Video Magnification to Perceive Anxiety

Adam J. Fisch
Department of Mechanical Engineering
Princeton University, Princeton, NJ 08540
afisch@princeton.edu

Max Shatkhin
Department of Electrical Engineering
Princeton University, Princeton, NJ 08540
shatkhin@princeton.edu

Abstract

To visually recognize possible feelings of anxiety in humans, video magnification and signal analysis is used to extract physically manifested cues normally imperceptible to the naked eye. Our algorithm uses a Hidden Markov Model with anxiety as states and detected heart rate and blushing levels as emitted evidence variables. The Viterbi algorithm is used to generate the most likely sequence of anxiety levels over the duration of the subject's video.

1. Introduction

> *"There's no art to find the mind's construction in the face"*
>
> – Shakespeare

When presented with a stressful, anxiety filled situation, the inner state of a person's mind is excited. Blood can start rushing to the face and ears, making them feel hot. The heart rate might quicken and grow stronger, and be felt pounding in the chest and ears. A normal outside observer, however, will likely never know the difference. The human vision system can only see the tip of the mental iceberg, and any physiological changes due to emotional changes are commonly imperceptible. Thus, as Shakespeare wrote in Macbeth, there is no way to read someone's mind.

Perhaps not with Shakespeare's own eyes. Current research in video magnification and computer vision, however, has yielded advances in developing a "visual microscope" for the world [11]. By magnifying tiny changes over time in video frames, amazing details can be extracted – and normally invisible changes can clearly be seen. These observations, such as vital signs, can help give a small glimpse into the psychological conditions behind the skin. In this project we seek to algorithmically obtain and leverage this information to help predict a person's general state of anxiety, embarrassment, or nervousness.

Convincing studies correlate certain autonomic reactions such as heart rate and blushing response to elements of stress [3]. Unlike most facial expressions, these subconscious changes are involuntary products of the fight or flight response, and thus cannot easily be suppressed or controlled. Changes in heart rate and redness can therefore act as a good, partial evidence variables for estimating the actual inner conditions of a person.

While concealed to the naked eye, these telling signals can be measured with video magnification. In particular, as the heart beats, blood pumped to the face causes capillaries close to the skin surface to swell and contract periodically. Consequently, the face becomes ever so slightly redder and paler over time. As someone blushes, the mean baseline redness of the face also slowly increases. By analyzing these tiny color variations over time, useful information such as the subject's heart rate [6] or small movements [5] can be determined.

Filtering video of a face over the human range of heart rate frequencies allows the cyclical pulse signal to be traced and converted to beats per minute. Similarly, when filtering at low-frequencies, the blush response can be detected by analyzing trends in the Hue-Saturation-Value (HSV) color space. Blushing manifests itself in low-frequency variations in the saturation magnitude of red-based colors.

Using those techniques in our implementation to record heart rate and blushing traces over time, we then take the output to build a Hidden Markov Model (HMM). Manipulations on the HMM probability tables are used to produce estimates of the evolution of the anxiety states over time. Although still a rough generalization with error, the generated classification output gives us the ability to reconstruct the emotional state of mind.

2. Related Work

Motion detection and magnification in videos has been a popular research topic for some time. Most approaches borrow either a Lagrangian or Eulerian type analysis from fluid dynamics when analyzing the "flow" of video pixel data. Lagrangian techniques analyze the motion of specific

feature points over time, while Eulerian systems look at the change in value of a fixed pixel location over time.

Examples of video magnification used to analyze tiny changes include Liu et al. [5] of MIT, who use extracted, Lagrangian, feature point trajectories to measure small motions, such as the deformation of a swing set beam under loading. Wu et al, [11], also of MIT, developed the first Eulerian model that magnified variations that occurred in special temporal and spatial frequency bands at specific locations in a window. This is particularly applicable to periodic motions that have a special frequency spectrum – such as a vibrating guitar string.

This research has also yielded non-invasive methods of detecting heart rate, respiration rate, and other vital signs. Heart rate extraction is a popular area of development, with pulse being detected by methods ranging from analyzing bobbing head motions [1], to using blind source separation and frequency analysis on the averaged traces of the video RGB color channels [6]. To our knowledge, however, using extracted physical cues to infer internal, emotional states has not yet been attempted.

3. Methods

To perform our analysis, we created our own entire pipeline in Matlab. The video is loaded and stored, processed to measure the heart rate signal and blush trend, and then converted into features to build an HMM and other trained classifiers.

3.1. Heart Rate Detection

Developing a robust algorithm for detecting the heart rate of the subject in the video proved to be quite challenging, and several approaches were explored. Heart rate detection has been proposed and shown in several previous works such as [6] and [8], however we found that our own implementations of those algorithms didn't quite perform as well as expected and needed.

In [6], the authors average the color components over each frame, giving a singly dimensioned trace for each red, green, and blue channel over the length of the video. Independent component analysis is then performed on these three signals, forming three new signals. Independent component analysis is a form of blind source separation that seeks to separate independent signals from linear mixtures in the absence of any other information. Fourier frequency analysis was then performed on second ICA signal, which was empirically found to give better results than the other two signals. In our implementation, we found that simply taking the maximum frequency response of the power density - frequency plots from the Fast Fourier Transform (FFT) gave unreliable results unless the signals were very clean. Slight movement of the subject in the video sequence, or what appeared to be aliased light signals from indoor lamps, at times added significant noise to our graphs obscuring the true heart rate frequencies. Interestingly, videos from [11] were remarkably noise free, which leads us to believe that they had a better video acquisition setup.

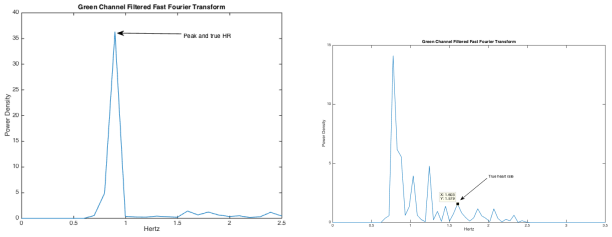


Figure 1. A clean FFT signal (left) vs. a noisy FFT signal (right). True HR response of the noisy signal is overshadowed. Clean signal video source from [11].

In another approach, the Eulerian magnification method in [8] avoided averaging over the whole region of interest by capturing many time varying signals of individual fixed pixels. These pixel signal peaks are then used in a pulse-onset detection process to deduce the heart rate. This process was also attempted for our implementation, however, again, extraneous motion in videos caused artifacts that interfered with our calculations.

Ultimately, our algorithm combines aspects of both approaches. Our algorithm's flow is shown in Fig. 3. In order to reduce noise in the color signal, several processing techniques are applied to the data. First obvious sections of the video frames are omitted by prompting the user to select the face region, and additionally, to box off any outlying areas such as the eyes, any facial hair covered regions, and the mouth. The frames are then further filtered spatially by blurring and downsampling the video sequence in a gaussian pyramid, which removed high frequency variations. After this is completed, the face is further segmented into "skin" vs "not skin" based on color. This was done in a basic manner by selecting pixels with a red channel intensity higher than the grayscale luminance, with the premise

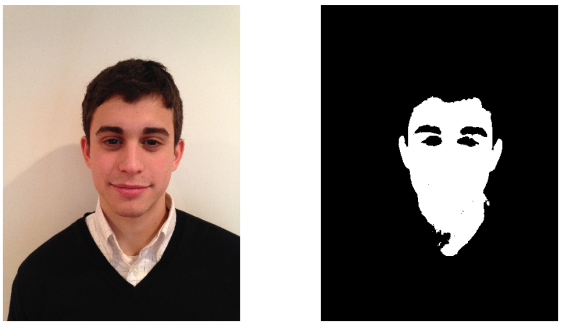


Figure 2. An example of basic image segmentation to select skin tones from a background.

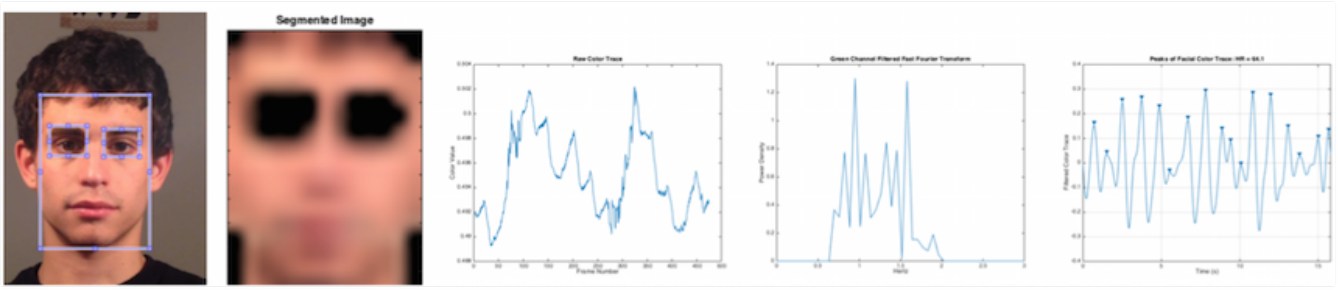


Figure 3. Pipeline: Bounding box selection - Spatial filtering with skin segmentation - Averaged trace - Temporal filtering - Peak detection.

that this simple metric represents candidate reddish pixels that are not too dark.

Finally, the green color channel is averaged over each frame, and this trace is selected to be used in the next parts of the process. Green is used over the red and blue channels as it has been shown to contain the strongest plethysmographic waveform [10].

While at this stage of the process much of the noise that is introduced from spatial differences has been removed, the green trace still needs to be processed further in the temporal domain. Often the signals experience some drift, thus to remedy this the trace is detrended by subtracting a low-order polynomial best fit from the data. This isolates the oscillatory motion. Then it is statistically normalized to have zero mean and unit variance by taking:

$$g_{norm} = \frac{g - mean(g)}{std(g)} \tag{1}$$

To temporally filter the signal such that only frequencies in the range of valid human heart rates are considered, the signal is converted to frequency space by taking a FFT. All frequencies outside a specified heart rate range are suppressed. This heart range can be given by the user the finer the window, the more accurate the heart rate measurement will be. If no prior information is known about the state of the subject, e.g. if they are an at rest adult, healthy male (<110 BPM), then a general range from 40 - 180 BPM can be used. The frequency filter signal is converted back to real space and is in its final form. Peak analysis is used to local maximums along the curve, under the constraint that they are spaced at least a minimum time span apart. This time span is determined by the upper limit of the heart rate range. Finally, the heart rate is calculated from the average time between peaks:

$$HR = 60/\Delta t_{avg} \tag{2}$$

To verify our method, we compared heart rate measurements with readings obtained by a commercial heart rate monitor. Results were also compared on videos taken from [11], where heart rate truths were provided. Four samples are shown in Fig. 4. Measurements were best for videos with good lighting, little movement, and wide faces with pronounced skin coloring. Poor video quality had a very adverse effect on measurement accuracy.

As can be seen in Fig. 4, the peak analysis yielded a greater accuracy in the heart rate measurements. Heart rates are not perfectly uniform, and arrhythmias in the signal are not uncommon. Additionally, the color traces sometimes show beat frequencies in magnitude. Thus, for such signals where several frequencies are represented as opposed to one strong, pure frequency, FFT analysis can be confused. Peak to peak analysis is more robust to these conditions.

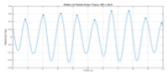
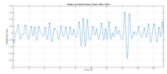
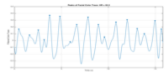
Input	Color Trace	HR: FFT	HR: Peak - Peak	HR: Actual
		152	152	154
		54	54	50-60
		61	94	100
		42	64	58

Figure 4. Comparison of detected heart rates to ground truths.

3.2. Blush Detection

While heart rate detection analyzes the frequency of peaks in the normalized color trace signal, the blush response is captured by looking at the general trending changes in magnitude of these color peaks over a larger time scale. In order to obtain this signal from the normalized color trace we experimented with a couple of different approaches.

To low-pass filter the signal and extract any trends that might be present, we first applied a best-fit low order polynomial to the trace. However, we saw that the polynomial fit lost accuracy quickly due to byproducts of the continuous, polynomial constraints on the problem. Filtering the

Figure 5. Illustration of hue and saturation metrics in the HSV color space [4]. Saturated colors in the red range are searched for when attempting to detect blush.

data directly with a moving median filter gave better fitting results. After smoothing, the signal was shifted and normalized based on a reference point. This was taken as the minimum value across the signal – so that all outputs were positive. The blushing response metric was calculated as the percent change from the baseline to obtain a dimensionless metric suitable for individuals with varying physical characteristics, such as skin tone.

The color space in which the blush response is perceived was a defining factor with respect to the effectiveness of our algorithm. Research was done about a number of color spaces, each having its own pros and cons when dealing with skin color domains [2]. The RGB color space is common and simple to manipulate, however the lack of a distinct luminance or intensity measurement makes it unreliable for applications where lighting is an important factor. Instead, our method analyzed the HSV color space, where colors are separated by their hue, saturation, and value. The HSV color space is similar to the way humans perceive color. Hue represents color, given as an angle from 0 to 360 degrees around a color wheel, illustrated in Fig. 3.2. Saturation represents the purity of the color. Colors with the same hue but lower saturations appear faded. Value is a metric for the brightness of the color.

For detecting color changed that might be corresponding to a blush, we separated out reddish colors with hue values in the range of 0 to 18 and 350 to 360 degrees. Filtering out the other colors, the change in saturation over time was then measured. Fig. 3.2 below shows the magnified result of decreasing and increasing saturation levels for colors in the red hue space, while keeping the value parameter constant. Visually, it is confirmed that this is indeed similar to the

Figure 6. Input image of a face and the result of filtering on the desired hue range.

Figure 7. Saturation levels of pixels in the desired hue range amplified to -50% and +50% illustrate an exaggerated blush response. Value is constant.

physiological color change that take place during blushing. Further testing was done using temporal analysis on videos of control subjects seen in the data acquisition section of this report. The saturation trend line proved to be a good fit for a high proportion of the test cases, using both user feedback, expected responses, and visual analysis, when applicable, as baselines.

3.3. Classification

3.3.1 HMMs and the Viterbi Algorithm

In our primary attempt to infer information about the subject's state of anxiety given the detected physical cues, we chose to design and train a Hidden Markov Model (HMM). An HMM is a temporal probabilistic model that can be used to predict the evolution of the state of the "world" at each point in time. As depicted in Fig. 8, an HMM is comprised of a hidden sequence of transitioning states, $x_0, x_1, ..., x_T$. At each time $t \rightarrow T$, these states emit certain evidence variables, $y_0, ..., y_T$.

In our case, the hidden states are a subject's true level of stress and anxiety, while the detected heart rate and blush response are the observable evidence variables. We construct the sequence of outputs by processing an input video and extracting a time-varying trace of heart rate and and blush measurements.

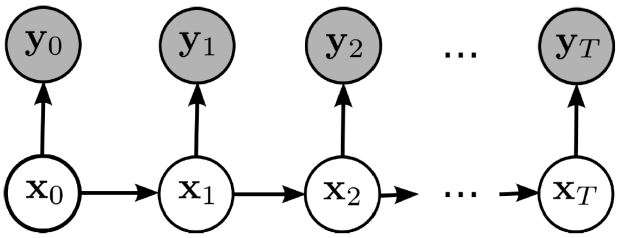


Figure 8. Sequence of a Markov Model with hidden states and their observable outputs [7]

If the anxiety state world is indeed a Markov process, which is a reasonable assumption, then the anxiety level corresponding to state t depends on the state or multiple states that came before it, and the outputs at time t depend on the state at t. Thus, an HMM like this is defined by there sets of probabilities:

1. $P(E[t] = y|X[t] = x) \rightarrow$ the probability of observing output *y* at a state *x*
2. $P(X[t+1] = x'|X[t] = x) \rightarrow$ the probability of transitioning from state *y* to state *x'*
3. $P(X[0]) \rightarrow$ the starting state probability distribution

Therefore, a HMM has two probability matrices as its main data structures: a state transition matrix giving the probability of going from x_i to x_j and an accompanying output probability matrix of seeing evidence y_i given state x_i. For our purposes, we assume that $P(X[0])$ is 1 for always starting in a non anxious state. If these conditional distributions are known, then given only a sequence of observations, we can then find the progression of hidden states that is most likely to have produced what was seen[9]. This can be accomplished with the Viterbi algorithm, which generates the most likely path of states by solving for the set that maximizes the conditional probability based on the observed evidence [9].

$$Path = \max_{\boldsymbol{x}_1 \ldots \boldsymbol{x}_t} \boldsymbol{P}(\boldsymbol{x}_1, ..., \boldsymbol{x}_2, \boldsymbol{X}_{t+1} | \boldsymbol{e}_{1:t+1}) \qquad (3)$$

Using Viterbi, we then generate our predictions based what was the most likely anxiety evolution was.

We chose to focus on using an HMM rather than a type of linear classifier such as a Neural Network or SVM because of the temporal properties that the HMM encompasses. The Viterbi algorithm uses information from both past and future points to generate the most likely state at a given time. This allows us to change the question from, "given the static output at time t, what is the state?" as it would be for a classifier, to "given the temporal *sequence* of outputs, what *was* the most probable state of nervousness at time t?" Seeing as the temporal variation of the heart rate and blushing metrics is likely as, or more, telling than the actual values themselves, we anticipated the HMM to be more useful and well suited to our application.

3.3.2 Constructing the HMM

In order to run the Viterbi algorithm, the HMM states, outputs, and conditional distributions had to be constructed and learned. We converted perceived anxiety into discrete states by giving levels of anxiety distinct scores. We experimented with both binary classifications nervous or not nervous as well as with multiple bins, such as a anxiety levels ranked from 1 - 5.

For the outputs, we combined both heart rate and blushing into a single "megavariable" tuple. We also discretized heart rate and blushing by binning them by magnitude ranges. In order to make the heart rate evidence less influenced by individual characteristics such as resting heart rate, we logged the evidence as the change in heart rate from a baseline taken at the first measurement rather than by its absolute value.

To generate the state transition and output emission probability matrices we trained on data gathered in Section 4. By logging, normalizing, and Laplace smoothing counts of how many times state *x* transitioned to *x'*, and likewise how many times *x* emitted output *y*, we obtained estimates for the matrices.

Due to time limitations, we were not able to collect as much training data as would be desirable for accurate results. Thus, our estimates for both state transition and output probability were still not yet convergent, rough estimates.

4. Results

4.1. Experimental Setup and Training

Training experiments were conducted on 9 individuals in a controlled setting. These individuals were asked to sit in front of an Iphone 5s camera while being asked a series of questions designed to elicit responses that indicated some level of stress or nervousness. The area where they were seated was well lit with a strong, continuous light source to mitigate variations of lighting conditions that might contaminate the videos. The Iphone camera was constrained by a mechanical clamp to avoid motion blurring and oscillations due to any hand tremors. Furthermore, our subjects were instructed to remain as still as possible in front of the camera, and to avoid making any sharp movements or changes in facial expression.

Questions were asked to them in a yes or no format and responses were given simply by giving a "thumbs-up" or "thumbs-down" out of the frame. Each video lasted approximately 30-40 seconds and 5 to 6 questions were asked

at evenly spaced intervals throughout. After the video was captured, we asked subjects to rate, on a scale of 1 to 5, how nervous they felt after being asked each question. It was explained to them that the measurement of nervousness could be due to self-noticed autonomic responses, such as feeling that their heart started to beat faster, or to just an overall level of discomfort with the question.

The sequence of questions was designed to establish baselines or reference points of stress signs by introducing them with simple questions such as "Do you attend Princeton University?" or "Is your favorite color blue?" From there, the questions increased in anticipated stress response by asking seemingly simple questions designed to make them possibly hesitate and become flustered. Examples included:

- Is 67 divisible by 9?
- Is 5! greater than 2^7?
- Have there been greater than 53 US Presidents?

From the initial training videos, we obtained output and state transition sequences by tracking heart rate and blushing over time, and matching up those outputs to the subject identified states. In order to extend the state transitions over the entire duration of the video, we spline interpolated between the labeled states, and then rounded the trajectories to states 1-5. Training results for the state transitions are shown in Fig. 9. Output results varied per individual, but many of the heart rate and blush trajectories looked to be promisingly correlated with the labeled data.

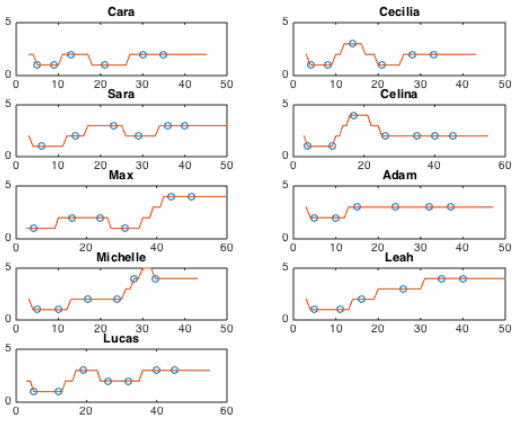


Figure 9. States from training videos

4.2. Testing

After gathering the training data was completed, new subjects were invited in to be tested. The experimental setup

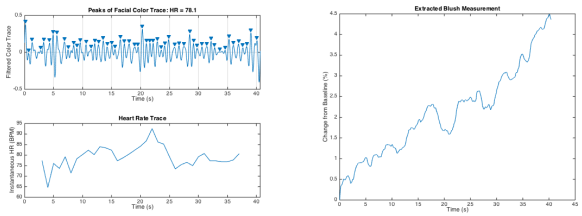


Figure 10. Test subject heart rate and blush response.

for the test subjects was the similar, with a variety of stressful and easy questions presented to each participant.

Processing the videos yielded encouraging results. While not perfect, the output from the HMM did match a good amount of the trend of the actual data. Example results for one of the test subjects is shown in Fig. 11.

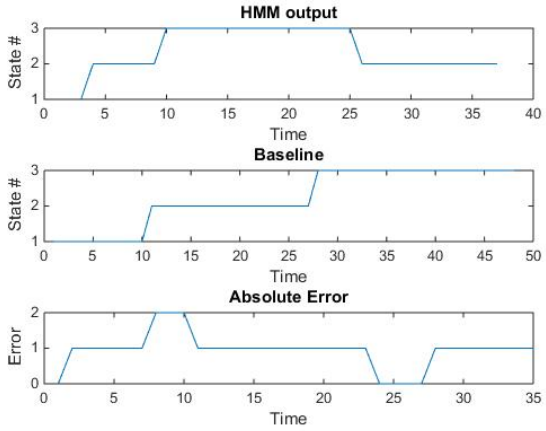


Figure 11. HMM anxiety states output for a test subject

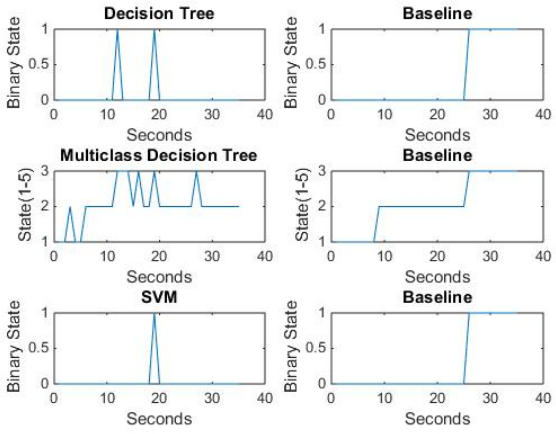


Figure 12. Alternative classifier results for the test subject

To check whether the HMM was indeed as good a choice for a predictor as we initially assumed, we also took the

training and trained three more types of classifiers. A binary output decision tree that distinguished between nervous and not nervous, a multiple class decision tree that further tried to discern the specific anxiety level (1-5), and a SVM were trained and used for testing. Results for the same test subject in Fig. 4.2 are shown in Fig. 12. While, again, the training and testing data was too sparse to get an accurate measure of how the algorithms truly performed, it did initially appear that the HMM does give a smoother, better hypothesis.

5. Conclusion

In some regards, in this project we were indeed able to take a peak behind the mask that it the human face. However, it is also clear that we are not yet reliably able to do so. We believe that the ambiguities in our results are primarily due to a shortage of data, unreliable user feedback data, and a simplification of the biological processes behind anxiety.

Although we asked a number of subjects to participate in our study, we believe that in order for algorithms such as a HMM to have realistic transition probabilities, or for classification algorithms to be tuned to the right parameters, a considerable data set must be obtained. We anticipate that if we had the time to gather a larger quantity and variety of data, our results would have been more positively reflected. Furthermore, our baseline for testing our results was based on user-feedback of how nervous they felt. This method was the only one available to us at the time, and so comparing the output of our HMM and classifiers does not reveal the actual error in our results. Because of the nature of uncertainty of feeling stressed, this feedback is hard to come by.

Finally, we attempt to model stress as a function of two autonomic signals, namely heart rate and blushing response, when the actual process is undoubtedly a function of a much wider variety of factors. This is, in itself, not a reason for the method to fail as the two vital signs could in fact hold enough information to detect stress within a reasonable margin, however without enough data to see the fluctuations of these variables in different situations, making a prediction is difficult.

Those realizations and factors notwithstanding, as described in our report, the results of our project were quite encouraging. Given more time, we would put more effort into solving the first two issues in the preceding paragraph. As in the case of the history of Neural Networks, acquiring a significant amount of data might give our method the resources it needs to make good state estimations. We would also attempt to find a more reliable method for testing whether our results were truly indicative of levels of stress. This might be done by measuring heart rate, blood pressure, or other known signs of stress with more invasive tools at the same time as capturing them on video.

With consistent iteration, comes more consistent results, and hopefully a robust "emotional microscope" with which to analyze video.

References

[1] G. Balakrishnan, F. Durand, and J. Guttag. Detecting Pulse from Head Motions in Video. In *2013 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR)*, IEEE Conference on Computer Vision and Pattern Recognition, pages 3430–3437. IEEE; IEEE Comp Soc, 2013. 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, JUN 23-28, 2013.

[2] A. Elgammal, C. Muang, and H. Dunxu. Skin detection - a short tutorial, 2009.

[3] S. G. Hofmann, D. A. Moscovitch, and H.-J. Kim. Autonomic correlates of social anxiety and embarrassment in shy and non-shy individuals. *INTERNATIONAL JOURNAL OF PSYCHOPHYSIOLOGY*, 61(2):134–142, AUG 2006.

[4] *Introduction to well, colors*. Retrieved on January 13 from Clear Support at http://clearsupport.freshdesk.com/support/articles/65805-introduction-to-well-colors.

[5] C. Liu, A. Torralba, W. T. Freeman, F. Durand, and E. H. Adelson. Motion magnification. *ACM Trans. Graph.*, 24(3):519–526, July 2005.

[6] M.-Z. Poh, D. J. McDuff, and R. W. Picard. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Express*, 18(10):10762–10774, May 2010.

[7] P. Protopapa. Lecture 19: Hidden markov models, 2014.

[8] M. Rubinstein. *Analysis and Visualization of Temporal Variations in Video*. PhD thesis, Massachusetts Institute of Technology, Feb 2014.

[9] S. J. Russell and P. Norvig. *Artificial intelligence: a modern approach (3rd edition)*. Prentice Hall, 2009.

[10] A. Rustand. *Ambient-light Photoplethysmography*. PhD thesis, Norwegian University of Science and Technology, Jun 2012.

[11] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.*, 31(4):65:1–65:8, July 2012.