minimum of five minutes allowing for the sensors to acclimate to their position on the skin and for the associated noise from the initial application of the device to dissipate.

Between each phase, participants read a neutral text passage out loud while being recorded. This provided an opportunity for any residual physiological reactions from any previous deceit to lessen so that further phases would not be corrupted by carry over from the previous. The neutral texts were exerts from a 1953 weather report and as such were
not eloquently written. Participants were told prior to these phases to keep reading even if they stumble over word pronunciation. This was especially important for the non-native English speakers who struggled with some of the vocabulary as any embarrassment might cause unwanted emotions introducing noise into the data in the following phase.

To ensure that any immediate and subsequent reactions prompted by a particular portion of the experiment were credited to the correct portion, there was a four second buffer of recording time between each progression throughout the experiment. Each of those progressions were then saved to separate files which kept the data segmented as it was collected. In each phase, all mentioned sensors recorded from the beginning of the phase unless otherwise specified below.

![Subject participating in study.](image)

**Figure 4.1: Subject participating in study.**

### 4.1 Phase One

The first phase consisted of a question and answering session composed of ten questions. The participants were provided the questions prior to the data recording session and were
instructed to choose two questions to answer dishonestly. The experiment facilitator was not informed of which questions were answered with a lie until after the experiment was completed.

1. What is your name?

2. How old are you?

3. Where are you from?

4. What is your major?

5. What is your father’s name?

6. What is your mother’s name?

7. What did you do over the winter break?

8. What is the name of best friend?

9. How long since your last haircut?

10. What is your favorite show?

The questions were prerecorded and played back during the experiment. The screen of the computer consisted of a pseudo cross hair to serve as a focal point for the eyes of the participant which they were asked to focus on. As the questions were being asked, all sensors except for the audio were recording the reaction of the participants and upon completion of the question playback, the audio began recording. The experiment facilitator manually instructed the experiment to progress to the next question so that it was ensured that the participant fully answered prior to moving on.
4.2 Phase Two

The second phase consisted of describing two images displayed individually on screen, one truthfully and the other incorrectly. Prior to the images being presented, each participant was instructed which picture to lie about and which to describe truthfully. The first image shown to the participants was Figure 4.2 and the second was 4.3. Even numbered participants, zero based, truthfully described 4.2 and odd numbered participants truthfully described 4.3 however the order of picture presentation remained constant. The images were chosen because they provide a significant amount of content to describe for the truthful portion of the phase and an equal amount of content to actively ignore at the deceitful portion. There was no time limit nor minimum for this phase.

![Figure 4.2: First image presented in Phase two.](image)

4.3 Phase Three

In this phase, the subject was asked to first describe a true story from their life. If requested, time was given to the participant to figure out what aspect of their life to describe. Once that was completed, they were then asked to fabricate a story. In an effort to simulate on the
spot lying, participants were given little to no time to come up with a lie. As in the previous phase, these two portions did not have a time limit nor minimum when the response was being recorded.

### 4.4 Phase Four

The next phase was measuring responses from participants as they watched a humorous video clip. The video clips chosen were a scene from “Bruce Almighty” and a scene from the television show “Who’s Line Is It Anyways”. Since not everyone in the experiment was familiar to the movie or the show, a brief synopsis was given to provide context prior to viewing the clip. One of the clips subjects were asked to watch and react naturally, as in laugh if they found the clip funny, don’t laugh if they don’t. The other clip they were instructed to suppress any sort of reaction, so keep a straight face even if they found the scene to be funny. Participants were split evenly so that half suppressed reaction on one clip and the other half on the other. The premise behind this phase was to isolate a non verbal form of deceit providing more robust data.
4.5 Participant Demographics

Twenty six subjects participated in the study. When the experiment concluded, each was asked to fill out a brief post survey questionnaire. The following questions were asked and all but the final question were multiple choice. The last question was open ended.

- Gender?
- Please select an age range
- Which questions did you lie on?
- Were you familiar with the movie ”Bruce Almighty” prior to today?
- Did you find the ”Bruce Almighty” clip funny?
- Were you familiar with the show ”Who’s Line is it Anyway” prior to today?
- Did you find the ”Who’s Line is it Anyway” clip funny?
- Please describe any strategies used when trying to suppress your reaction in the video clip.

The goal of the gender ratio was to keep it even so that there was little bias in the resulting pushed towards one gender or another. The final breakdown can be seen in Figure 4.4.

As most of the participants were students, the age range was skewed towards the younger end of the spectrum. This was not ideal, but it was the largest percentage of the university population who were available to participate. Figure 4.5 shows the complete ratio of participant age ranges.

There was a concern that not all participants would find the video clips funny due to lack of context and cultural differences. Less than 10% of participants in either video stated that they did not find the clip funny. An argument could be made that if a participant who did not find the clip to which they were supposed to suppress their reaction funny that their
data would not be valid since there was no actual suppression. This study operated under the assumption that the forethought of suppression would invoke a similar physiological response so no data was removed from the study for that reason.
Chapter 5

Analysis

The data collected from the participants consisted of 7,207 items for a total of 7.3 GB once all associated files were generated. Needless to say that not all of the data collected was good data especially considering that it was being collected from sensors that are known to be sensitive to noise.

Each modality was examined and considered individually which coincided with an effort to build parallel classifiers of each feature space which fed into a final classification taking the outputs of each individual model and making a final determination of whether or not an individual was telling a lie or not. Therefore all models generated were binary classifiers with 0 representing the truth and 1 indicating a lie.

Given the 26 participants in the study, 10 questions asked in Phase 1 and 2 samples collected from each of the subsequent phases, that totaled 16 instances to classify per subject and 416 total. Of the 416 total, 31.3% were lies and the remainder truth. This is a relatively small sample size with a disproportionate amount of negative instances given the requirements of the machine learning techniques tasked to model the data. In order to compensate for the size and ratio of the dataset, data augmentation techniques were used in the development of the training and testing data. However, training and testing data was split prior to augmentation so no derivation of testing data was present in the training set. The ratio was 75% training to 25% testing for all modalities including the final model.
5.1 Convolutional Neural Network (CNN)

A CNN is a feed-forward type of neural network that is modeled on the visual cortex of animals. It is similar in some ways to a multilayer perceptron model, however because that type of model is based upon multiple fully connected layers, it does not scale well as resolution increases. Conversely, a CNN typically only uses fully connected layers at the end of the model which allows for less computation. This, coupled with the decreasing size of each layer the deeper traversed into the CNN, allows for shorter training times than other types of neural networks. The order of traversal in the network typically follows the pattern below.

1. Convolutional Layer
2. Pooling Layer
3. Rectified Linear Unit Layer
4. Fully Connected Layer

Deep learning occurs when multiple levels of these layers are stacked together. The types of layers don’t typically change aside from the associated parameters, but the architecture may vary depending on the particular problem being addressed. A CNN is trained by a method called backpropagation which calculates the loss gradient and by use of the chain rule updates weights in each of the previous layers appropriately.

5.1.1 Convolutional Layer

As its namesake implies, it is almost always the first layer in a CNN. In this layer the input image is manipulated by a filter of n x n size, where n is an adjustable parameter. The filter or kernel is the weight of the layer that is adjusted through the training process by way of back propagation. The filter begins as an arbitrarily designated matrix of values but essentially becomes a feature detector. Training the CNN refines the weights into said
detector, but the end state of the detector is unknown prior to the completion of training. Depending on the size of the filter, the resulting convolved image is \( n - 1 \times n - 1 \) pixels smaller in dimension. The result is known as a feature map and is fed to the next layer of the network.

### 5.1.2 Pooling Layer

It is common to insert a pooling layer after a convolutional layer as it provides a means of sub-sampling otherwise down as down-sampling the results of the previous layer. This injects non-linearity and also allows for variation in the feature found by the filter in a given layer. The process works by sliding a window over the feature map of \( m \times m \) size with a stride of \( x \). The maximum value found within the window is taken and placed respectively in the output. The window is then slid across the feature map by \( x \) and the maximum is taken again. This process is repeated until the entire map has been pooled. Some common variation, aside from \( m \) and \( x \) are whether, and by how much, the pooling overlaps and the choice between max pooling versus average pooling.

### 5.1.3 Rectified Linear Unit (ReLU)

The ReLU is an activation function defined as

\[
f(x) = \max(0, x)
\]

The function works by sliding a \( n \times n \) window along the feature map, the incoming layer, and finding the maximum value of that window with a hard lower limit of zero. This type of activation function has become the most popular for use in deep neural networks as of 2015. As[7] states, this is likely due to the vanishing gradient problem seen with other activation functions like sigmoid and tanh which have range limits of \((0, 1)\) and \((-1, 1)\) respectively. As backpropagation makes its way towards the original input layer the loss gradient shrinks exponentially with each layer thus causing the changes in the front of the network to be much smaller and take longer to train. Since ReLU does not have an upper
limit, the shrinking isn’t as drastic and the front of the network can be developed much quicker. In general ReLU layers are relatively easier to optimize, converge quicker and are faster to compute than their counterparts in deep learning. [13]

It is of note, that ReLU can suffer from the very thing that causes them to excel in what is known as the knockout problem. During training, a ReLU unit can update a weight in such a way that the neuron is never activated again due to the the \((0, \infty)\) range. If that scenario occurs, then that filter essentially becomes an unused aspect of the model. It may be found that up to 40% of a network can be "dead" [5]. Fortunately, this phenomenon occurs predominately when the learning rate is set too high, and with a proper setting, the amount of neurons that are affected are kept to a minimum.

### 5.1.4 Fully Connected Layer

The final layer acts as the voting layer in which the output label is determined based upon some combination of activated values that has been learned during the model training. This layer has connections to all values in the previous layer, hence its name and may at times be connected to another fully connected layer, though, as mentioned earlier, this type of layer is computationally expensive. In a fully trained model each potential label evaluates its probability with the most likely option being selected as the output. Since the output is based on probabilities, a threshold parameter can be added so that a label is not selected unless the likelihood is above a certain value as a form of a degree of confidence.

### 5.2 Audio Model

The audio data was collected at a rate of 44,100 samples per second in two channels. This provided enough information in both channels to build two distinct models, one for each channel. The first step in the audio pipeline was to convert the audio samplings into a spectrogram image, examples of which can be seen in Figures 5.1 and 5.2. At a glance, the two channels seem almost identical which is to be expected, however there are variations and since each channel is trained in its own model, the similarities do not affect the efficacy
of the classification.

A spectrogram is a visual representation of sound in which the x axis is time, the y axis frequency, and the z axis amplitude. All of the data associated with the audio recording is present in a spectrogram so no data loss occurred during the conversion. The spectrograms provided a standardized image format, which was resized down to 32 x 32 pixels, despite the difference in length of each audio clip.

It was decided that Phase 4 from the study would not be included in training the audio models. The sound of the video clip was present in each video and dominated any recorded audio so they would not provide any benefit to classification. This left 364 instances of subject recordings for audio analysis, 28.65% of which were lies and the remainder truth.

![Figure 5.1: Sample channel 1 audio spectrogram from the study.](image)

Of the potential audio augmentation techniques, Salmon et al. found that pitch shifting had the greatest positive impact on performance without negative impact [10]. Accordingly, each sample was augmented using pitch shifting and then stretched or shortened back to the original length of the audio clip. The total number of instances were brought up to 5200 and the ratio of truth to lies was made even at 50/50.

The data was then trained using deep learning in a CNN passing through 4 convolutional layers. The images were sent through the CNN in 3 color channels and trained for 1000
epochs.

5.3 Video Model

Three video modes were collected during the experiment, full frame from a webcam, face frames based on a facial algorithm run on the full frame, and a thermal image camera. Each form of video was preprocessed the same way and each model was built in the same fashion as well. The only caveat is the exclusion of a single thermal video recording as the file was corrupted and could not be read.

The first attempt to develop a video classifier involved a convolutional long short term memory recurrent neural network. This method required videos of the same length otherwise videos of less length would need to be zero padded to accommodate the missing frames. In addition to zero padding, videos needed to be trained one at a time due to memory limitations of the system being used in relation to the size of all videos combined into a single input. For these restrictions and the amount of time training would have taken, a different approach was taken.

The approach chosen was to convert the video sequences into a single image which could be trained using a CNN. In order to accomplish this, important features from each
frame would need to be included in the final image. Thus the method of video to image conversion was as follows.

1. Convert frame to gray scale image

2. Apply Canny edge detection on frame using a lower threshold of 25 and a high threshold of 200

3. Sum the values of each frame onto a new image creating a voting space

4. Normalize the values of the new image and multiply by 255 creating a final gray scale image

As with the audio analysis, augmentation was used to provide more data to train the model. The augmentation method was image rotation. The image was rotated around the center pixel \( n \) times from negative \( n/2 \) degrees towards the first quadrant with a 1 degree step. The total number of instances were brought up to 5720 and the new ratio of truth to lies was 50/50.

### 5.4 Eye Tracking Data

The eye tracking data experienced several flaws that were noticed during the data collection portion of the study. The SMI eye tracking device used was not able to calibrate to a high degree on all individuals. Even once calibration was completed, the device would lose track of the eyes and many samplings would result in zero because of that. This was especially true of those who wore glasses as the device relies on reading the reflection from the eye as a part of its measurement.

Another issue with the eye tracking data was that participants should have kept their head relatively still once calibrated. However, it was noted that upon finishing a question, participants would frequently look to the study facilitator which could cause calibration to be lost for the following portions of the experiment. This type of error would be difficult
to clean and for use in these results and would require multiple assumptions. For these reasons an eye tracking model was not used.

5.5 GSR Data

Upon completion of the study and further analysis of the GSR data, it was found that many recordings were erroneous. The Shimmer 3 device would record a stream of values for a single recording that were out of range of expected values not only for typical GSR readings but for the individual being assessed, as compared to other readings from different portions of the experiment.

Potential reasons for the faulty readings are shortcomings in the device and participant fidgeting. The device was cycled through on and off states repeatedly during this experiment and it was possible that this would sometimes cause the device to fail upon reset and calibration prior to starting the next recording. While participants were instructed to leave the hand to which the device was connected still during the experiment, it was observed that many participants would moved that hand while answering questions.

After removing the bad data from the training and testing set, efforts were made to train a long short term memory recurrent neural network on the remaining data. Augmentation was performed by making the assumption that a signal was universal across the recording and thereby slicing the signal into two pieces and taking the first portion and concatenating that signal onto the end of the second portion.

This model was not able to perform better than random and was therefore not included in the final classifier.

5.6 Final Model

The final classifier was evaluated using a decision tree and a neural network. Based on preliminary testing, a decision tree was better able to fit to the data and was chosen as the method for final classification. To elaborate, a decision tree is a tree like structure that is
trained by using gini impurity calculation defined as

\[ gini\_impurity(f) = \sum_{i \neq k} f_i f_k \]

to split the data set into subclasses where final classification is possible.

The input to the final classifier was the outputs of the individual models which was the probability between 0 and 1 of an instance being a lie. A visual representation of the 2 tier model can be seen in Figure 5.3. The output of the final classifier then is the cumulative confidence of a lie taking into account all modalities.

Figure 5.3: Full classification flow of the model.

The data for the decision tree was pure. That is to say, there was no augmentation performed for training. As with the audio analysis, Phase 4 was not included due to fact that the audio results would be determined by the video clip being played in the background. This left 364 instances to classify, again broken into 75% training and 25% testing.
Chapter 6

Results

The metrics calculated to evaluate the models were accuracy, recall, precision, specificity, false positive, and F1 score. Accuracy is the probability of the classification being correct. Recall is the proportion of instances that were lies and were labeled by the classifier as such. Precision is the proportion of instances that were classified as lies that actually were. Specificity is the number of classifications of truth when they were telling the truth. False positive is the number of times when instances were classified as lies when they were truth. F1 score is a measure of a test’s accuracy that considers both recall and precision in its calculation which is evaluated as

\[ F1 = 2 \times \frac{precision \times recall}{precision + recall} \]

The decision tree generated can be seen in Figure 6.1. It is of note that through multiple trainings this was not the only tree developed, however it did produce the best results. There were conditions in which another variation of the tree was derived and attributes used for decision higher in the tree were used later, this resulted in much less accurate results. This happened at the point in the tree seen in Figure 6.1 that the thermal feature is evaluated against. If audio channel 1 or audio channel 2 were used again instead, the resulting tree would perform much worse on the testing set.

The above mentioned metrics can be seen for each classifier in Figure 6.2. While no individual classifier performed well on its own, including multiple modalities increased prediction accuracy. Overall the final classifier was able to achieve 92.3% accuracy.
Figure 6.1: Visualization of the final decision tree where $x[0]$ is audio channel 1, $x[1]$ is audio channel 2, $x[2]$ thermal, $x[3]$ face, and $x[4]$ full frame.
Figure 6.2: Results of each classifier.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Recall</th>
<th>Precision</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Audio Channel 1</td>
<td>0.47</td>
<td>0.69</td>
<td>62.9%</td>
</tr>
<tr>
<td>n = 1300</td>
<td>514</td>
<td>130</td>
<td></td>
</tr>
<tr>
<td>Actual: Truth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual: Lie</td>
<td>346</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>Face Isolation</td>
<td>0.52</td>
<td>0.68</td>
<td>63.9%</td>
</tr>
<tr>
<td>n = 1446</td>
<td>556</td>
<td>164</td>
<td></td>
</tr>
<tr>
<td>Actual: Truth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual: Lie</td>
<td>503</td>
<td>223</td>
<td></td>
</tr>
<tr>
<td>Full frame</td>
<td>0.43</td>
<td>0.71</td>
<td>62.5%</td>
</tr>
<tr>
<td>n = 1446</td>
<td>503</td>
<td>127</td>
<td></td>
</tr>
<tr>
<td>Actual: Truth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual: Lie</td>
<td>415</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>Final Classifier - Decision Tree</td>
<td>0.73</td>
<td>1.00</td>
<td>92.3%</td>
</tr>
<tr>
<td>n = 91</td>
<td>65</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Actual: Truth</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actual: Lie</td>
<td>7</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 7

Ethics

Some might argue that the removal of the ability to lie is an invasion of privacy. Perhaps it may even be considered another form of surveillance in which it is not your phone lines being tapped or your web browsing history monitored but rather a glimpse into your very thoughts. It could further be argued that telling the truth is a personal decision that one must make on their own and that another individual should not be able to take that decision away.

In reality, we all make judgments about truthfulness everyday on less scientific determinations. Consider the dynamic of a relationship. A man might be late coming home and his wife asks where he has been. That man might say that he had to work late to meet a deadline when in reality he stopped at a bar on his way home to grab a drink. The wife then has to choose whether or not he is being truthful or lying to her. She uses past actions and the current situation to make a determination if what he is saying is true. Based on the way that she reads the situation she reacts by either moving on with their night or fighting with him for lying. Perhaps she believes him and turns a blind eye to a drinking problem or decided to divorce him after acknowledging the lie and choosing not to stay in the marriage due to a continuation of alcohol abuse.

Lying is so well integrated into our society that people are expected to engage in the act to spare someone’s feelings and are considered rude for failing to do so. A common scenario is the accepted response to someone asking “Do I look fat in this?” is of course not. To say otherwise might hurt their feelings. A honest person however, might break
convention if they believe the answer to be yes. This provides the person asking the oppor-
tunity to make adjustments to their outfit and, practically speaking, is the correct thing to
do. Political correctness however would disagree and that, it seems, is the direction society
has chosen to move.

We as individuals are continuously faced with situations where veracity affects our life.
It only stands to reason that better decisions can be made if total truth is known. Mak-
ing decisions on falsehoods are irresponsible at best. Astronauts would not lie to mission
control, nor should mission control have to contemplate the possibility. Unfortunately,
interpersonal relationships are now plagued with frequent lies in an effort to avoid respon-
sibility for one’s actions. This is especially evident when people are interrogated by law
enforcement personnel.

There would be a drastic improvement in law enforcement efficiency and justice in
general if there was a reliable way to determine the truth of suspects. Not only would it
serve as a major deterrent as would be criminals would be aware that they could not simply
lie their way out of punishment, but it would also allow for the prosecution of guilty parties
and the vindication of the innocent.
Chapter 8

Future Work

Individual classifiers performed better than average, however there is room for improvement. If the legs on which the final classification stands improve, it logically follows that final results will amplify that enhancement.

Audio analysis may be improved by adjusting resize dimensions of the spectrogram. 32 x 32 was the only size tested, other sizes may yield better results. Larger sizes will retain more of the original information from the full sized spectrogram, the more size reduction, the less information is available for the CNN to train upon.

The video to a single image process works, however there is certainly information lost in the conversion that may be relevant to increasing the accuracy of video classification especially with the thermal modality as it barely scored higher than average in accuracy. Implementation of a variable length sequence convolutional long short term memory recurrent neural network might provide the best results so future work would definitely include work either within the Keras/Theano framework or development of a supplemental library capable of the necessary architecture.

Other modalities should be included in order to provide more accurate results. Based on the issues with the eye tracking device. Gaze estimation might be a substitute which wouldn’t necessarily rely on calibration beforehand. If multiple cameras are used, it could help account for relative position of the eyes despite head movements. GSR, a component of the polygraph, was not able to provide any useful input to the final classifier in the context of this study. It was expected that this signal would have the highest correlation to deceit. Whether through faulty equipment or sensors that did not maintain contact with the
skin throughout the entirety of the experiment, the GSR protocol should be adjusted so that it has the opportunity to provide valid input to the overall process.
Chapter 9

Conclusions

The described parallel classification pipeline performs quite well on the limited sample size of data collected. Individual classifiers were able to increase their accuracy to above random in respects to their own modality. The final tier classifier achieved an accuracy rate of 92.3% without any false positives and only 7 failed lie classifications. That is not to say however that this model could be applied universally. There is still much further testing and additional modalities to consider, not to mention improvements on those already included in the outlined process. It does seem to be a starting point however to develop a system that could out perform the industry standard, the polygraph. If a system could prove to be robust and reliable, perhaps the predictions made by said device may become admissible in court.
Bibliography


