# INTERNSHIP REPORT - DECEPTION DETECTION THROUGH VIDEO ANALYSIS

**Siddhartha Rao Kamalakara** [*]
Department of Computer Science
Manipal Institute of Technology
Manipal, India 576104
srk97c@gmail.com

June 12, 2019

## ABSTRACT

Deception Detection is a task that has high significance in real life applications ranging from law enforcement to social media texts. Automating deception detection is a challenging task since human behavior is very diverse and subjective. Traditional Machine Learning approaches are neither scalable nor powerful enough to capture the input distribution. Deep Learning approaches have been shown to scale very well provided that there is a lot of data to train on. Unfortunately, acquiring labelled data for deception is very hard due to the nature of the task. Most attempts at deceit in real life go unnoticed or they are acquired in confidentiality. Video data contains many sub-modalities that are useful for deception detection: Micro expressions, pose estimation, Gestures etc. Recent research has revealed that it's possible to acquire physiological signals from videos by detecting subtle variations in the input. This work first provides an overview of the work done in video analysis so far and then, explores the task of extracting pulse rates from raw videos through a method called remote photoplethysmography.

***K*eywords** Deception Detection · rPPG

## 1 Introduction

Deception is common in our daily lives. Some lies are harmless, while others may have severe consequences and can become an existential threat to society. For example, lying in a court may affect justice and let a guilty defendant go free.Therefore, accurate detection of a deception in a high stakes situation is crucial for personal and public safety. The ability of humans to detect deception is very limited. In (Bond Jr and DePaulo 2006), it was reported that the average accuracy of detecting lies without special aids is 54%,which is only slightly better than chance. Automated Deception Detection has been the focus of recent research due to significant advances in statistical analysis techniques over the past few years. Detecting deception in humans is a difficult task because of the complexity in underlying factors that characterize it. Behavioral changes vary across subjects across multiple modalities. Physiological factors like Heart Rate, Respiration Rate also play a major role in detecting deception . Capturing and understanding these changes at scale requires powerful models. According to Vrij et al (2000), there are three approaches to lie detection: non-verbal (observing an individual's gaze, movements etc), verbal (analyzing the speech content), and physiological (heart rate, muscle activity etc).

Recent advances in video analysis techniques make it possible to extract physiological factors from raw videos. This makes videos a great source for detecting deception due to the fact that they combine non-verbal cues and physiological factors into one modality. Disentangling each of these factors from video and building predictive models on top of these features is a promising research direction for deception detection. This report details the work done during the 4 months internship period at MICL, NTU.

[*] Work done during Internship at NTU, Singapore

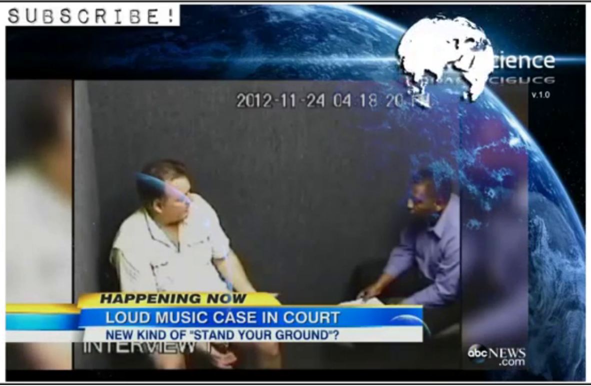


Figure 1: Frame from a selected video that is unfit for analysis

## 2 Related Work

Availability of videos for deception detection with ground truth is very limited because of the nature of the data. Existing approaches have relied on datasets of limited size and simple statistical models for prediction. This means that the existing approaches are not scalable due to the severe lack of diversity in the available data and the limited expressive power of simple features and models.

### 2.1 Dataset: Real-Life Trial dataset

[Pérez-Rosas(2015)] created a dataset consisting of real-life trial videos that are publicly available on YouTube channels and other public websites. The dataset also contains statements made by exonerees after exoneration and a few statements from defendants during crime-related TV episodes. The speakers in the videos are either defendants or witnesses. The video clips are labeled as deceptive or truthful based on a guilty verdict, not-guilty verdict and exoneration. The final dataset consists of 121 trial videos, of which 61 are deceptive and 60 are truthful. There are 21 unique female and 35 unique male speakers. Each clip in the dataset is annotated as either Innocent or Guilty. Each video also has annotations for a limited number of gestures. Although there are 121 videos, the real life trial dataset is far from ideal for a video analysis. Many of the videos include the same subject or witness which presents a problem when deploying end-to-end models for learning. The dataset also contains videos where the subject does not appear or is out of focus. Existing video-based approaches have removed these "outlier" videos thereby further reducing the size of the dataset. An example of such a video is shown in Figure 1.

### 2.2 Previous work on Real Life Trial data

[Zhe Wu(2017)] use Improved Dense Trajectories (IDT) to extract features from consecutive frames and eliminate camera motion using the RANSAC algorithm. The feature points are then densely sampled at multiple spatial scales and tracked through a span of multiple frames such that drifting does not occur. [Zhe Wu(2017)] in addition to IDT, manually annotated the dataset with micro-expressions and computed high level features for detecting these micro-expressions. These feature vectors from IDT and micro-expression detectors along with features from Audio and Text are encoded into a fixed length vector through Fisher encoding which uses a k-component Gaussian Mixture Model to compute the encodings.

$$G_{\mu_i} = \frac{1}{T\sqrt{w_i}} \sum_{t=1}^{T} \gamma_t(i) \frac{(x_t - \mu_i)}{\sigma_i} \tag{1}$$

$$G_{\sigma_i} = \frac{1}{T\sqrt{2w_i}} \sum_{t=1}^{T} \gamma_t(i) (\frac{(x_t - \mu_i)^2}{\sigma_i^2} - 1) \tag{2}$$

where $\gamma_t(i)$ is the posterior probability. $\{x_1, x_2, ..., x_T\}$ is the bag of features.

[Zhe Wu(2017)] used various predictive models like LinearSVM, KernelSVM, Random Forests etc to evaluate the performance on the task and found that LinearSVM performs best with IDT and Micro-Expression detectors.

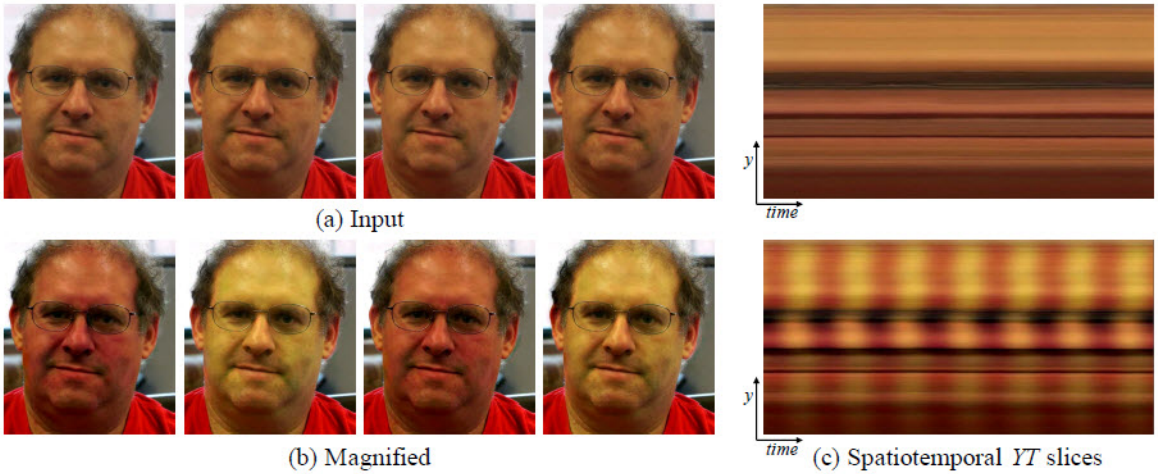


Figure 2: Eulerian Video Magnification used to magnify color changes [Wu et al.(2012)Wu, Rubinstein, Shih, Guttag, Durand, and Freeman]

[Krishnamurthy et al.(2018)Krishnamurthy, Majumder, Poria, and Cambria] take a deep learning approach to detecting deception on the real-life trial dataset. Like Wu et al, they use Video, Transcriptions, Audio and Micro-Expression features computed from an end-to-end approach to classify deception. The video features are computed from a 3D convolutional neural network which leads to a 300 dimensional vector which is then combined with the feature vectors extracted from other modalities through a hadamard product. The concatenated feature vector is connected to a Dense layer with 1024 output units which is then used as the final feature vector to classify deception. The network is trained end-to-end with Stochastic Gradient Descent by minimizing the binary cross-entropy. The paper claims to have achieved 94% accuracy using only videos. A potential problem with an end-to-end approach to learning deception from the trial dataset is that there are multiple videos of the same subject in the dataset which could influence what the network learns. The authors use K-fold cross validation to prevent the presence of the same subject in training as well as the testing set. Another concern is that the network used by Krishnamurthy et al is prone to overfitting due to the massive number of parameters and very limited amount of data.

Almost all the other approaches that analyzed the Real Life Trial dataset have either not used the video or have relied on features computed from statistical techniques. This does not help the general case of deception detection because of the limited modeling capability of traditional statistical techniques. For an algorithm that can generalize well, lot of diverse data is required to capture the variance across human behavioral traits. This means that we also need models that are powerful enough to understand this pattern in the data. Deep learning is indeed a promising direction for solving deception detection but suffers the drawback of less data and interpretability which is important if we intend to deploy these systems in the real world. Existing work so far has not leveraged the physiological signal present in videos to detect deception. One reason is that the Trial dataset is not good enough for an accurate physiological measurement due to issues previously mentioned (absence of subject, subject out of focus, blurry and inconsistent videos) along with the absence of ground truth.

## 3 Incorporating Physiological Measurements

### 3.1 Eulerian Video Magnification

Eulerian Video Magnification [Wu et al.(2012)Wu, Rubinstein, Shih, Guttag, Durand, and Freeman] aims to magnify subtle changes in objects or beings that are usually invisible to the naked eye. This includes both subtle color changes and motion changes. Human skin color varies slightly with blood circulation. This variation, while invisible to the naked eye, can be exploited to extract pulse rate. Motion magnification can be used to observe the respiration rate which is given away by the movement in lungs. EVM can thus be used to compute physiological factors from raw videos which in turn can be used for detecting deception. EVM involves spatial decomposition of the input frames followed by temporal processing of the pixel values to estimate the displacement signal which is then amplified and added back to the original frame to obtain the magnified frame. An example of color magnification is shown in Figure 2

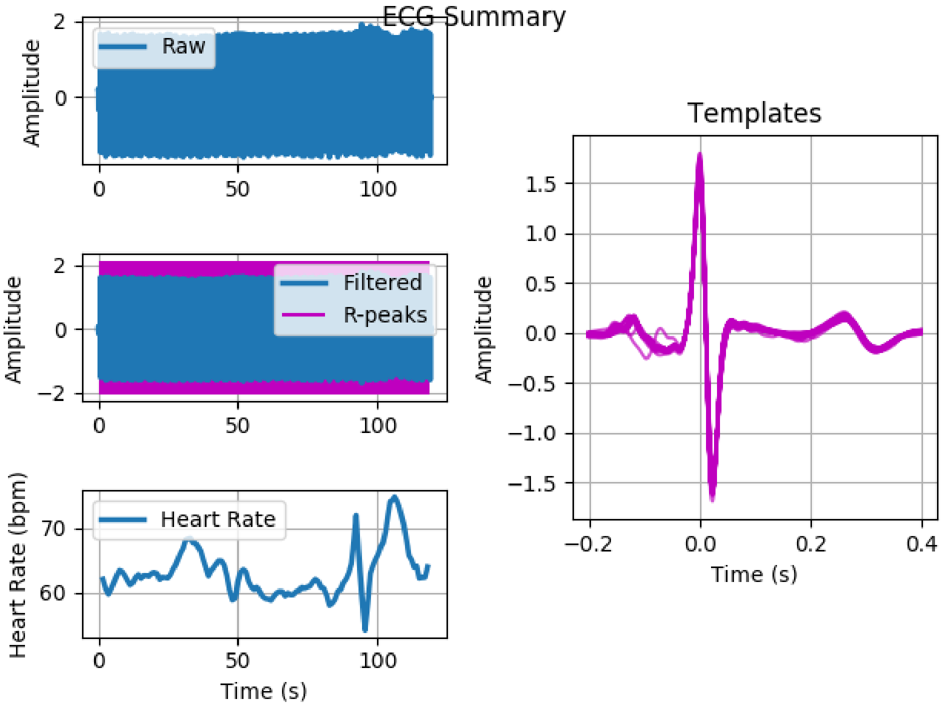


Figure 3: Features of ECG signal extracted from a subject

### 3.2 Project Creton

Project Creton was a study conducted at NTU to understand the various factors that give away deception. The data collected as a part of Project Creton includes EEG, ECG signals along with the videos of the subjects and Eye tracking data. There were 30 subjects in total and each subject was classified as either Guilty or Innocent. 14 subjects were classified as Guilty and 16 were classified as Innocent. Each subject was presented with around 170 stimuli to which a binary response was obtained. The stimuli were of different categories viz General Lie, Relevant, Irrelevant, Alpha, First Instructions Slides etc. For each stimulus, the corresponding EEG, ECG signals, Eye tracking data and the video was recorded. iMotions software was used to synchronize the data obtained from multiple sensors. The ECG sensor was a BioPac manufactured device with a sampling frequency of 500 Hz. The video was recorded with a resolution of 640x480 at 30 frames per second. The order of stimuli presented was randomized and some of the stimuli were skipped.

Biosspy PyPi package was used to filter the raw ECG signals, obtain the templates and calculate the pulse rate. The resulting pulse rates were obtained for each subject across all stimuli presented. Figure 4 shows the histogram distribution of pulse rates across all Guilty (Left) and Innocent (Right) subjects. It was observed that the Guilty subjects, in general, have higher pulse rates.

A surprising finding is that the distribution of pulse rates is similar even when considered at the individual stimuli level. Figures 5 , 6 and 7 show the pulse rate histograms for different kinds of stimuli.

## 4 rPPG

A photoplethysmogram (PPG) is an optically obtained signal that can be used to detect blood volume changes in the microvascular bed of tissue. A PPG is often obtained by using a pulse oximeter which illuminates the skin and measures changes in light absorption. While photoplethysmography commonly requires some form of contact with the human skin, remote photoplethysmography(rPPG) allows to determine physiological processes such as blood flow without skin contact. This is achieved by using face video to analyze subtle momentary changes in the subject's skin color which are not detectable to the human eye. Most approaches use the subject's face to detect these subtle color variations. The PPG waveform is composed of a quasi-periodic AC-component and a static DC-component, where the AC-component is superimposed on the much larger DC-component. The AC-component reflects cardiac synchronous changes in the blood volume with each heartbeat(i.e., the component of interest), while the DC-component is related to the intrinsic

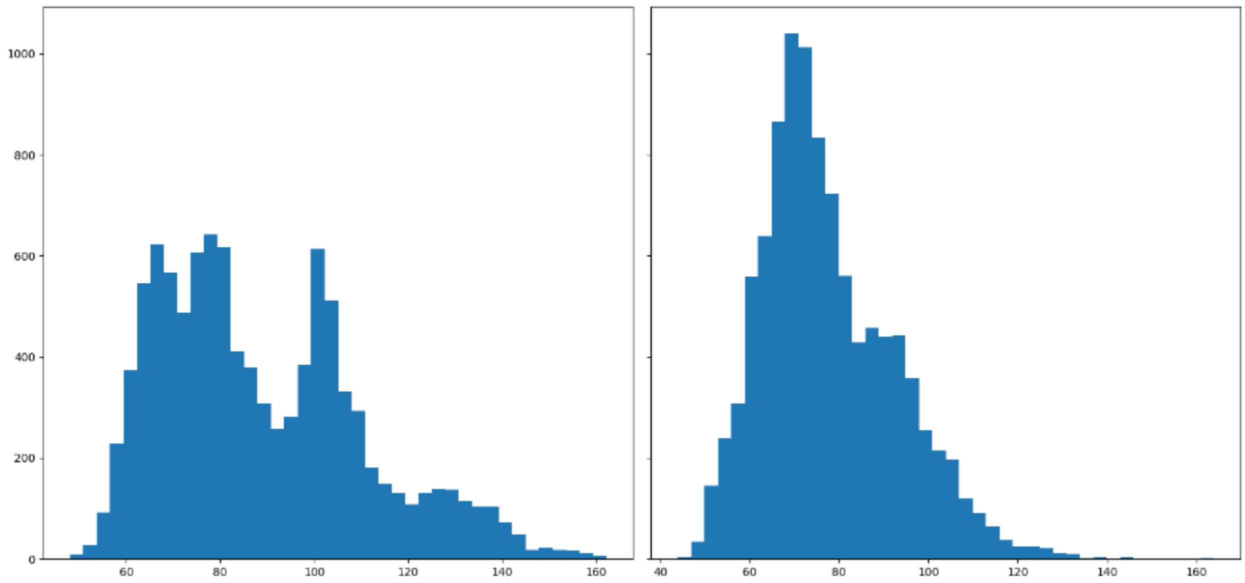


Figure 4: Distribution of pulse rates(ECG) across subjects - Guilty vs Innocent

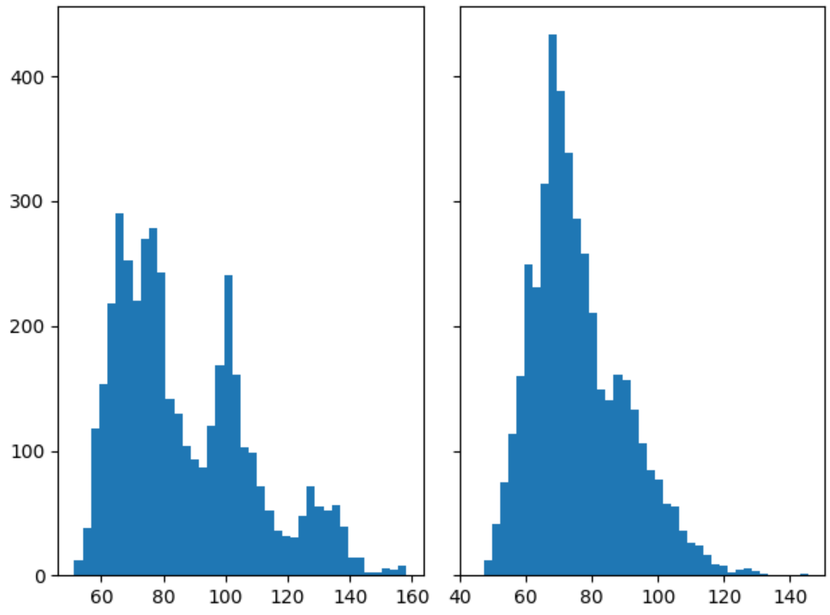


Figure 5: Distribution of pulse rates for General Lie - Guilty vs Innocent

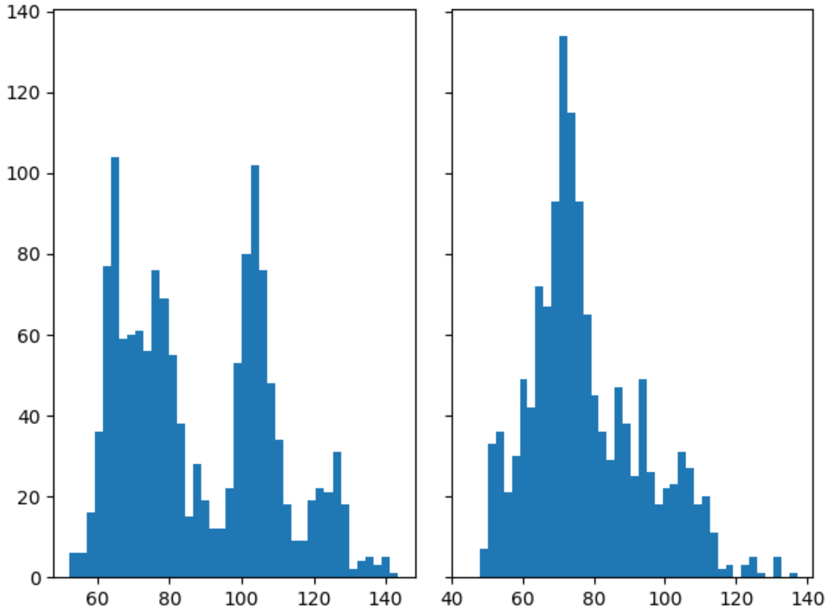


Figure 6: Distribution of pulse rates for Alpha - Guilty vs Innocent

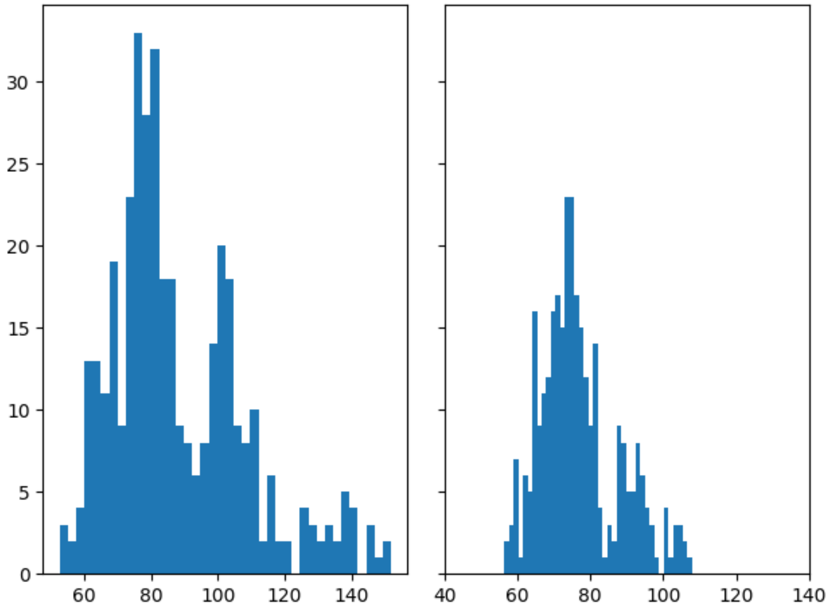


Figure 7: Distribution of pulse rates for relevant - Guilty vs Innocent

skin-color (i.e., melanin concentration), light spectra, and camera gain [Wang(2018)]. Heart rate and Pulse rate are technically different measures. In healthy individuals, these two terms have the same readings. Heart rate measures the rate at which the heart expands or contracts while Pulse rate measures the rate at which blood pressure changes throughout the body. rPPG and PPG are used to measure pulse rates .

### 4.1 Process

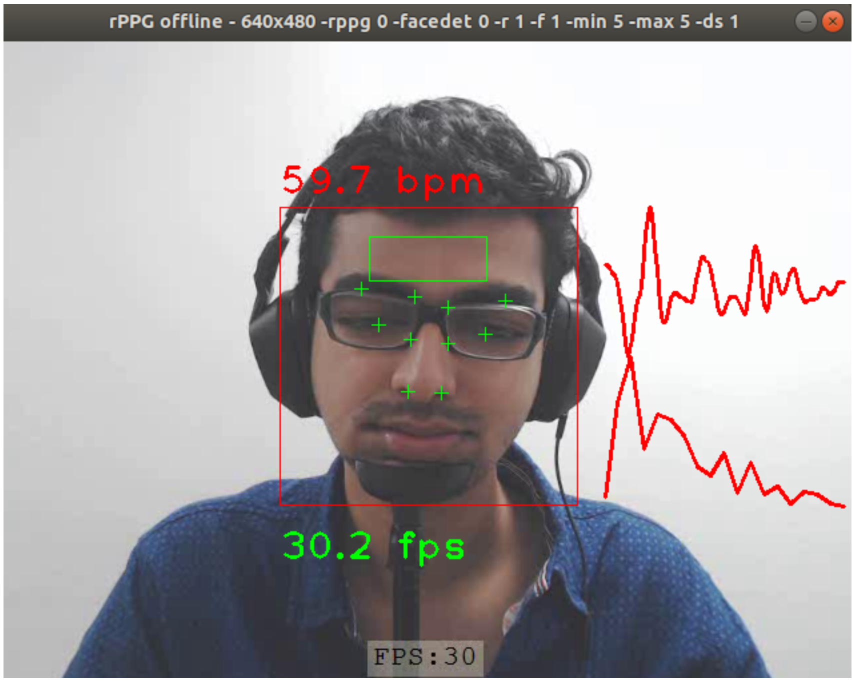
rPPG can usually be deployed in two ways, depending on the positions of the light source and photo detector. The first uses the Transmission model where the LED and the photo detector are placed on the opposite sides of the skin. The Reflective model requires the LED and the photo detector to be on the same side of the skin. The camera based rPPG method relies on the reflective model. There are a few constraints that need to be imposed when deploying camera based PPG. The automatic image enhancement functionalities have to be disabled during the measurement as they will change the pixel values and potentially corrupt the pulse induced color change. The images also need to be processed in raw and uncompressed format to prevent the loss of information. It has been found that the forehead region contains the maximum signal-to-noise ratio that is useful for detecting the PPG signal. This implies that that the rPPG algorithm needs to extract the region of interest (ROI) that corresponds to the subject's face. Recent advances in deep learning enable robust detection of facial landmarks and boundaries. Movement of the subject may cause noise that renders the signal useless for analysis. So, it is necessary to employ ROI tracking to ensure that the pixels contained in the ROI belong to a skin region invariant to subject motion. One of the methods of tracking the ROI is to re-detect it for every frame. This can be computationally expensive especially with deep learning based face detection. So, a set of feature points can be computed from the ROI that can help detect the ROI of the subsequent frames through an affine transformation. After ROI detection in multiple frames, the raw RGB signal is extracted from these regions. Despite ROI tracking, the raw extracted signal might contain unwanted noise from illumination changes, motion and other factors. Since the Heart rate signal lies in a particular frequency range [0.7 Hz, 4 Hz], digital filters are applied to remove unwanted noise from the signal. The filter is applied after normalization of the series of RGB values. The cutoff frequencies of the filter can be dynamic based on previous estimated Heart rate. It is a fair to assume that the raw signal contains a one dimensional plethysmographic signal p(t), which can be represented as a linear combination of the raw signals using a convex combination. In practice, Blind source separation, ICA (Independent Component Analysis), PCA (Principal component analysis) algorithms are used to break the raw signal down to lower-dimensional components and an optimal combination of these components that leads to a good estimate of p(t) is found. After estimation of the plethysmographic signal, the Heart rate can be measured by performing a frequency analysis on p(t). The signal, p(t) is converted into frequency domain using the Fast Fourier Transform (FFT) Algorithm. In the frequency domain, frequency corresponding to the index with the highest spectral power is chosen as an estimate for the Heart rate.

### 4.2 Application on Creton data and Related problems

The results of rPPG were supposed to be synchronized with the ECG signal's output. This would have allowed for improvements in the unsupervised rPPG setting. The plan was to also try a supervised learning version of rPPG since the ground truth is available as the ECG signal. iMotions failed to synchronize the video recordings corresponding to the stimuli which led to erroneous and broken videos in the data. Only 6 out of 30 videos have the complete recordings. Most of the videos are less than 3 minutes long with random segments missing. This makes the data illegible to be analyzed in the context of rPPG. The output from rPPG cannot be matched with the ground truth because of these missing segments.

### 4.3 New Data

To overcome the problems with the data collection of Creton, new data has been collected with only ECG and the corresponding video. Three subjects were shown three movie trailers of different genres viz Flipped (Drama), Kung Fu Panda (Humor) and Hush (Thriller) in the same order. A blank slide was shown at the beginning for 30 seconds in order to establish a baseline measurement. The BioPac ECG sensor was used to collect the measurements at a sampling rate of 1000 Hz. The total duration for which each subject was monitored is 428 seconds. The ECG features extracted from one of the subjects is shown in Figure 3. The data and the sample output can be found **here**. The rPPG sample output is shown in Figure 8 . The green rectangle in the figure represents the Region of Interest from which subtle color changes are captured. The two red graphs on the right represent the extracted PPG signal (top) and Power Spectrum (bottom). The Power spectrum signifies the decomposition of the PPG waveform into component frequencies. This is done by performing frequency analysis on the PPG waveform (Fourier Transform). Ideally, there would be a single peak in the power spectrum that corresponds to the pulse rate. Evaluation of the previous rPPG approach on the new data has revealed that the pulse rate predicted by the rPPG algorithm differs from the one extracted from the ECG signal by an average of 20 beats per minute.



rPPG offline - 640x480 -rppg 0 -facedet 0 -r 1 -f 1 -min 5 -max 5 -ds 1

Figure 8: rPPG on a subject

The current rPPG approach is sensitive to the skin tone of the person. With darker skin tones, the colour signal gets weaker, hence the algorithm's performance will be reduced. Spatial averaging did not yield significant improvement in the performance.New approaches demonstrated an adaptive extraction of PPG signal which the current implementation doesn't deal with. Another area that could be improved upon is the dimensionality reduction step of the algorithm. A PCA based compression might be too lossy to reveal the underlying PPG component.

## 5 Adversarial Attacks

Neural networks, despite their significant task performance and inspiration from biological systems, often behave in ways that challenge the intuition of human observers; signalling a peculiar fragility and brittleness present in the models that are anticipated to take over the field. This brittleness is often highlighted in the context of model sensitivity to very specific changes in inputs known as adversarial examples. In simpler terms, adversarial examples are inputs that are specifically crafted to fool machine learning classifiers. In most cases, adversarial examples are indistinguishable from their "real" counterparts to humans. The nature of adversarial examples is quite concerning since an attacker could feed inputs that are designed to yield a certain outcome from the machine learning model. It turns out that adversarial examples generated from a specific model are transferable to other machine learning classifiers as well. This property, referred to as "transferability of adversarial examples" is particularly problematic because malicious inputs could cause significant damage when high stake decisions are made or driven by machine learning models. Fortunately, attacking text based models is more complicated than attacking image classification models due to the form of textual input. Text based models map each word in the input to a vector, referred to as Word Embeddings. Due to the discrete mapping of each word to a fixed vector, it is hard to craft adversarial inputs without making them different from the source inputs. So, adversarial inputs in the text domain instead focus on retaining the semantic information in the input. Adversarial examples in this case are crafted by replacing a chosen set of words with the words that correspond to the nearest embeddings.

The CNN implemented in [Britz(2015)] is used to craft adversarial examples. By replacing a set of words with the nearest learned embeddings, we were provided with interesting insights into the dynamics and learnings of the deep learning model. We demonstrate the results of the word replacement by crafting the adversarial example from a truthful review

**Original sample**:
*my husband and i arrived at the swissotel chicago to celebrate our 13th wedding* ***anniversary****. upon arrival at the*

*given and advertised check in time* ***our*** *room was not ready we waited for more than an* ***hour with*** *our bags*

**Sample with nearest embedding replacement**:
*my husband and i arrived at the swissotel chicago to celebrate our 13th wedding* ***it****. upon arrival at the given and advertised check in time* ***they*** *room was not ready we waited for more than an* ***were up*** *our bags*

The words highlighted correspond to the words that have been replaced with the nearest embeddings. It is interesting to note that the CNN embeddings layer has found the word *it* to be the one closest to the word *anniversary*. It is clear that the crafted example is not semantically consistent, yet, the model classifies this as a truthful review with a very high confidence of $0.99$, the same result as the original sample. Hence, we are able to successfully tamper with the input without affecting the result.

## 6 Conclusion

This work explored the possibility of measuring physiological signals remotely through video analysis. During the intial phase of the internship, the feasibility of video analysis on Real Life Trial dataset was studied. It was concluded that the dataset is not ideal for video analysis due to various issues ranging from inconsistent quality to scalability. Video with corresponding physiological signals (ECG, EEG etc) was collected during the study for Project Creton. Interestingly, analysis of ECG signals revealed that gulity subjects, in general, have a higher average heart rate than innocent subjects. Moreover, this pattern was valid at an individual stimuli level. Remote Photoplethysmography (rPPG) offers a non-invasive method for predicting pulse rate from raw videos. It involves detecting and analyzing subtle color changes in the skin. Project Creton's data was deemed unfit for rPPG prediction due to a problem with the data collection and export process. It resulted in broken videos with missing stimuli. To overcome this problem, new data was collected with only the video data and the corresponding ECG signal. rPPG analysis on the new data yielded promising results although the predictions could be improved. Finally, an adversarial attack on text model was performed and the ground truth has been succesfully altered while keeping the model's predictions intact. This revealed insights into the neural network's learnings. In conclusion, initial steps that look promising have been taken to measure pulse rate remotely through rPPG. With better algorithms and compression techniques, the results could be improved upon. In addition to rPPG, vulnerability of current machine learning models was explored and demonstrated.

## References

[Britz(2015)] Denny Britz. Convolutional neural network for text classification in tensorflow. 2015. URL `https://github.com/dennybritz/cnn-text-classification-tf`.

[Krishnamurthy et al.(2018)Krishnamurthy, Majumder, Poria, and Cambria] Gangeshwar Krishnamurthy, Navonil Majumder, Soujanya Poria, and Erik Cambria. A deep learning approach for multimodal deception detection. *CoRR*, abs/1803.00344, 2018. URL `http://arxiv.org/abs/1803.00344`.

[Pérez-Rosas(2015)] Abouelenien Mohamed Mihalcea Rada Burzo Mihai Pérez-Rosas, Verónica. Deception detection using real-life trial data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 59–66, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3912-4. doi: 10.1145/2818346.2820758. URL `http://doi.acm.org/10.1145/2818346.2820758`.

[Wang(2018)] Wenjin Wang. Robust and automatic remote photoplethysmography. 2018. URL `https://drive.google.com/file/d/0B6HctXoTd1wFQUt2X0l5M2NRY2s/view`.

[Wu et al.(2012)Wu, Rubinstein, Shih, Guttag, Durand, and Freeman] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William T. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (Proc. SIGGRAPH 2012)*, 31(4), 2012.

[Zhe Wu(2017)] Larry S. Davis V. S. Subrahmanian Zhe Wu, Bharat Singh. Deception detection in videos. *CoRR*, abs/1712.04415, 2017. URL `http://arxiv.org/abs/1712.04415`.