Facial Boundary Recognition and Heart Rate Extraction From Remote Video Source

Vlad Slyusar

ECE 5831: Pattern Recognition & Neural Networks
University of Michigan

December 16, 2016

0.1 Introduction

The goal of this paper centers on investigating techniques for heart rate extraction from an ordinary video source. While such biometric information is theoretically present within any uncovered region of the body that contains a pulse, the implementation focuses on attempting to extract the heartbeat specifically from a human face. A forward-facing head region provides less of a challenge in detection and tracking throughout the video than an arbitrary body part with exposed skin. The first frame of the video is used for initial face detection with user prompt for choice in the case of multiple targets. The selected face is tracked sequentially across all frames via its representative feature vectors. The tracking produces coordinates for a facial boundary for each individual frame. The identified boundaries are extracted, rotated upright, and scaled to the same size producing a cutout of the face adjusted for rotation and distance variation. In order to reduce noise from movement or breathing, the tracked facial boundary can also be optionally stabilized using a phase-based approach[WRDF13]. Next, the Eularian video magnification technique can be applied to amplify frequencies within a selected range corresponding to a valid heartbeat[WRS$^+$12]. The amplified video is processed to determine the color variations within linear progression of the frames and the most prominent frequency is identified as the candidate heart rate and adjusted from Hz to Beats/Minute. In the face of significant background noise, the process can be repeated with a user specified region that does not contain faces or exposed body parts. The results are used as a baseline in an attempt to remove the bias frequencies extracted from the face and isolate the heartbeat.

0.2 Problem Statement

While camera based heart rate detection systems already exist, the most popular implementation of the technology requires direct contact to skin, generally by placing a finger directly over the camera lens. Such an approach has become quite common with the relative abundance of personal devices encompassing both a camera and an LED within relative proximity. By applying a strong illumination to the skin pressed firmly against the camera, the color intensity fluctuations induced by blood circulation can be tracked to extrapolate an approximate rate of heartbeat[Mel13]. While most applications with such functionality strongly discourage using results for medical work, the simple and unobtrusive nature of the method provides a good approximation with little special equipment or complicated procedures. Since the heartbeat can be extracted in this way using only the input data from a camera, the approach should also apply to a video stream without direct skin contact with the lens. Skin contact is only required in the first place as a means of noise rejection by filling up entire camera view with the region of interest. If skin region can be identified and properly tracked throughout all frames, even a non-contact video will contain such biometric data. Furthermore, even if the boundary can be properly extrapolated from the video, the region of interest will encompass a fraction of the total pixel count of the lens and will contain a significant amount of noise from the environment. The challenge lies within amplifying the frequencies corresponding to the potential heartbeat signal and rejecting the frequencies introduced by noise within the video.

0.3 Related Work

While video-based heart rate detection systems are not yet widely implemented, the concept is not new. In November 2012, a collective team of MIT and Cambridge researchers demonstrated the capability to amplify sub-threshold frequencies within a video using a technique they called Eularian Video Magnification[WRS$^+$12]. The technique decomposes the video into the spacial domain and applies a temporal bandpass filter to the frames. The filter provides amplification of a specified range of frequencies within the spatio-temporal domain. Their work was used as one of the main foundations for the implementation in this paper. While the demonstrated method can extract imperceptible frequencies from a video, this project expands on the idea by first extracting only the face to be amplified and then comparing to the amplification of non-facial region within the same video. Furthermore, this paper focuses on amplifying a wide frequency range encompassing resting and active heart rate, in comparison to the relatively narrow-band filters specifically selected for each implementation by Wu et al.

In addition to amplifying color changes, a similar approach can be applied to movement amplification. In this case, the goal was to reduce small motion across the video frames. While the Eularian Video Magnification technique works in this case, a more efficient solution to the problem was proven using a Phase-Based approach. Another team at MIT Computer Science and Artificial Intelligence Lab introduced a method of achieving this through analysis of phase variations of the coefficients of a complex-valued steerable image pyramid over time, corresponding to motion[WRDF13]. Their work was used within an optional intermediary step of the overall process and provided a more stable and clean final output image. However, outside of the visual results, the extra step provided little improvement in heart rate prediction. Since this step incurs a significant computation cost, the option to skip such movement attenuation altogether is provided within the code.

0.4 Method

0.4.1 Processing Stage 1 - Facial Boundary Recognition and Tracking

This stage of processing takes in the raw video directly from recording device, and produces an output of just the face adjusted for rotation and distance from camera. In order to achieve this, the Viola-Jones Object Detection Framework is applied to the first frame[Wan14]. All potential faces are identified via detecting prominent features and if more than 1 face is identified the user is prompted to pick one of their choice. An example of the feature set and facial boundary is provided in figure 2. Once a face is selected, the Kanade-Lucas-Tomasi(KLT) algorithm is applied to track the face over the entire video based on the eigenvectors of the selected face within the first frame[Bir07]. This produces a boundary region of the face for each frame, the coordinates are stored in an array for generating a new video. The coordinates from KLT execution scale with head rotation and adapt to size differences that result from changes in distance to the camera. In order to compile a final output video, the image within the coordinate box in each frame must be cropped, rotated, and re-sized. The rotation relied on simple geometry to determine angle between the selected region and horizontal with respect to the image. Cropping, however required a more advanced approach when the image was rotated. One approach would rely on extracting all the pixels within the box and applying a transformation, but this would produce jagged edges without additional methods for compensation. Instead, the implemented method first computes a region for initial crop which perfectly encompasses the rotated region. Figure 3 illustrates the crop region in red, and facial boundary in yellow. The outer coordinates of the face ROI lie on exactly on the lines of the selected box for cropping. After performing the first crop, the new image is rotated, and finally cropped to eliminate the dead regions.

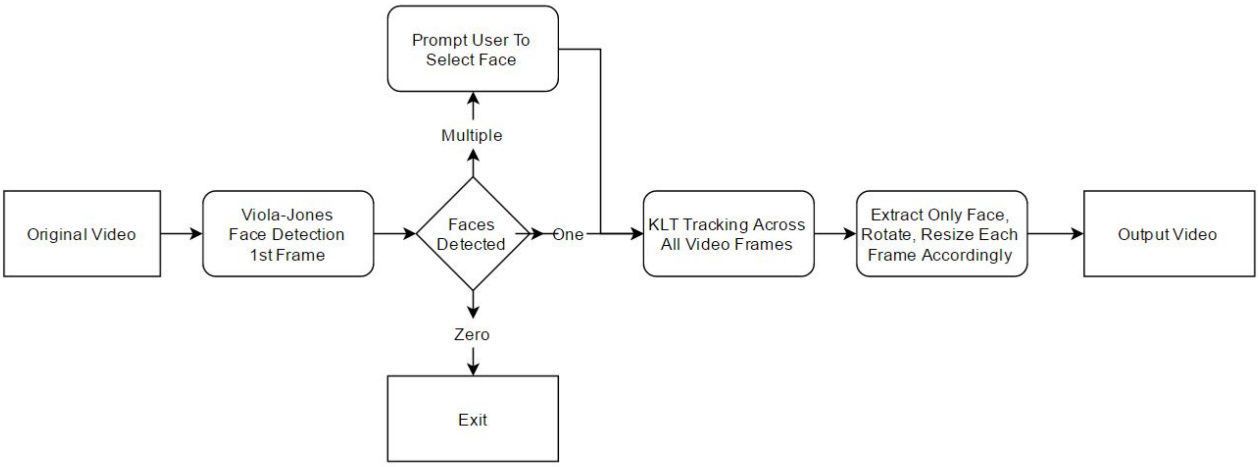
Prompt User To Select Face
Multiple
Original Video
Viola-Jones Face Detection 1st Frame
Faces Detected
One
KLT Tracking Across All Video Frames
Extract Only Face, Rotate, Resize Each Frame Accordingly
Output Video
Zero
Exit

Figure 1: High level flow chart of the system

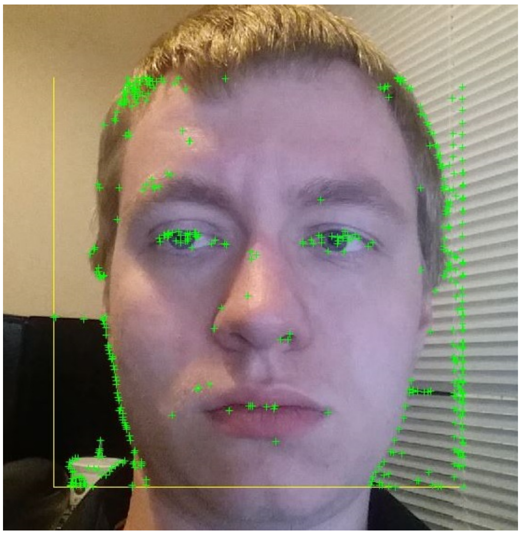

Figure 2: Feature extraction and face recognition

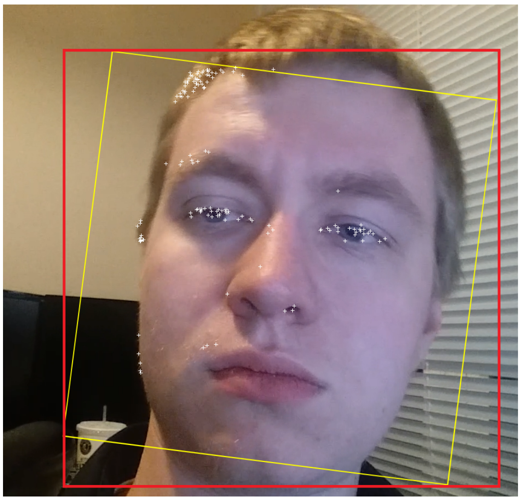

Figure 3: Initial crop parameters for rotated ROI

0.4.2 Processing Stage2 - Phase Based Video Motion Processing

This stage of processing was initially explored as a means of generating nicer visual output of the heartbeat over the face extracted from the video. By suppressing small vibrations, the further processing stages produce output files with less visible jitter and nicer color transitions. The work done by Wadhwa et al. was used as the main framework for implementation[WRDF13]. The output from stage 1 in fed to into a phase-based motion attenuator to produce an output video which is stationary relative to the face. The process works by analyzing local phase signals over time within different spacial scales and orientations. An example of the approach is provided in figure 4. Complex-numbered steerable pyramids are used to decompose the video stream and separate the amplitude from the phase for the local wavelets. Temporal filtering is applied to the phases independently with an optional stage of spatial smoothing. Once the phases are bandpassed in the temporal domain, they are attenuated and the video is reconstructed.[WRDF13].

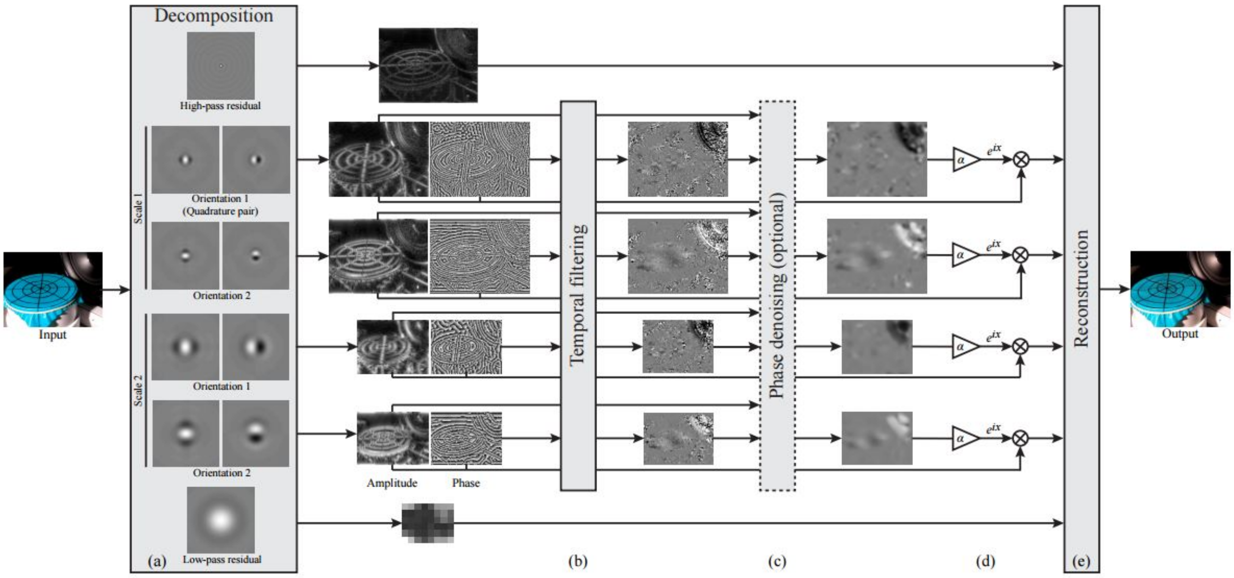

Figure 4: Phase-Based approach to movement attenuation within video

0.4.3 Processing Stage3 - Spatio-Temporal Frequency Amplification

Within this stage of processing, the cropped, adjusted, and optionally stabilized image undergoes frequency amplification. The basic concept relates to the way a bandpass filter may be used within signal processing of a standard electromagnetic wave to isolate specific frequencies within a signal. A Fourier transform can be applied to convert such a signal from the time domain into its frequency domain representation. A bandpass filter is defined by the low and high cutoff values and allows for selecting the portion of the signal within those frequencies as illustrated in figure 5. Unlike an electromagnetic wave, which can only vary in amplitude within one dimension, individual video frames are composed of two-dimensional arrays of pixels. As a result, a two-dimensional discrete Fourier transform must be applied to convert a single frame into the spacial domain. The frequency representation of the current frame varies with time as the video progresses. The basic idea of Eularian Magnification approach is to decompose the video into different spatial frequency bands and then apply temporal filtering to each of them. The filtered bands are now left with primarily signals within the frequency range defined by the filter and each is amplified by an amplification factor α. The amplified signals are added on top of their original forms and then converted from individual spatial representations back into a standard video [WRS+12]. An overview of the process is illustrated in figure 6. The difficulty within this stage lies within selecting optimal cutoff values for the filter, setting the amplification factor α, and picking a good number of spatial levels to decompose the video into.

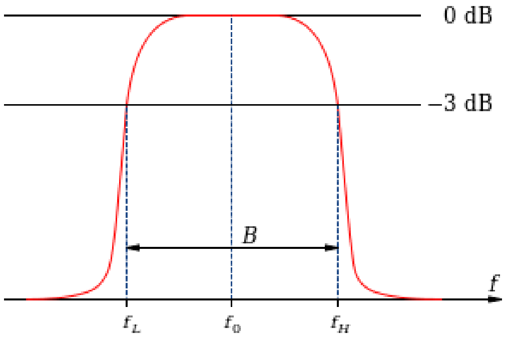

Figure 5: Bandpass filter example diagram

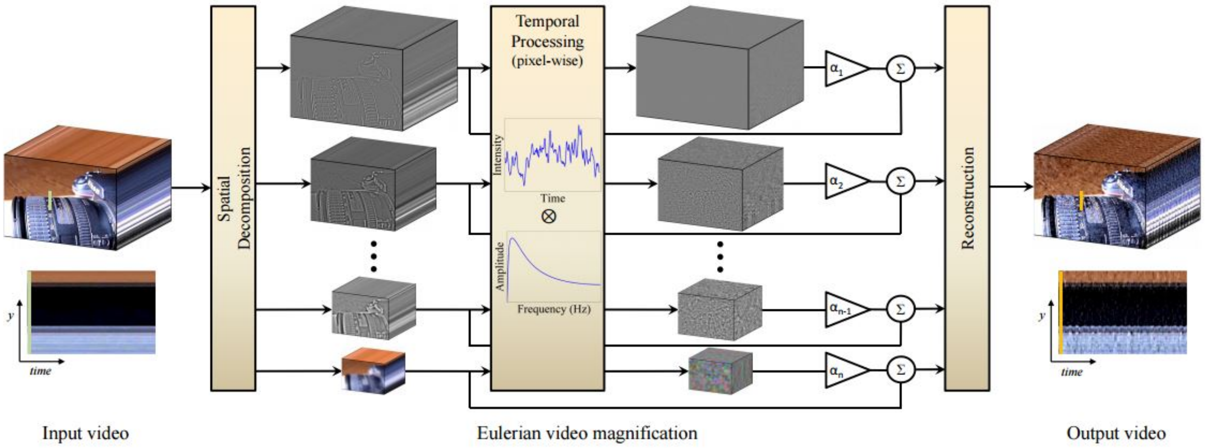

Figure 6: Eularian Magnification Framework Overview

0.4.4 Processing Stage4 - Automated Heart Rate Calculation

Once the amplification is performed and the results are collapsed and written out to a new video, the last step of the process lies within determining the likely heart rate automatically. The video from the last stage will generally exhibit a countable pattern of color variations which can be analyzed by hand to extrapolate an approximation of heart rate. In order to automate this step, several techniques were tested. The average value of all pixels within a single frame was calculated for each of the RGB channels. The progression was plotted within the temporal domain for analysis and it was determined that an overall average of all three color pixel values for each frame would serve as a good representative measure. Future work could focus on exploring various other methods including selecting a specific color, or nonlinear combination. The initial approach to counting heart rate from the sinusoidal color progression focused on identifying average peak-to-peak and trough-to-trough distances and adjusting by the video's frame rate to determine a heart rate value. As later discovered, this approach failed to show the entire picture and only worked well only for near perfectly sinusoidal signals. As an updated approach, the discrete Fourier transform was applied within the user specified frequency range and the highest peak was selected as the heart rate approximation.

0.4.5 Baseline Adjustment - Attempt to Remove Non-heartbeat Frequencies in Presence of Significant Noise

A final addition to the process was created to further mitigate noise frequencies often introduces into the video. Under low lighting conditions especially, the skin color fluctuations corresponding to heart rate are often significantly weaker than general noise, especially when a very wide bandwidth bandpass filter is used in stage 3. An experimental solution to mitigating this issue and providing potential of isolating frequencies corresponding to heart rate from the noise focuses on re-running the four stage process again. This time user input is requested to determine a boundary without any faces or body parts to be used as a baseline.

0.5 Implementation Details

The implementation breaks down into a number of custom functions written to support the various stages of process execution. The supporting function for Eularian Magnification as well as Phase-Based processing are isolated in individual folders: '/AmplificationSupport' and '/PhaseBasedSupport'. The call-graph for the full system is provided in figure 7. The program can be run from 'main.m' after configuring several parameters. The default folder for new videos to process is '/input' and the processed files from each stage of execution are written to '/output'. The filename must be set including extension. The parameter 'mode' is used to enable('show') and disable('hide') figures detailing the face tracking and extraction process. The 'stages' array defines which stages of execution to enable for the given run. This approach allows for individual stages, such as the optional Phase-Based attenuation to be disabled. Furthermore, once an output for a given stage is generated, it is stored to disk. Hence for testing different filter parameters in stage 3, the first 2 stages can be set to '0' saving much valuable time. The 'baselineCheck' parameter was one of the latest additions and allows for runninng the Baseline Adjustment stage. This portion of code is still experimental but can produce some rather interesting results. Finally, the amplification filter parameters must be set for use within stage 3. The amplification factor 'alpha' defines how much to bring out the frequencies of interest. The number of spacial levels within the decomposition structure for Eularian Magnification is set in 'level'. Typically a value between 4-6 delivers good results. The most important parameters to set are the bandpass filter low frequency cutoff 'fl' and the high frequency cutoff 'fh'. These values represent the frequency in herz, thus are stored as BPM/60 for simple comprehension. With the addition of 'baselineCheck', ranges as wide as 40bpm to 200bpm have demonstrated promising results with the noise frequency cancellation features.

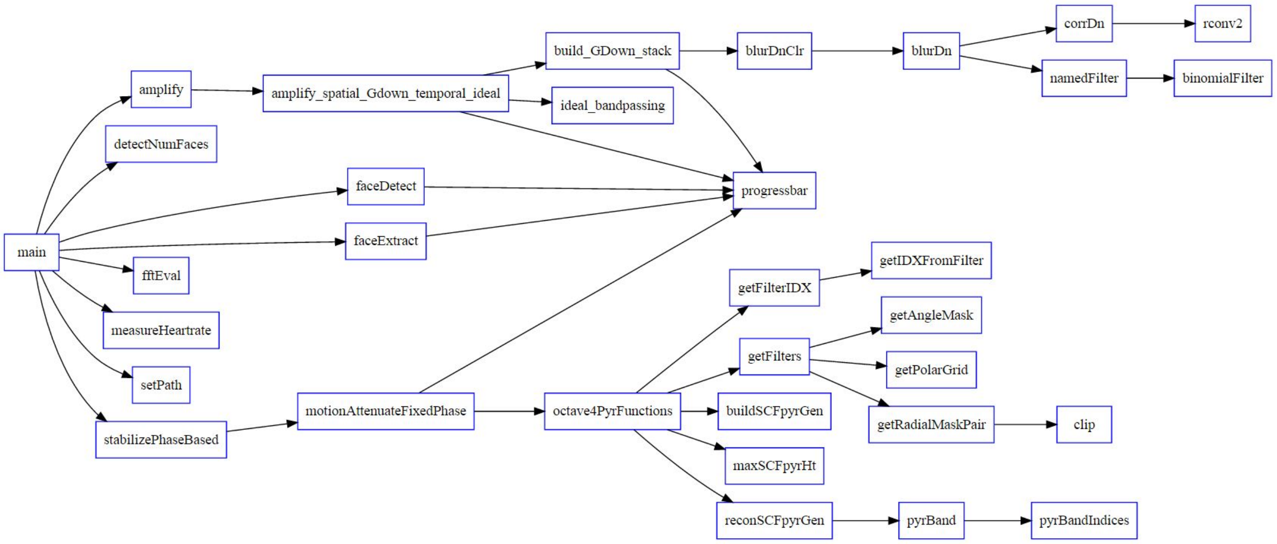

Figure 7: Matlab callgraph of the full system

0.6 Dataset Used

The testing dataset consisted of numerous videos recorded under different conditions. Generally a video would be recorder, followed up by a short workout to increase heart rate for the next video. Various length videos were used ranging from 4 seconds to half a minute.

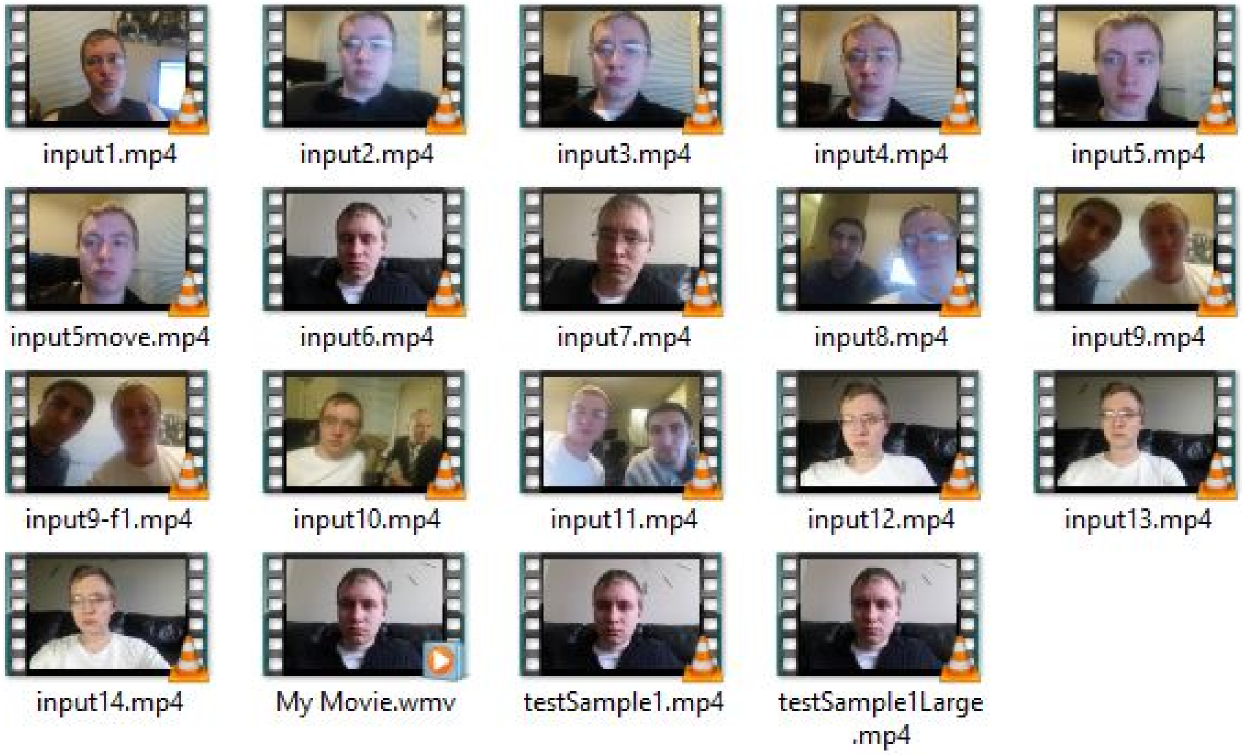

Figure 8: Dataset used for system testing

0.7 Results

The results from two closely related test videos are provided within this section. Although narrow filters provide more accurate results, these tests were run under the following conditions:

- stages=[1 0 1 1]
- baselineCheck=1
- alpha=150
- level=4
- fl=40/60
- fh=200/60

Effectively, the entire range of possible resting and active heart rate frequencies, from 40-200 beats/minute, was amplified by 150 times. Due to lack of access to professional equipment, the heart rate calibration was performed using a touch-based censor on a smartphone. The video used for results below had a control heart rate of 81bpm.

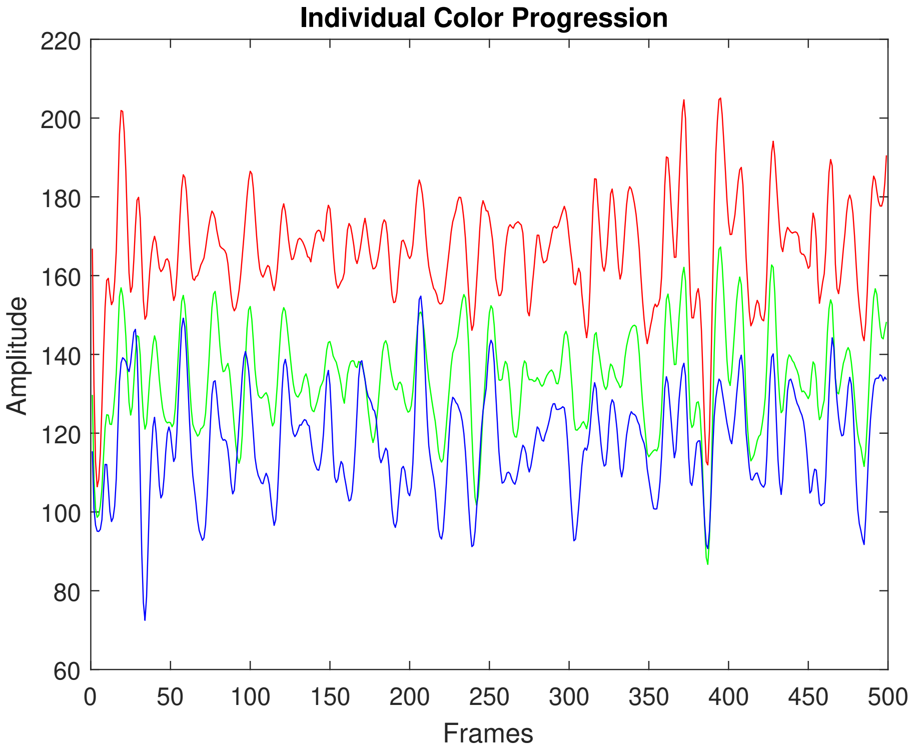

Figure 9: Individual color progression through frames

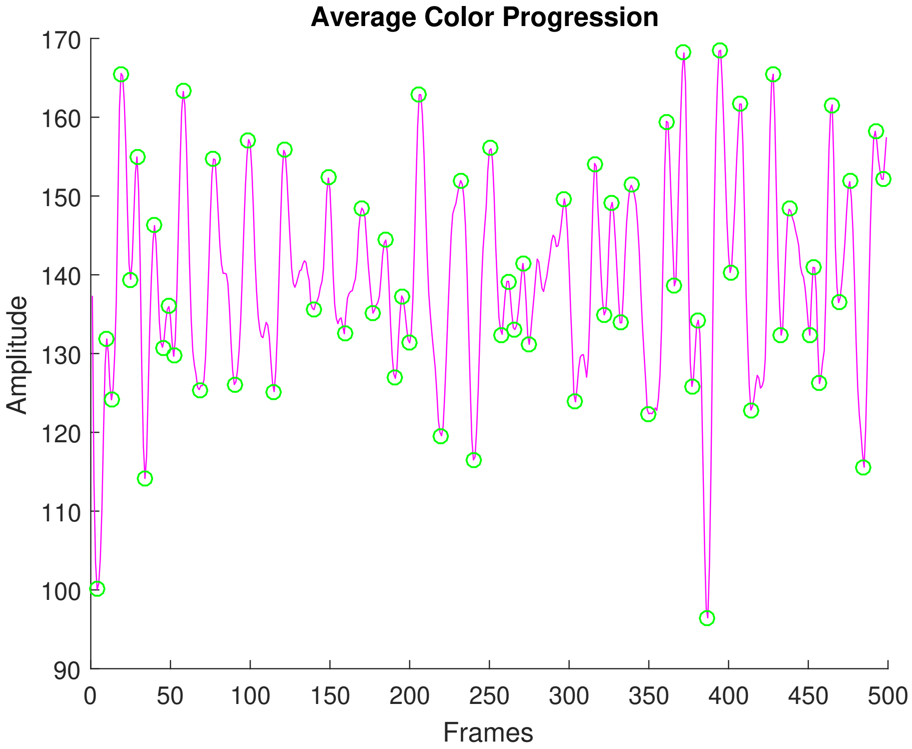

Figure 10: RGB average progression through frames

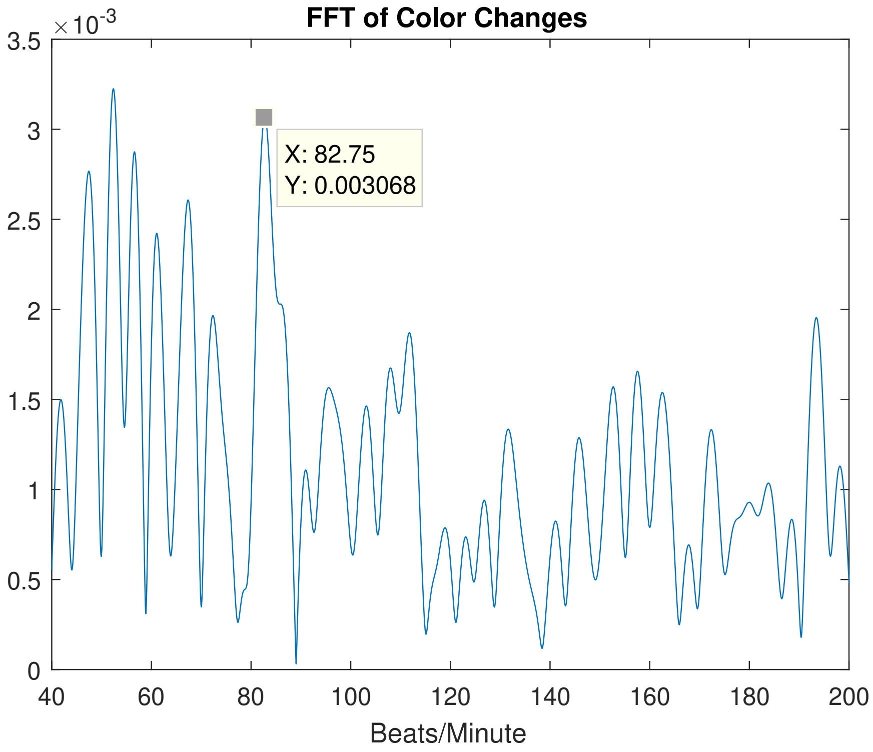

Figure 11: RGB average represented in frequency domain

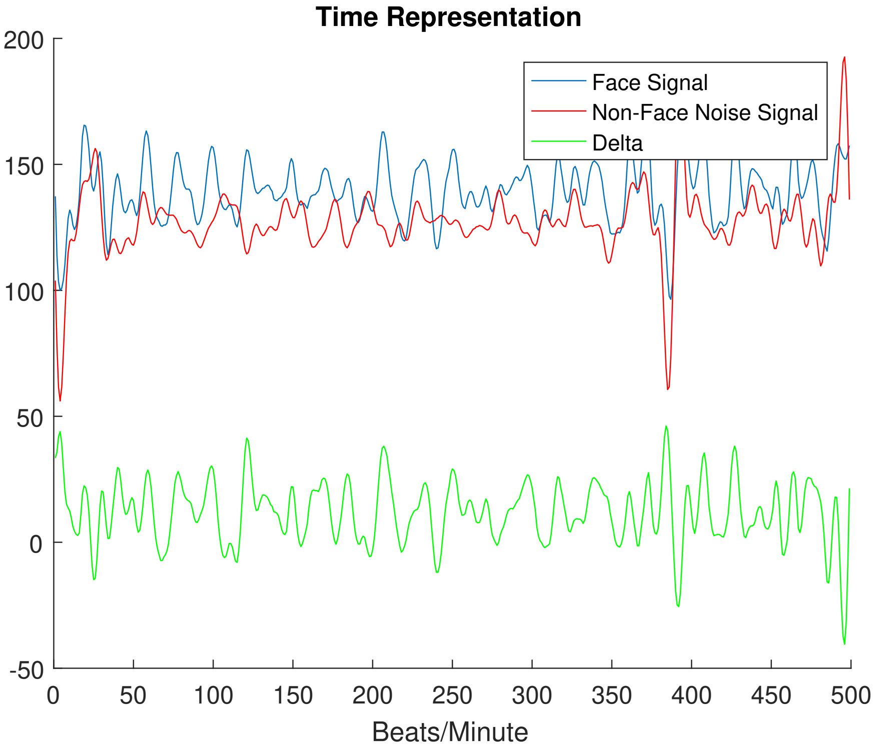

Figure 12: Time representation of face, non-face region, and the difference

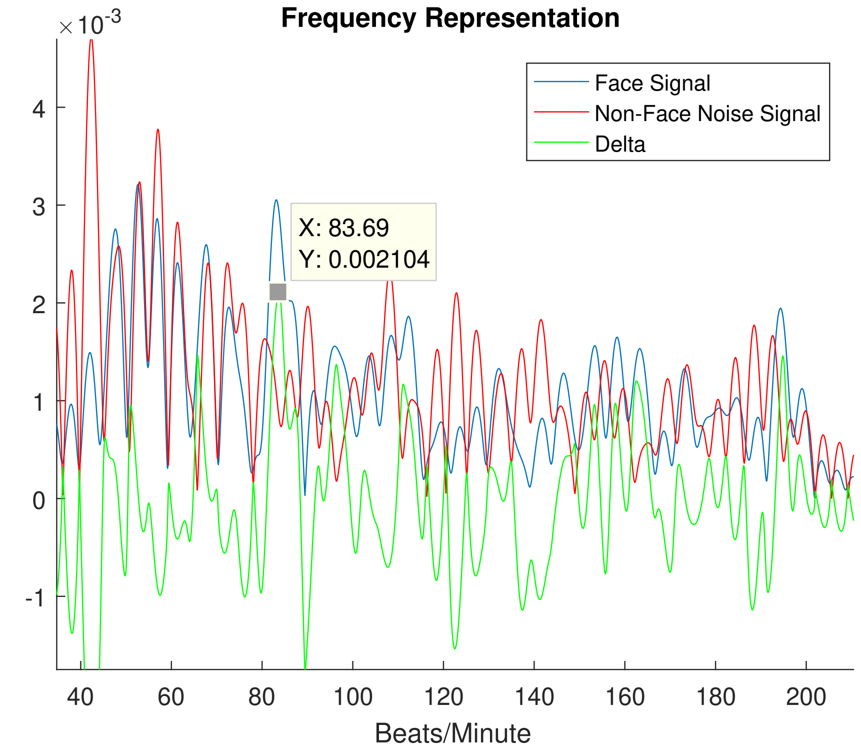

Figure 13: Frequencies of face, non-face region, and the difference

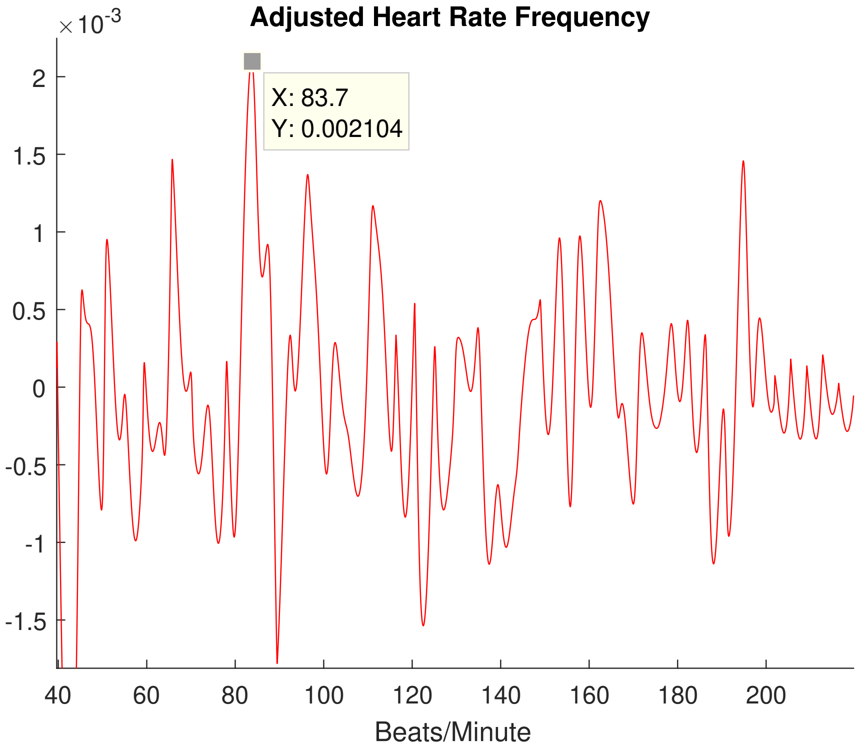

Figure 14: Heart rate FFT adjusted for noise

The next test video was recorded shortly after under similar conditions. An intense workout raised the control heart rate to 122bpm using touch-based sensor. The face detection program produced the following results:

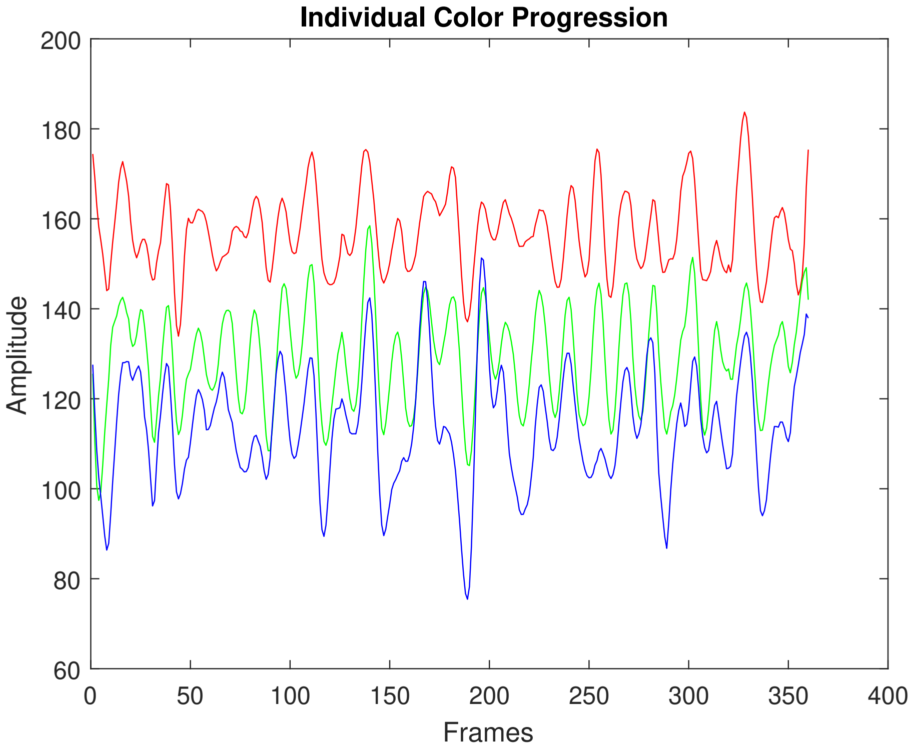

Figure 15: Individual color progression through frames

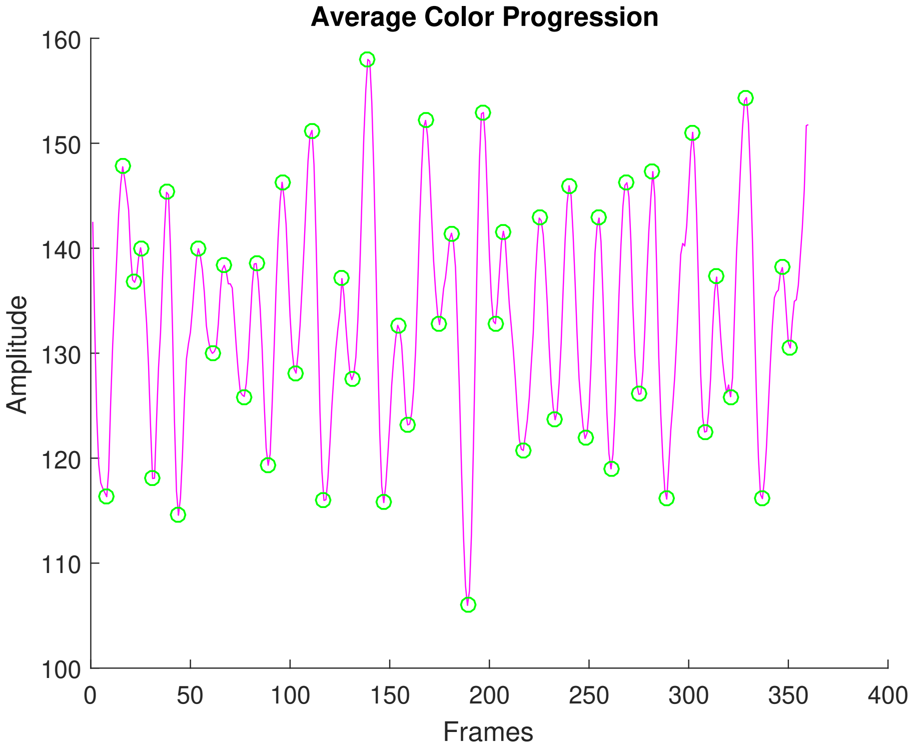

Figure 16: RGB average progression through frames

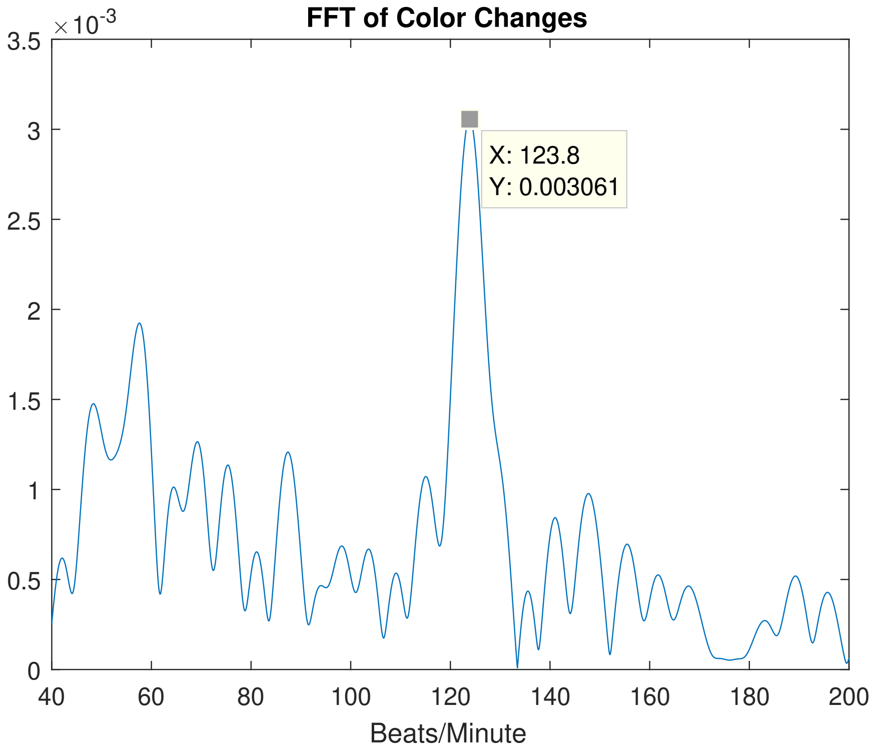

Figure 17: RGB average represented in frequency domain

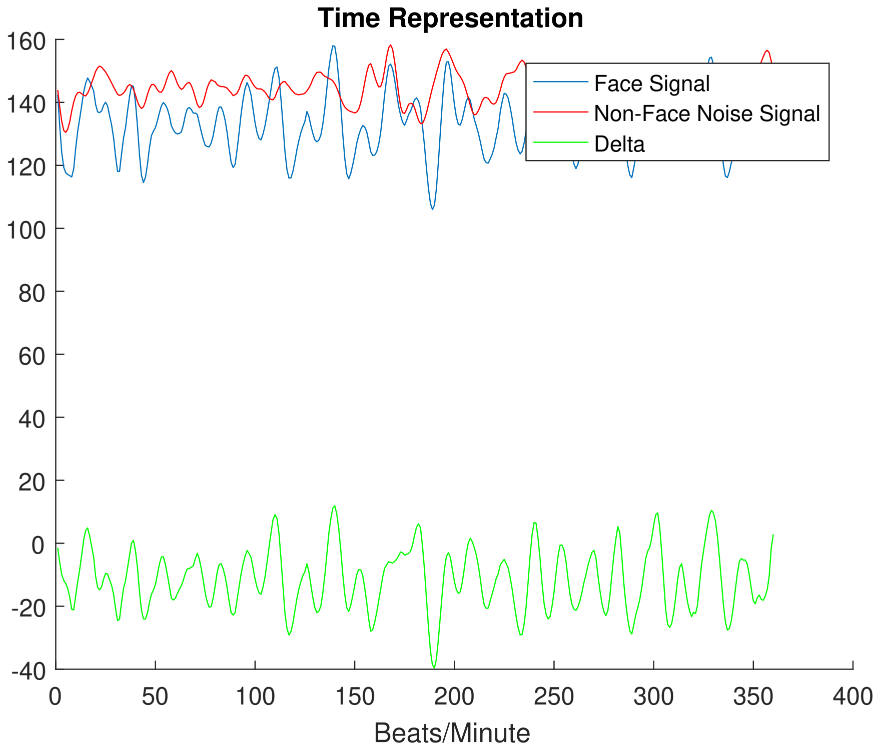

Figure 18: Time representation of face, non-face region, and the difference

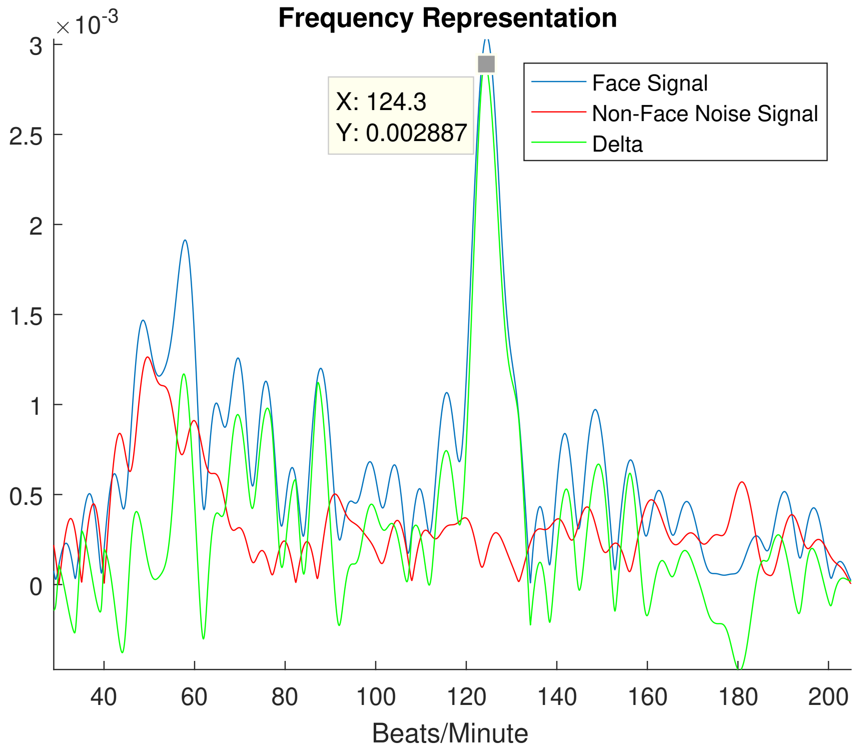

Figure 19: Frequencies of face, non-face region, and the difference

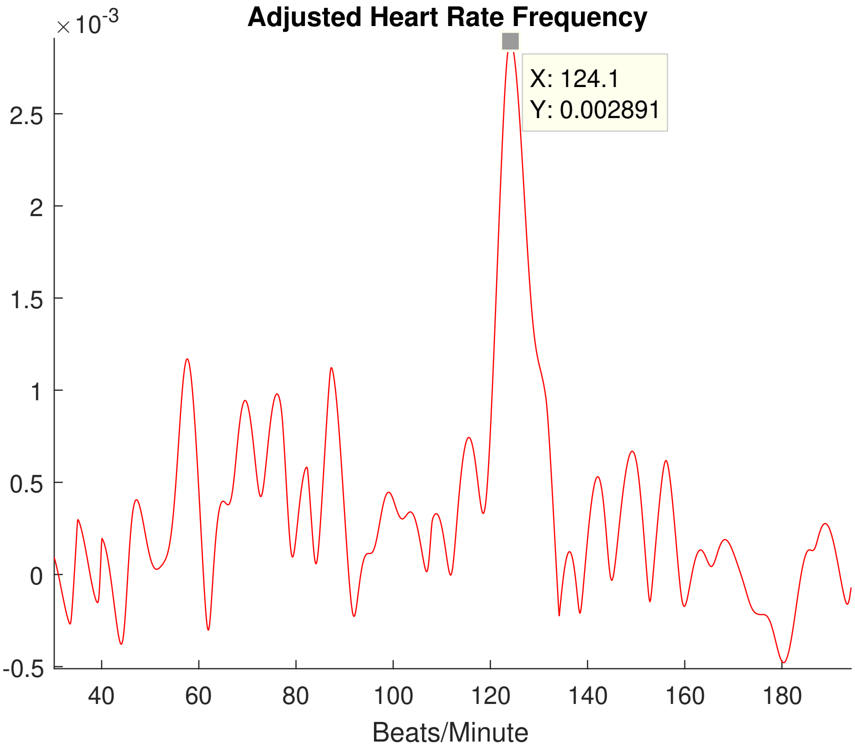

Figure 20: Heart rate FFT adjusted for noise

0.8 Discussion on the Results

Although the results did not match exactly with touch-based measurement devices, the values extrapolated from just the face were relatively close. In the case of the test videos presented, the prediction was slightly higher than the measurement. Furthermore, with the first test, the correct heart beat frequency was only identified after canceling for noise. As can be seen in figure 11, the highest point on the frequency plot is actually around 55 bpm. By exploring figure 13, it can be seen that the 55bpm frequency appears strongly within the noise data represented by red. The real data, represented by blue only differs significantly from the red near the frequency 83.7bpm. The difference, represented in green, is also plotted individually in figure 14 illustrating this real value for hear rate. In the case of the second video, the initial prediction was quite accurate in figure 17 and further backed by noise cancellation technique producing figure 20. It is important to note that within much more narrow filters, close results were often achieved without the requirement for noise cancellation technique. Nonetheless, using a wide-band filter provides a massive benefit of minimized requirements in terms of initial input which aligns more with the final goal of fully automated measurement. Among numerous other trials performed, the system was able to accurately extract heart rate for high BPM values disproportionately more than for low measurements. A possible explanation is the strength of the signal as blood pulsates with more intensity at higher heart rate. Furthermore, direct lighting of the face was identified to have a large impact on system performance. Additional testing is required under much more strictly controlled environments and medically accurate data for system validation and true comparison of various techniques.

0.9 Conclusion and Future Directions

The initial goal of the project set out to determine the feasibility of extracting heart rate from a video simply containing a face. Although limited time prevented fully testing the system under extreme conditions such as movement, face tilt, and lighting, the results obtained from the test cases provided a high degree of optimism in a future full-proof implementation. A significant among of new knowledge was acquired throughout the various phases of research, development, and testing of the project. The identification of bandpass filter parameters, as well as final interpretation and noise cancellation of the signal proved the most challenging. Many initial tangents to determine the ideal approach were tested and further research is necessary within this direction. A potential solution could involve scanning with a high bandwidth filter first, then narrowing the frequency range for more accurate results. The long-term vision represents a small embedded device with built in camera and onboard real-time processing for heart rate metrics. Such a device would compute real-time heart rate data for any target within its field of view, hence providing valuable photoplethysmographic data for further analysis.

Bibliography

[Bir07] S. Birchfield. Klt: An implementation of the Kanade-Lucas-Tomasi feature tracker. http://www.ces.clemson.edu/stb/klt/, 2007.

[Mel13] Ignacio Mellado. Measuring heart rate with a smartphone camera, September 2013.

[Wan14] Yi-Qing Wang. An Analysis of the Viola-Jones Face Detection Algorithm. *Image Processing On Line*, 4:128–148, 2014.

[WRDF13] Neal Wadhwa, Michael Rubinstein, Frédo Durand, and William T. Freeman. Phase-based video motion processing. *ACM Trans. Graph.*, 32(4):80:1–80:10, July 2013.

[WRS$^+$12] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. Eulerian video magnification for revealing subtle changes in the world. *ACM Trans. Graph.*, 31(4):65:1–65:8, July 2012.