Report 17-UT-058

NSTSCE

National Surface Transportation
Safety Center for Excellence

Video Magnification to Detect Heart Rate for Drivers

Abhijit Sarkar, Zachary Doerzaph, A. Lynn Abbott

Submitted: November 13, 2017

Technology
Impairment
Infrastructure
Mobility

Housed at the Virginia Tech Transportation Institute
3500 Transportation Research Plaza • Blacksburg, Virginia 24061

ACKNOWLEDGMENTS

The authors of this report would like to acknowledge the support of the stakeholders of the National Surface Transportation Safety Center for Excellence (NSTSCE): Tom Dingus from the Virginia Tech Transportation Institute, John Capp from General Motors Corporation, Chris Hayes from Travelers Insurance, Martin Walker from the Federal Motor Carrier Safety Administration, and Cathy McGhee from the Virginia Department of Transportation and the Virginia Transportation Research Council. The authors would also like to thank Dr. Natalia Henao-Guerrero and her students from the Virginia-Maryland College of Veterinary Medicine, Dr. Devi Parikh from the Bradley Department of Electrical and Computer Engineering at Virginia Tech, and the following individuals who provided assistance with research conducted at the Virginia Tech Transportation Institute: Dr. Jon Hankey, Kayla Sykes, Rebekah Duke, Kelly McGowan, and Dr. Suzie Lee.

The NSTSCE stakeholders have jointly funded this research for the purpose of developing and disseminating advanced transportation safety techniques and innovations.

EXECUTIVE SUMMARY

Heart rate is a strong indicator of a person's psychophysiological state. For this reason, many applications would benefit from the ability to measure noncontact heart rate. The present work describes a new procedure for estimating blood volume pulse from video of a person's face, with an emphasis on real-life scenarios like driving. The approach builds on the algorithm known as Eulerian VidMag, which has shown promise under laboratory conditions, but exhibits problems when applied in naturalistic situations. In particular, problems arise due to movement by the subject, changing illumination conditions, and low-frame-rate video. This work describes methods developed to address some of these problems, including working with video rates down to 10 frames per second. The methods were tested using videos of indoor subjects, as well as videos of drivers in naturalistic situations. We assessed the method through analysis of different stress levels using the extracted heart rate information for a driver on the road by comparing heart rate variability at different stress levels. Experiments showed that a systematic post-processing strategy can improve the accuracy of the VidMag algorithm's raw output and achieve a good correlation in both instantaneous heart rate and average heart rate with the ground truth heart rate measurements. This, in particular, improves with higher quality video sources and more controlled experimental conditions. However, the robust broad application of the methods on existing lower-quality naturalistic video data remains a challenge.

TABLE OF CONTENTS

LIST OF FIGURES iv

LIST OF TABLES vii

LIST OF ABBREVIATIONS AND SYMBOLS ix

CHAPTER 1. INTRODUCTION 1

NON-INTRUSIVE MEASUREMENT OF HEART RATE 2
EULERIAN VIDEO MAGNIFICATION (VIDMAG) 3

CHAPTER 2. IMPLEMENTATION AND TESTING OF VIDMAG 5

FLOW OF ALGORITHM 6
PREPROCESSING: FACE DETECTION AND COMPENSATION FOR NATURAL MOVEMENT OF THE HEAD 7
Face Detection Using the Histogram of Oriented Gradient (HOG) Method 7
Intra-face Face Detection Methods 10
Face Registration 12
SKIN DETECTION AND IMPROVED BEAT ACCURACY 14
Patch Selection 14
Skin Detection 15
VIDEO MAGNIFICATION: PEAK DETECTION AND RESAMPLING 21
HEART RATE VARIABILITY (HRV) 22
Overview of HRV 23
Data Collection for HRV Verification 23
POST PROCESSING OF MAGNIFIED VIDEO 24
Pulse Signal Extraction 24
Savitzky-Golay (S-G) Filter Analysis 25
Lomb-Scargle Periodogram 26
Error Analysis 28

CHAPTER 3. DATA COLLECTION 31

PHASE I – STATIC CONDITIONS 31
PHASE II – DYNAMIC CONDITIONS 31
RAW-DATA ANALYSIS 32
EXPERIMENTS AND RESULTS 33
HEART RATE VARIABILITY (HRV) ANALYSIS 35

CHAPTER 4. DISCUSSION – CHALLENGES AND CONSIDERATIONS FOR FUTURE WORKS 39

HRV ANALYSIS FROM BVP SIGNAL: AN ANALYSIS 39
FACIAL MAP FOR HEART RATE MEASUREMENTS 40
ILLUMINATION 40
Challenges 40
Use of Vehicle's Stationary Parts for Illumination Correction 41
Use of Lambertian Surface 42
PROBLEMS RELATED TO NATURALISTIC DRIVING VIDEOS AND SHRP 2 43
AUTHENTIC DATA FOR BEHAVIOR ANALYSIS 44
RECENT DEVELOPMENTS IN THE FIELD OF REMOTE MEASUREMENTS OF HEART RATE 44

APPENDIX A. AVAILABLE FACE DETECTORS 47

APPENDIX B. CRASH – SHRP 2 49

REFERENCES 51

LIST OF FIGURES

Figure 1. Previous work: (a) Pulse rate sensor embedded in steering wheel (Lin et al., 2007). (b) PPG sensor in the ear (Poh et al., 2010) (c) Pulse measurement sensor installed on the seat by Ford. (d) pulse rate estimation from subtle head movement, (e) Pulse rate using blind source separation (Poh, McDuff, & Picard, 2011), (f) Video magnification for pulse rate measurement (Wu, 2012). 2

Figure 2. Diagram. Eulerian video magnification (VidMag) techniques – flow of work (Wu et al., 2012). 4

Figure 3. Diagram. VidMag application workflow. 5

Figure 4. Images. Optical flow tracking 8

Figure 5. Graph. Improved time efficiency. 9

Figure 6. Graph. Localization accuracy. 10

Figure 7. Screenshots. Results from intra-face: (a) daytime, (b) different pose, (c) nighttime infrared frame selection. 11

Figure 8. Screenshot. Face detection with 49 fiducial points 12

Figure 9. Screenshots. Selection of frame sequence based on the movement of fiducial points. 12

Figure 10. Screenshots. Face registration. a & b) Two randomly chosen faces with automatically detected landmark points. Superimposed frames: c) before registration, d) after registration. 13

Figure 11. Screenshots. Frame registration showing improvement in consecutive frames. The original scatter sequence shows how fiducial points move with time. After homography, the movement of fiducial points has been minimized. 13

Figure 12. Diagram. Skin detection and heart rate estimation 14

Figure 13. Images. Sample images from the SFA dataset: This image set comprises face images with different level of pigmentation, gender poses and variation of facial hair, use of glasses, and different hairstyles. 16

Figure 14. Images. Sample images with skin ground truth from the ECU dataset. Images from this dataset were taken with different indoor and outdoor backgrounds and more challenging situations 17

Figure 16. Diagram. The skin detection algorithm. 17

Figure 17. Graph. Precision and recall curve for threshold selection 18

Figure 18. Diagram. Heart-rate extraction flowchart of the new system. The VidMag algorithm is used to extract a raw signal that is analyzed further for noise reduction, interpolation to address low frame rate, peak detection, and heart rate extraction. 22

Figure 19. Graph. Raw output of BVP. 22

Figure 20. Graph. R-R interval. A sample ECG signal from the MIT-BIH arrhythmia dataset Goldberger et al., 2000). 23

Figure 21. Photos. Experimental setup for data collection phase III. (a) PPG pulse meter is worn on the fingertip and one of the three leads of the ECG measurement kit is attached to the participant’s left wrist. (b) The S5 machine with the data recording software. 24

Figure 22. Screenshot. Manual selection of skin area of interest. 24

Figure 23. Graph. Raw heart rate estimation from VidMag. 25

Figure 24. Graph. Heart rate after filtering. 26

Figure 25. Graphs. Comparison for average of heart rate. The heart rate has been averaged for every 5 seconds. 26

Figure 26. Graph. Comparison of instantaneous heart rate in frequency domain. The power spectral density is computed using the Lomb-Scargle periodogram. 28

Figure 27. Graph. Comparison of VidMag algorithm by spectrogram analysis. (a) shows the spectrogram from the pulsemeter and (b) shows the results from the VidMag algorithm. 28

Figure 28. Graph. Comparison of instantaneous heart rate. 29

Figure 29. Graphs. Error analysis of heart rate. 29

Figure 30. Photo. Garmin wrist watch and HRM. (Image source: http://www.garmin.com/en-US). 31

Figure 31. Screenshots. Example of collected video 32

Figure 32. Screenshots. Sample screens from the collected video data using MiniDAS. 32

Figure 33. Graph. Sample run using Garmin HRM: Participant 1 – (A1 and A2 – inside city, B – inside city secondary task and extra cognitive load, C – interstate driving – less cognitive load) 33

Figure 34. Graph. Sample run using Garmin HRM: Participant 2 – an instance of panic during hard braking. 33

Figure 35. Graph. Heart rate estimation using the new procedure for a subject indoors. 34

Figure 36. Graph. Heart rate estimation using the new procedure for video obtained while the subject was driving. 34

Figure 37. Graph. Heart rate estimation using the new procedure. The subject was performing a secondary task while driving in city conditions. This corresponds to segment B in Figure 33. 35

Figure 38. Graph. Frequency domain analysis of segment C. 36

Figure 39. Graph. Frequency domain analysis of segment B. 37

Figure 40. Diagram. Illumination gradient on the face. 41

Figure 41. Screenshot. Stationary part of car interior. 41

Figure 42. Graph. Interior illumination variation with time. 42

Figure 43. Screenshot. Face detection problem in SHRP 2 video. 44

LIST OF TABLES

Table 1. Face detection accuracy with (top row) and without (bottom row) temporal information. 10

Table 2. Example of successful cases of skin detection from the SFA database. 20

Table 3. Failure cases from the SFA dataset – Top-left and top-right: hair is confused as skin, top-middle: part of shirt has been selected, bottom-left: illumination and background confusion, bottom-middle: effect of wrong-face detection, bottom-right: inadequate grayscale sampling from mask 21

Table 4. Algorithm 1 – Select narrow bandpass frequencies. 25

Table 5. Time domain analysis of HRV for segment C. 36

Table 6. Time domain analysis of HRV for Segment B. 36

LIST OF ABBREVIATIONS AND SYMBOLS

bpm	beats per minute
BSS	blind source separation
BVP	blood volume pulse
ECG	electrocardiogram
FFT	fast Fourier transform
fps	frames per second
GPS	Global Positioning System
HF	high frequency
HOG	histogram of oriented gradient
HRM	heart rate monitor
HRV	heart rate variability
Hz	Hertz
LF	low frequency
MRF	Markov random field
NDS	naturalistic driving study
NHTSA	National Highway Traffic Safety Administration
PPG	photoplethysmography
RGB	red, green, blue
S-G	Savitzky-Golay
SHRP 2	Second Strategic Highway Research Program
VidMag	Video Magnification
VLF	very low frequency
VTTI	Virginia Tech Transportation Institute

CHAPTER 1. INTRODUCTION

Driver behavior plays a pivotal role in crash and near-crash scenarios. Factors such as the driver's cognitive workload, stress level, and inattention, including drowsiness, fatigue, driving under the influence, and on- and off-road distractions, can all be contributing factors leading to a crash or near-crash. A report from the National Highway Traffic Safety Administration (NHTSA) states that 35,092 people lost their lives and more than 2.44 million people were injured as a result of road accidents in 2015 (NHTSA, 2015). More than 6.3 million crashes, including those involving injuries and property damage, were reported in the same year, resulting in a total economic and comprehensive cost to society of about one trillion dollars. These statistics have led researchers to conduct naturalistic driving studies (NDSs), which are an effective way to understand the role of drivers' performance and behavior as they relate to traffic safety. For example, the Second Strategic Highway Research Program (SHRP 2) data contain numerous variables showing drivers' interactions with both their vehicles and traffic. These were collected by monitoring variables such as speed, acceleration, following distance, maneuver selection, braking patterns, etc. However, one shortcoming of the study is that direct measurement and monitoring of the driver's physiological state was not feasible (a frequent limitation of NDSs). Therefore, visual cues from face videos (e.g., percentage of eye closure, head droop, head movements) and secondary parametric data (e.g., braking pattern, lane deviation, change in heading) are the only methods available to data analysts for assessing a driver's psychophysiological state (e.g., fatigue, anxiety, etc.). Although some of these methods have been proven to be effective, they are often costly in terms of required manual labor and time commitment. Also, all of these methods rely on secondary behavior rather than direct measurement.

Objective measures of physiological variables, like blood pressure, heart rate, and respiration rate may be better indicators of a driver's true psychophysiological state. However, the major constraint for measuring physiological variables is that these objective measurement methods are generally intrusive in nature, often requiring devices with wires and heavy instruments, and interfere with the driver's natural behavior. As such, these methods cannot be implemented in naturalistic scenarios, particularly NDSs. In order to investigate the driver's psychophysiological condition, it is necessary to implement strategies that do not interfere with the natural driving scenario.

However, recent advances in computer vision and signal processing show that SHRP 2 video data may yet have significant potential to indicate drivers' psychophysiological conditions. This research effort was undertaken to evaluate the possibility of using pulse rate as a nonintrusive measurement variable in naturalistic conditions and considers its potential to indicate various psychophysiological driver states, particularly the following: fatigue, cognitive load, panic attack, and influence of alcohol. It has been well established in previous studies that psychophysiological conditions are often reflected in the functioning of the autonomic nervous system (Picard, Vyzas, & Healey, 2001).

NON-INTRUSIVE MEASUREMENT OF HEART RATE

Although a number of in-vehicle efforts involving collecting heart rate data have been reported in the literature, most of them were intrusive in nature. An ideal system for a naturalistic study should be nonintrusive, compact, robust, low cost, low in power consumption, safe, easy to use, and should not affect the driver's behavior. In this section, we will review some of the previous efforts and discuss their benefits and shortcomings.

Every cardiovascular pulse sends a volume of blood through the body. Photoplethysmography (PPG) measures the blood volume pulse (BVP) by passing light through skin. This process is not often convenient in naturalistic scenarios due to its intrusive nature; the subject must wear a plethysmograph device and, in most cases, a fingertip pulse meter. The literature shows that over the last decade, several other methods and measuring instruments have been engineered. Figure 1 shows some of these methods and instruments. Lin, Leng, Yang, and Cai (2007) used a pulse rate sensor embedded in the steering wheel. Poh, Swenson, and Picard (2010) presented an ear-wearable plethysmograph for a motion tolerant pulse measurement. The other notable contributions were from video monitoring—a less intrusive approach.

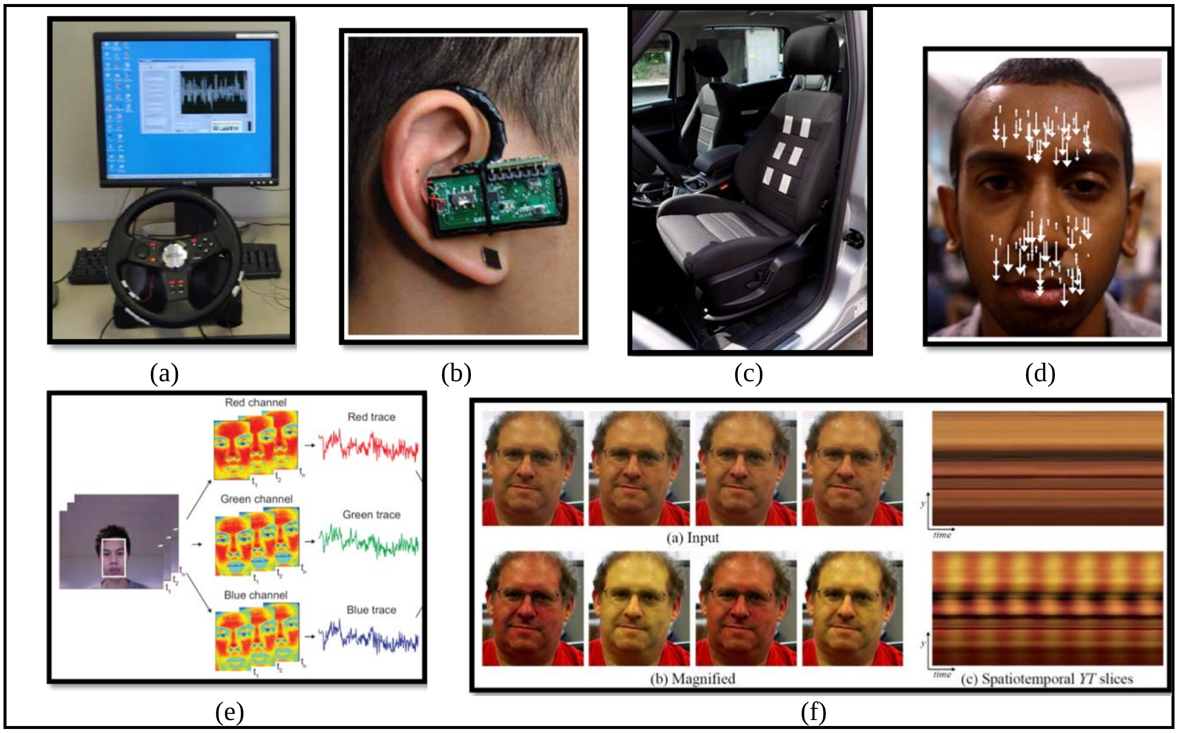


Figure 1. Photos. Previous work: (a) Pulse rate sensor embedded in steering wheel (Lin et al., 2007); (b) PPG sensor in the ear (Poh et al., 2010); (c) Pulse measurement sensor installed on the seat by Ford; (d) Pulse rate estimation from subtle head movement; (e) Pulse rate using blind source separation (Poh, McDuff, & Picard, 2011); (f) Video magnification for pulse rate measurement (Wu, 2012).

Balakrishnan, Durand, and Guttag (2013) reported a novel method to track the motion of the head. With every cardiovascular pulse, the thrust of blood causes a Newtonian reaction in the veins around the neck and head, resulting in a small pulsating movement in the head. This motion is not generally visible, but by carefully tracking points in the head and face, can be measured. The main merit of this method is that it does not depend on ambient illumination; however, the accuracy of the results depends largely on the natural motion of the head.

There are two other video monitoring strategies that were developed to allow direct measurement of BVP from a stream of video data. Poh et al. (2011) used a blind source separation (BSS) method to extract this signal. They assumed that the BVP signal was embedded in all three color channels (red, green, blue [RGB]) in different proportions and mixed with other signals linearly. They used BSS to transform the RGB channels to three independent channels with the assumption that one of the channels had a raw BVP signal. Although this method shows interesting results, there lies an ambiguity in the choice of correct channel in the required independent component analysis (a class of BSS) from video data. Secondly, this method is designed for color video, whereas in many existing naturalistic datasets, such as SHRP 2, face videos are in grayscale only. Poh et al. (2011) used three different color channel coding schemes based on face skin and applied an independent component analysis to extract the pulse signal. McDuff, Gontarek, and Picard (2014a), in their related work, showed that this method returns the best results while using a five color-band camera and 12-bit image depth. Thus, it also requires very high resolution images with a large number of face pixels. This level of high-quality video is not feasible in real-life scenarios or for large-scale affordable naturalistic implementation.

Finally, Wu et al. (2012) used a method of Eulerian video magnification (VidMag) to extract a BVP signal from a video sequence and magnify it to make it visible to normal eyes. We determined that this method was more suitable for naturalistic conditions for several reasons. First, this method traces the changes in pixel intensities, and hence can be used for SHRP 2 grayscale video data as well as future studies employing color video. We tested a video in RGB and in a transformed grayscale version for comparison. In both cases, the estimated blood volume changes were identical. Secondly, the inexpensive SHRP 2 camera sensor induces noise in the video that, in general, interferes with the video processing. VidMag uses a Gaussian pyramid to reduce such noise. The VidMag approach does not suffer from issues of ambiguity and is more tolerant of lower-quality video sources such as SHRP 2 video data. This method, which is the focus of the research project described herein, had not yet been refined for use in the complex real-world driving environment.

EULERIAN VIDEO MAGNIFICATION (VIDMAG)

Based on the findings from a literature review, this work was devoted mainly to the applicability of VidMag in naturalistic scenarios (development, testing, and evaluation). Wu (2012) successfully introduced a VidMag algorithm that combines spatial and temporal processing. The flow of the algorithm is shown in Figure 2. First, each frame is down-sampled using a Gaussian pyramid in order to reduce spatial noise in the frame. Then each pyramid level is temporally filtered using a bandpass filter to extract the frequency of interest. Finally, each level is concatenated with gain factors; these dictate which frequency components will be amplified and which will be attenuated in the final version. Each of these three steps will be further described in the following chapter along with the additional methods developed as part of this effort.

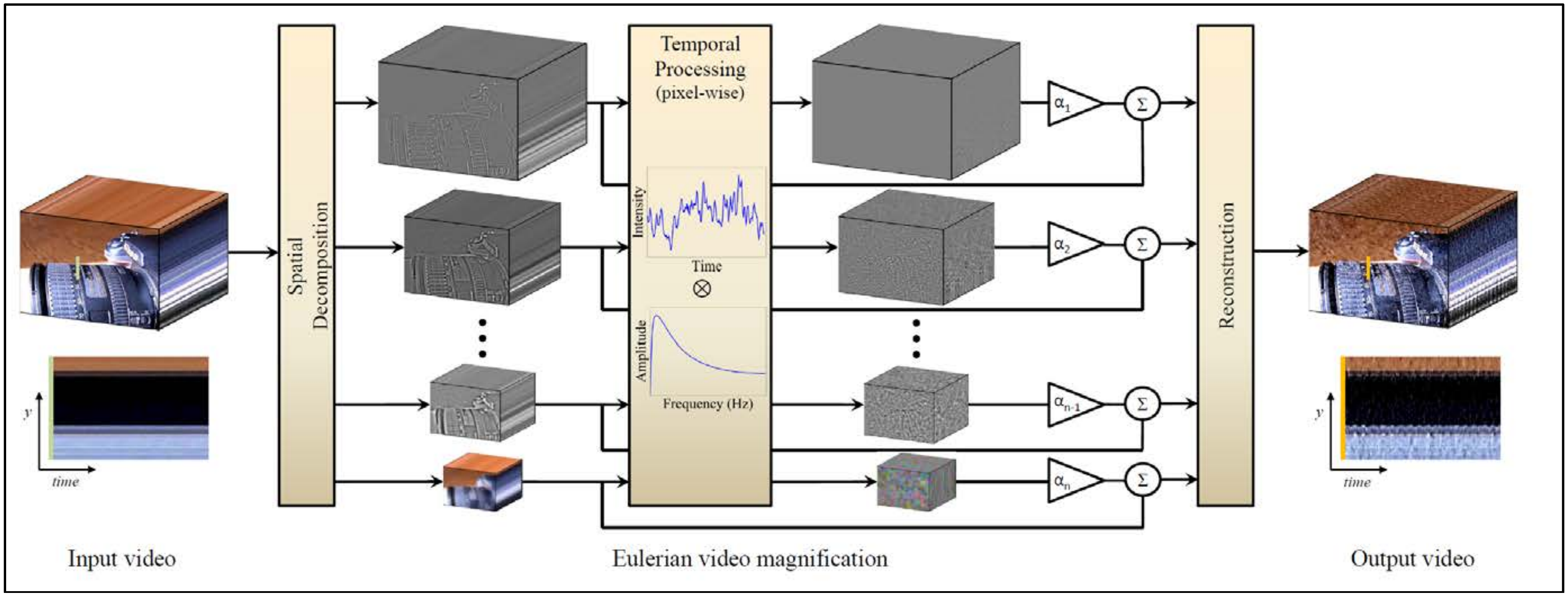


Figure 2. Diagram. Eulerian video magnification (VidMag) techniques – flow of work (Wu et al., 2012).

CHAPTER 2. IMPLEMENTATION AND TESTING OF VIDMAG

The basic strategy for applying VidMag is shown below in Figure 3. The main process is divided into four major steps: pre-processing, video magnification, post-processing, and physiological analysis. This section introduces the methods applied in each of these steps and is followed by a detailed discussion of each method in the subsequent subsections.

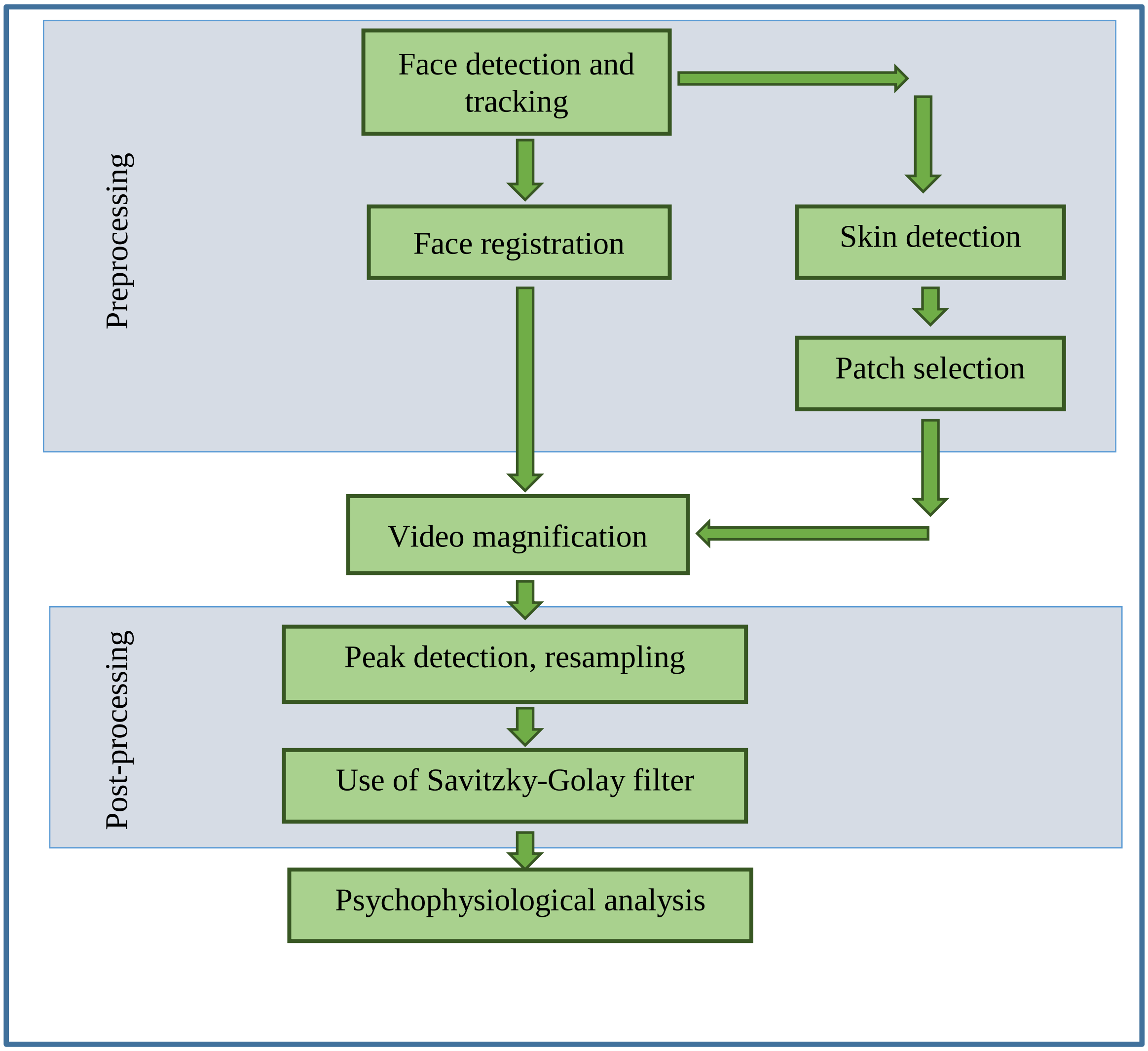

Preprocessing
Face detection and tracking
Face registration
Skin detection
Patch selection
Video magnification
Post-processing
Peak detection, resampling
Use of Savitzky-Golay filter
Psychophysiological analysis

Figure 3. Diagram. VidMag application workflow.

VidMag needs a stable video sequence to maximize pulse detection performance; however, development thus far has focused on optimal lab conditions. Therefore, the NDS video needs several preprocessing steps before they are fed into the VidMag algorithm. Face detection, alignment of faces from consecutive frames, and skin detection are some of the major steps in preprocessing. Automatic skin detection is needed so that the algorithm knows which pixels to track in order to obtain the vital signal from a consistent patch of skin. Once the video patch is selected and magnified, it passes through a bandpass filter that captures fundamental frequency components of feasible heart rate. To improve reliability, vital signals are extracted from different patches of the magnified video and combined into a single measure. The post processing step includes peak detection, alignment of the peaks from different patches, and finally peak filtering to eliminate high frequency noise and spurious peaks.

FLOW OF ALGORITHM

As mentioned in the previous section, we divided the implementation into different steps. First, we applied a face detector to identify the faces in the video view of the vehicle cabin. The face detector localizes the position of the face in each video frame. Standard face detectors can also detect the position of fiducial points like eyes, mouth, nose, and eyebrows. These features help us track the 3D orientation of the face in the 2D video image coordinates. During normal driving conditions, the driver moves his/her face in all three translational and three rotational directions. Therefore, a single point on the face is not stable, as it moves within the 2D image and may even disappear from view at times. Hence, we need to stabilize the face in order to track each pixel over time (later we will introduce patches to improve robustness over single-pixel tracking). This process of understanding the face orientation and tracking the areas of interest is known as image registration. We automatically selected a reference frame in which all other frames were registered. This calibration process mitigates noise that would otherwise result from both the driver’s intended face motion as well as motion induced by the vehicle dynamics. Once the faces over different frames were registered to their reference locations, the video was ready for the VidMag algorithm and the resulting signal magnification. Detailed parameter settings for VidMag, such as the filter parameter and gain control for the magnification, are discussed later in this chapter.

Not all parts of the face contain the information from which the pulse rate may be extracted. Only the skin provides such information; features such as the eyes and facial hair introduce undesirable noise to the assessment. Therefore, before extracting the pulse rate from the magnified videos, we implemented an automated skin detection algorithm. An effective algorithm was not identified within the existing body of work, requiring us to develop a novel skin detection algorithm that could be applied to grayscale skin pixel localization. We then selected some square patches from the known locations of the skin pixels.

For the post-processing step, we extracted pulse information from each patch, and then, by using a maximum voting technique, we selected the time instances of the pulse beat’s peak locations. We developed an algorithm to compensate for the effect of low frame rates to increase resolution of the extracted signal. A Savitzky-Golay (S-G) filter, specialized for filtering non-uniformly sampled data, was used to filter the final heart rate data. This step eliminated most of the high frequency noise in the heart rate sequence. Finally, to determine drivers’ physiological characteristics, the heartbeat sequences were analyzed in the time or frequency domain using the Lomb-Scargle periodogram, which is one of the most commonly used frequency analysis tools for understanding the psychophysiological condition.

In the following sections, we will discuss all of the steps in more detail.

PREPROCESSING: FACE DETECTION AND COMPENSATION FOR NATURAL MOVEMENT OF THE HEAD

The results reported for the VidMag algorithm in previous studies considered heart rate measurements in laboratory settings under restricted conditions; however, naturalistic situations are quite different from these settings. VidMag tracks the changes of a pixel over time at a fixed location on a person's face. (The face is well suited for analysis, as it typically provides a consistent region of exposed skin). In naturalistic driving scenarios, subjects frequently change the position and orientation of the head. Any kind of natural vibration also adds to the random motion, which is not well tolerated by VidMag. In standard NDS videos, drivers need to look at their vehicle's mirrors and scan their surroundings to actively monitor road and traffic conditions. Substantial head movement is also natural during specific maneuvers, such as lane changing, turning into intersections, or interacting with passengers. The frequencies of such driver head movements are also governed by traffic conditions, road conditions, and individual driving patterns. For example, while driving in a city, drivers need to scan the road, other vehicles, and pedestrians more frequently than while driving on an interstate or freeway. In comparatively stable driving scenarios, drivers tend to look forward and often use eye movements to scan the roadway. These latter cases are particularly favorable for our application. Thus, we implemented a face detector and tracker to identify scenarios where the face is comparatively stable and the driver is looking forward so that heart rate measures are more robust. In the next section, we discuss how a state of the art face detector was leveraged and extended to this problem space. Some additional face detectors that were not considered worthy of detailed evaluation and application, but which were initially considered, are reported in Appendix A.

Face Detection Using the Histogram of Oriented Gradient (HOG) Method

While face detection and tracking has been a much studied issue in computer vision for the last few decades, researchers have yet to find a face detector that can work reliably in all physical and environmental conditions. Due to the deformable structure of the face and the large variation of possible poses, solving the problem is quite difficult. Zhu and Ramanan (2012) introduced a tree-based approach using a histogram of oriented gradient (HOG) template for detecting faces in the wild; to date, it is considered one of the most state-of-the-art face detectors. We tested several collected datasets to determine the HOG template's face detecting accuracy in the desired driving context. In this particular study, we attempted to determine how temporal information improves the face detector's performance.

During evaluation, we found that this approach had some limitations. While testing the face detector with different video data, we determined that frame-by-frame calculation is computationally costly, which may make the method unsuitable for large NDS datasets. Secondly, for different illumination conditions and poses during the driving scenario, it sometimes became difficult for the face detector to locate the face independently. Thirdly, this method did not necessarily determine the face position in cases of occlusion, which often occurs in the vehicle due to head movements and various actions. Lastly, the localization accuracy depends on the number of parts in each of the models. Zhu and Ramanan (2012) used three models for face detection, whereas we used a 146-part model, which is moderate in terms of evaluation time and accuracy.

Improvement Strategies

We incorporated some modifications to the method in terms of time and accuracy and compared the results with the frame-by-frame evaluation. The steps taken in that process are described below.

Step 1: We assumed that in the first frame of the video, face detection was accomplished correctly. After the face was detected from the first frame, we determined the spatial boundary in which the face was located to provide the reference for the face in the next frame. We then extended the face boundary from Frame 1 by 70% on both sides and conveyed this spatial information to the next frame for an initial crop. For the next calculation, we ran the face detector only for the cropped region. Finally, we experimented with our dataset to make sure that a 70% extension was enough for two frames at a minimum of 7.5 frames per second (fps).

Step 2: With the first detection of the face in Frame 1, we also gathered information about the pose and the scale of the face in the image. This gave us the flexibility to dictate that, in the next frame, the face detector should search only for adjacent poses and the scale of the last detected face. In this regard, we assumed that there was minimum distance traveled along the optical axis (perpendicular to the face plane). This served to reduce the computation time.

Step 3: The algorithm used a tree-based search method where the base of the nose was the root of the tree. We used an optical-flow-based (Liu, 2009) tracking method to force the algorithm to search in a specific location of the cropped image, such that there was consistency in the results of the two frames. Figure 4 shows the two consecutive frames. As we knew the location of the nose base in the first frame, we computed the optical flow vector (magnitude and direction) and then moved the base point estimation accordingly. Next, we biased the detector to search only around the estimated point in the given pyramidal scale. This improved accuracy and helped reduce the number of tree searches. Figure 5 shows gradual improvements for all the steps and indicates that knowledge of the fiducial points from the previous frame drastically reduced the computation time for the next frame. The generation of optical flow increased computation costs somewhat, but they were much smaller compared to the original frame-by-frame face detection.

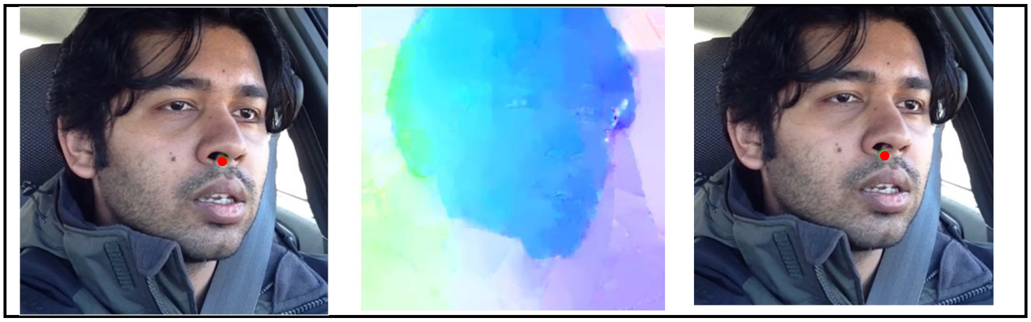


Figure 4. Images. Optical flow tracking.

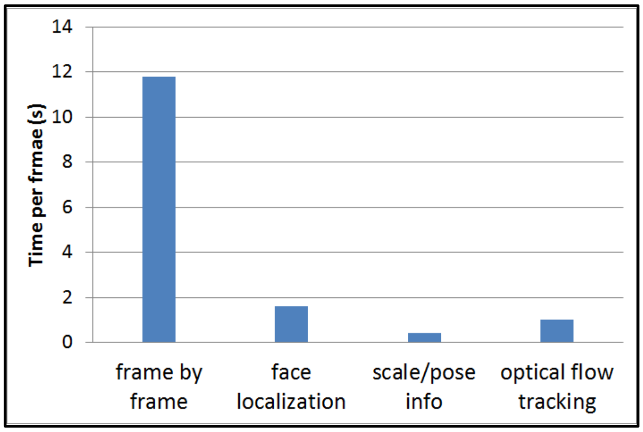


Figure 5. Graph. Improved time efficiency.

Experiment and Results

For this study, videos came from naturalistic driving data collected previously by the Virginia Tech Transportation Institute (VTTI) and their partners and self-collected on-road driver data using an iPhone camera. The second set of data allowed controlled collection under various lighting conditions. A total of around 9,000 frames from VTTI and approximately 2,000 self-collected frames were tested. The results will be discussed here in two ways: (1) in terms of face detection accuracy as a whole and (2) in terms of localization accuracy.

After testing all of the videos, we found that use of temporal information, given that the first frame was accurate, resulted in more than 99% accuracy. In all of the cases, the proposed method successfully tracked the face irrespective of the situation. Secondly, there were no false positives in the results. Some of the results can be seen in Table 1. In this table, the first row shows results from the original method (Zhu & Ramanan, 2012), and the second row shows the results after the inclusion of temporal information. This shows robustness for false negatives (case a), successful face detection in high glare (case b), face detection for infrared cameras (case c), and successful detection of the face in case of partial occlusion (case d).

The second method of testing involved comparisons utilizing the localization error. We annotated face data from VTTI videos collected during NDSs. This video data has seven points marked on the face: the four eye corners (four points), the mouth corners (two points), and the nose tip (one point). When we compared our results, either with or without temporal information, we found that temporal information led to better average accuracy (as shown in Figure 6). Accordingly, we concluded that the use of video temporal information helps detect faces more accurately and quickly.

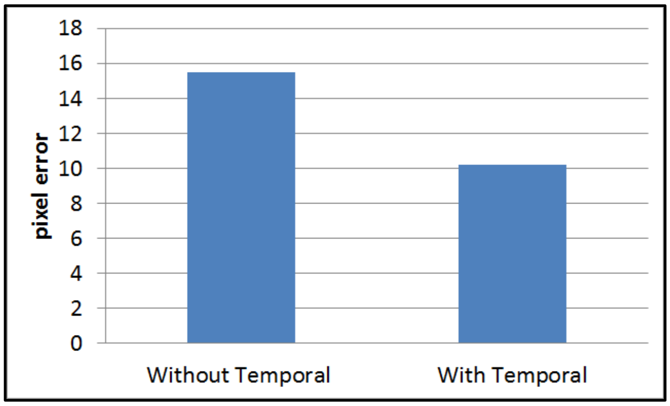
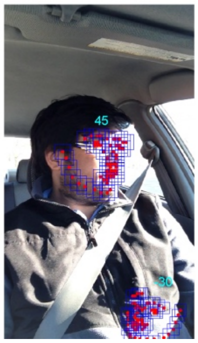
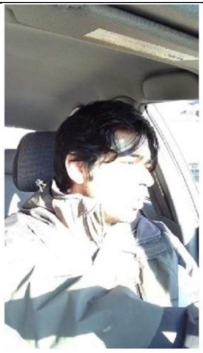
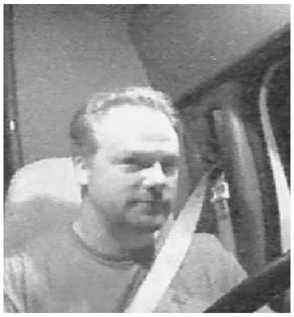
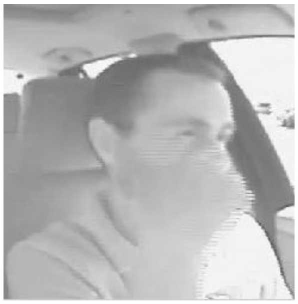
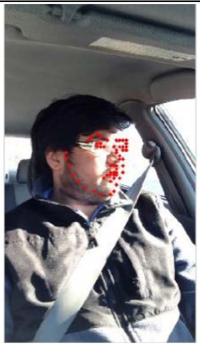
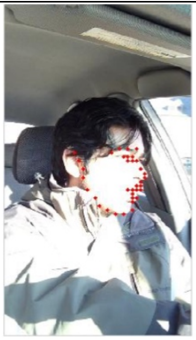
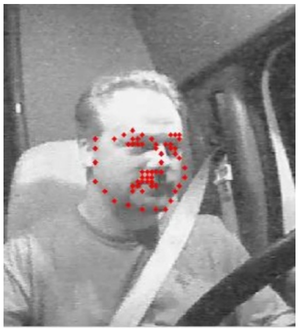
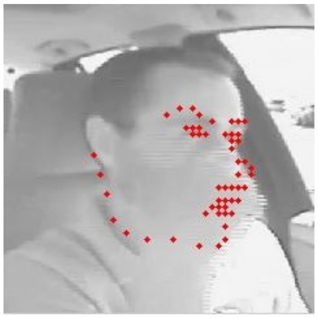


Figure 6. Graph. Localization accuracy.

Table 1. Face detection accuracy with (top row) and without (bottom row) temporal information.

 a	 b	 c	 d
 a	 b	 c	 d

Intra-face Face Detection Methods

Xiong and De la Torre (2013) introduced a method built on the Viola and Jones face detector. This method uses a supervised descent method for robust localization of fiducial points. The algorithm has a built-in tracking facility for face points, and the authors' method uses 49 points on the face. We tested this algorithm with some of the quality images from VTTI and found that this method had comparatively low localization error relative to the method used by Zhu and Ramanan (2012).

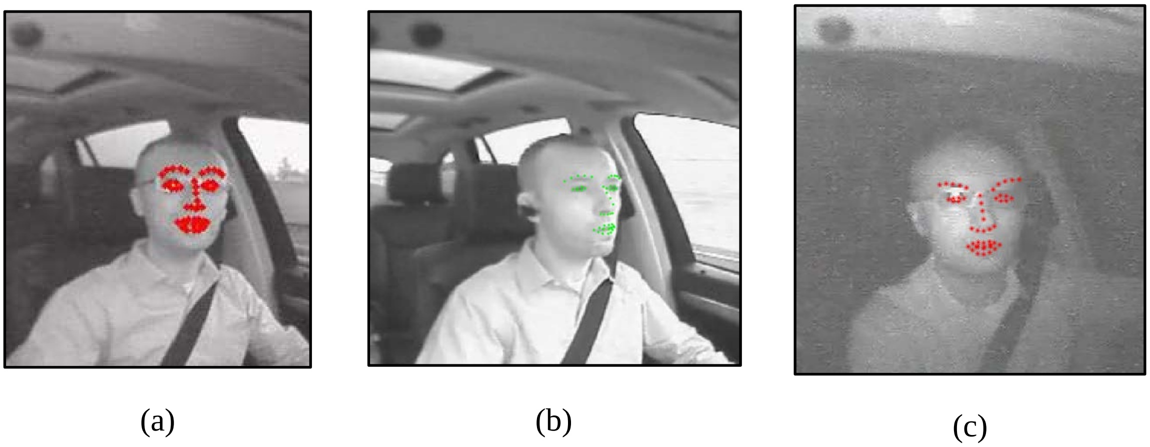


Figure 7. Screenshots. Results from intra-face: (a) daytime, (b) different pose, (c) nighttime infrared frame selection.

We extracted landmark points from each frame using the face detectors (Zhu & Ramanan, 2012; Xiong & De la Torre, 2013). Figure 8 shows 49 landmark points extracted for the face, including eyebrows, eyes (including eyeballs), nose-bridge, lips, and mouth. In almost 98% of the tested cases with MiniDAS (VTTI's proprietary data acquisition system) video frames, the method was able to detect faces with fiducial points in the daylight condition. These face videos contain three color channel and the video is recorded at a frame speed of 10 fps. The VidMag tracks changes in color intensities of a fixed point on the face over time. Therefore, we had to identify the same location on the face and track its changes. Because drivers often change their head position and orientation with respect to the camera, tracking a single point requires exact registration of faces between frames. Linear methods like homography show very good results for small movements, but the error increases as the rotation angle increases. This is partly due to interpolation of the missing points and partly due to loss of weak perspective.

Therefore, we selected a sequence of frames where the head movement was limited. This limitation did not particularly affect the testing scenario, as drivers tend to look forward most of time (roughly greater than 90% of the driving time). Selecting these particular frames allowed us to estimate the heart rate without losing most of the frames due to head movement. In order to recognize a sequence of frames with limited head movement, we tracked the fiducial points for consecutive frames and calculated the average shift in the Euclidian distance for the 49 points over time. If that value was within a defined limit, we accepted the results as an allowable sequence of frames. If the value was not within the defined limit, we discarded that sequence. In this study, we selected frame sequences where the mean-square shifts of the landmark points were limited to 17 pixels, for an average face dimension of 125 × 125 pixels. Figure 9 plots the fiducial points for two sequences of frames, showing the spread of the points over time. The face images are for illustration purposes. Figure 9(a) shows a compact plot of the points. We qualified this as a potential frame sequence on which to apply VidMag for heart rate extraction. The points in Figure 9(b) have very high scatter and so we discarded these frames. These discarded sequences mostly consisted of events like mirror checking, blind spot checking, etc.

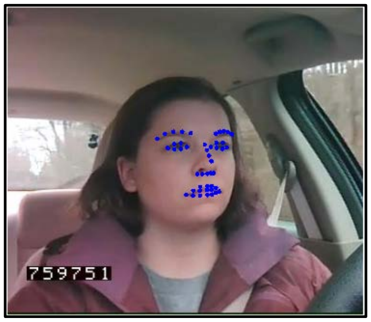
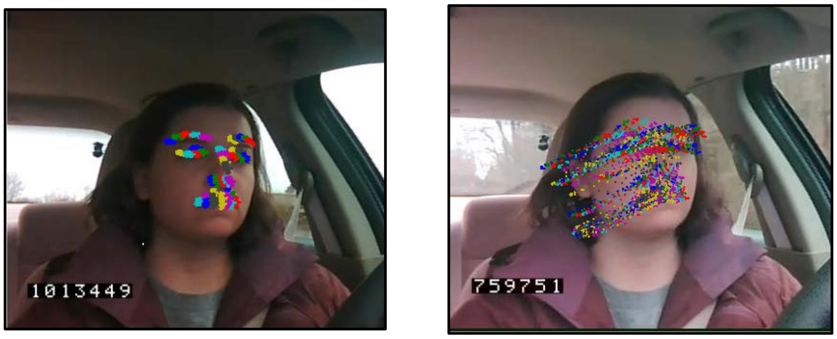


Figure 8. Screenshot. Face detection with 49 fiducial points.



(a)

(b)

Figure 9. Screenshots. Selection of frame sequence based on the movement of fiducial points: (a) for a sequence where the points are relatively restricted, (b) for a sequence where the points move extensively.

Face Registration

Natural vibrations of the face can be reduced by registering frames. Registration generally refers to alignment of an object from two different frames, where the position of the object may change with respect to the camera. In many of these cases, translation-only registration, which reduces the complexity of image registration, can be performed. As an example, we extracted landmark points from consecutive frames and registered them by translation (Szeliski, 2006). Figure 10 shows registration of two randomly chosen frames from a video sequence. Figure 10(a) and (b) show the frames with their corresponding landmark points. Figure 10(c) shows two superimposed frames before the registration process, and Figure 10(d) shows the results after linear sub-pixel registration (Szeliski, 2006). The purple and green colors indicate registration mismatches from the two frames. In a more challenging video sequence where the user is very active, the task of registering a face becomes more challenging.

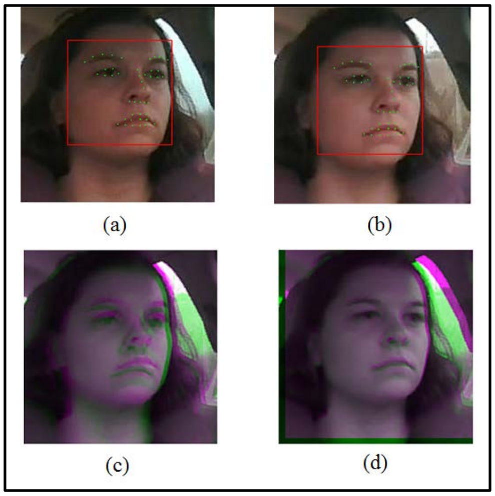


Figure 10. Screenshots. Face registration. (a) & (b) Two randomly chosen faces with automatically detected landmark points. Superimposed frames: (c) before registration, (d) after registration.

We had the option of either using direct homography for registration or of first registering faces by linear shift and then using homography on the shifted image. In general, we employed an incremental registration process, which first aligns the fiducial points on the face for consecutive frames by translation and then uses this estimation to align the points with a single frame in the video sequence. Our algorithm automatically selected the frame to which all other frames should be registered by computing a distance matrix. Next, homography was applied to register faces more robustly. Homography calculates a linear transformation between the two frames on the basis of the 49 landmark points. The leftmost image of Figure 11 shows the points in the original frame sequence. The middle and the right images consecutively show the results of translation and homography. This method allowed us to improve the mean deviations of pixels from a 34-pixel unit to a 5.5-pixel unit.

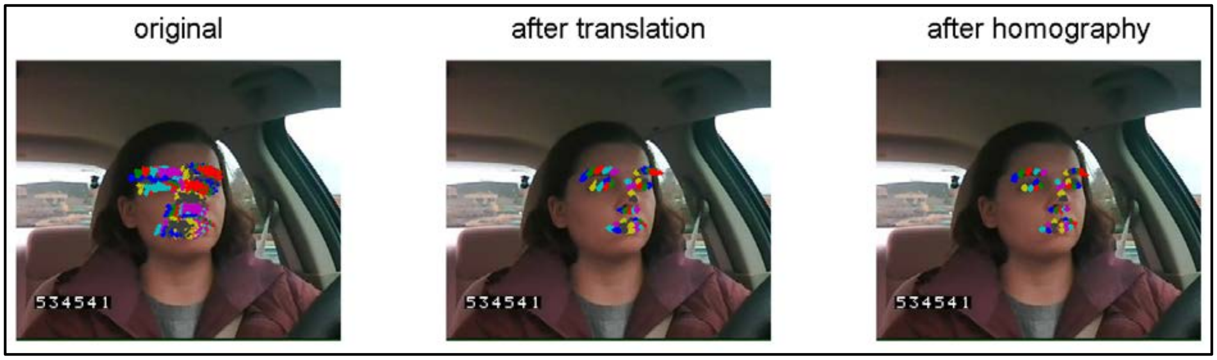


Figure 11. Screenshots. Frame registration showing improvement in consecutive frames. The original scatter sequence shows how fiducial points move with time. After homography, the movement of fiducial points has been minimized.

SKIN DETECTION AND IMPROVED BEAT ACCURACY

After we applied the video magnification algorithm to the registered face video, we identified the skin pixels, a necessary step to ensure that no unwanted face pixels contributed to the pulse rate information. For example, pixels from the eyebrows or nostrils introduce noise into the measurement, and therefore need to be eliminated during post processing. Because there was no standard algorithm available that detected skin pixels in grayscale images, we developed a novel skin detection algorithm.

For detection in grayscale, we automatically sampled some pixels from the face which we believed to represent skin based on their geometric location (e.g., skin patches near the forehead, nose, and cheek). Then we created a statistical model with the grayscale and texture information from the sample space. Finally, using a prior distribution of face location from the face detector, we applied the statistical model in a Bayesian framework to identify skin pixels. Figure 12 shows an example of skin detection and subsequent processing. Once the skin pixel locations were known, we divided them into small patches. In most of our work, we used square 10 × 10 patches of pixels. We extracted heart rate information from each of the small patches, detected the pulse location, and aligned them together (right part of Figure 12).

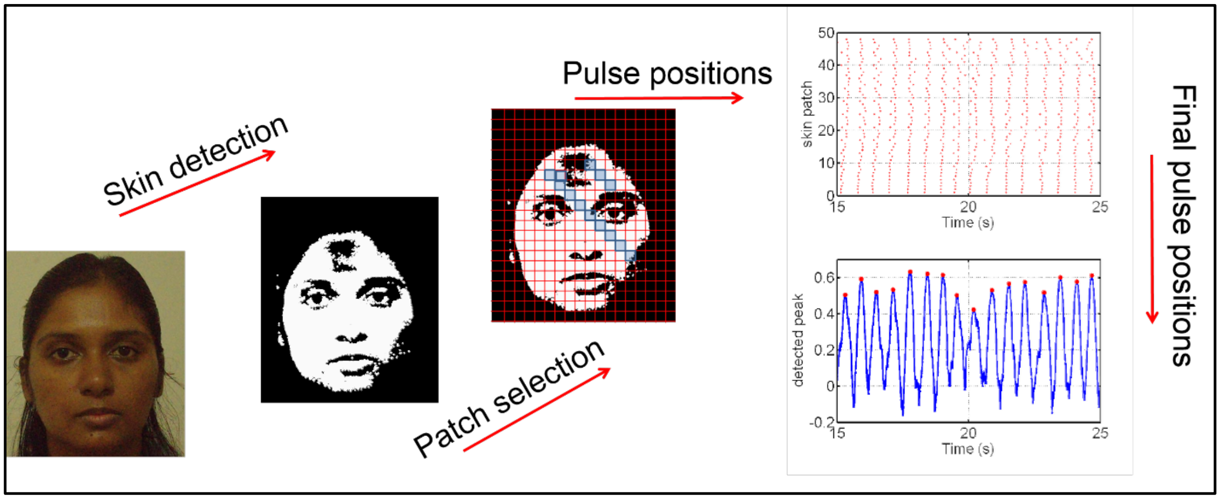


Figure 12. Diagram. Skin detection and heart rate estimation.

Patch Selection

In practice, each pixel position from the detected skin may contribute noise of different magnitude and phase to its extracted temporal pulse rate signal. Therefore, extracting the final signal by averaging all the pixel values of a frame in the magnified video may lead to inclusion of noise in the final estimate. During the registration process, we saw that the mean error was 5.8 pixels. Accordingly, we divided the face's skin pixels into a number of small square patches (i.e., 10 × 10 pixels) and extracted the pulse rate signal from each to minimize the effect of noise from registration error. Formally, we divided the skin pixels into n_p skin patches. Each patch from the magnified video independently estimated the pulse signal, S_{np}. We detected the pulse positions from each S_{np} by using a peak detection algorithm. The pulse positions, P_{np}, were then aligned together to determine the real signal and eliminate the outliers. Here the basic assumption was that all the patches should ideally produce exactly the same signal without any noise or phase delay.

Skin Detection

Skin detection using color information is an established field of study. The primary motivations for detecting skin in images and videos include face detection (e.g., Terrillon, Shirazi, Fukamachi, & Akamatsu, 2000; Chai & Ngan, 1999), gesture and action recognition, surveillance, and adult content screening (Fleck, Forsyth, & Bregler, 1996). The primary concentration of most research has been the investigation of different color cues for invariance to lighting conditions, pose, and ethnicity. A number of authors (Vezhnevets, Sazonov, & Andreeva, 2003; Kakumanu, Makrogiannis, & Bourbakis, 2007; Phung, Bouzerdoum, & Chai, 2005) have provided comprehensive surveys of skin color modeling and skin segmentation techniques using different color cues and color spaces. However, the emphasis in these studies is on color-based analysis, and SHRP 2 face videos are in grayscale. The camera operates in normal mode during the daytime and uses infrared for nighttime video, but in both cases the video is in grayscale. As we did not have any color information from the face videos, we used a novel skin detection method to identify the skin region on the face.

Theoretically, if it is possible to locate a face in an image and find the locations of the eyes, mouth, nose, and other parts, it should be easy to find the skin pixels from the relative location of a pixel with respect to all those parts. But in practice, identification of skin depends on two major factors: visibility and local appearance. Facial hair, tattoos, hairstyle, tanning, use of glasses, headbands, scarves, and winter-wear are some of the external factors which influence the visibility of a skin pixel, while differences in skin pigmentation, which cause a wide variation in the appearance of the skin, are the major intrinsic factor. Similarly, external factors like illumination conditions, shadows, and reflections also change the skin’s appearance. Conventional methods reported in the literature (Phung et al., 2005) have shown that color information shows excellent results for skin detection. A particular subspace in the three-dimensional color spaces successfully represents a wide variety of skin pixels under different conditions. But, for grayscale images, that subspace is not accessible. Histogram analysis alone does not show any distinctive behavior that may lead to a universal skin classifier.

Though the appearance of skin can vary in different images, with the exception of cases where there are harsh illumination conditions, the characteristics of skin pixels of a single face in a single image are close to each other. Based on this premise, we developed a single-image-based skin detection algorithm. One of the major advantages of using an image-specific skin detection algorithm is that its variability lies within that particular image and, with prudent strategies, reasonable information about the variability can be obtained. The sketch of our algorithm is shown in Figure 15. We first used a face detector to detect the face and its fiducial points. Next, with the help of the points, we sampled an area of the face with a higher probability of being skin. Such areas include part of the forehead, the nose, and part of the cheeks. We call this skin sample the “mask.” Once we developed the skin mask, we also determined the non-skin mask. The non-skin mask comprises the near black parts in the image, including the eyes, mouth, eyebrows (as selected by the face detection algorithm), and any dominant edges.

We evaluated our results on two standard skin databases: the ECU skin database (Phung et al., 2005) and the SFA skin database (Casati, Moraes, & Rodrigues, 2013). Both datasets contain images of men, women, and children with different skin colors, different head poses, with and without facial hair, and under different illumination conditions. Images from the SFA dataset were collected with restricted backgrounds and in a studio-like conditions. Each image contains only one face in close camera proximity. Figure 13 shows sample images and their ground truth from the SFA database, which has 1,118 images in total. Our face detector worked for 701 images; all statistics and results are discussed based on these images. The ECU database contains 4,000 images in more naturalistic conditions with a variety of backgrounds. Unlike the SFA database, the images in the ECU database contain skin from different parts of the body. This database was more challenging because of the varying scale, pose of the person, and several occlusions. Figure 14 shows a sample of the ECU dataset with segmented skin ground truth.

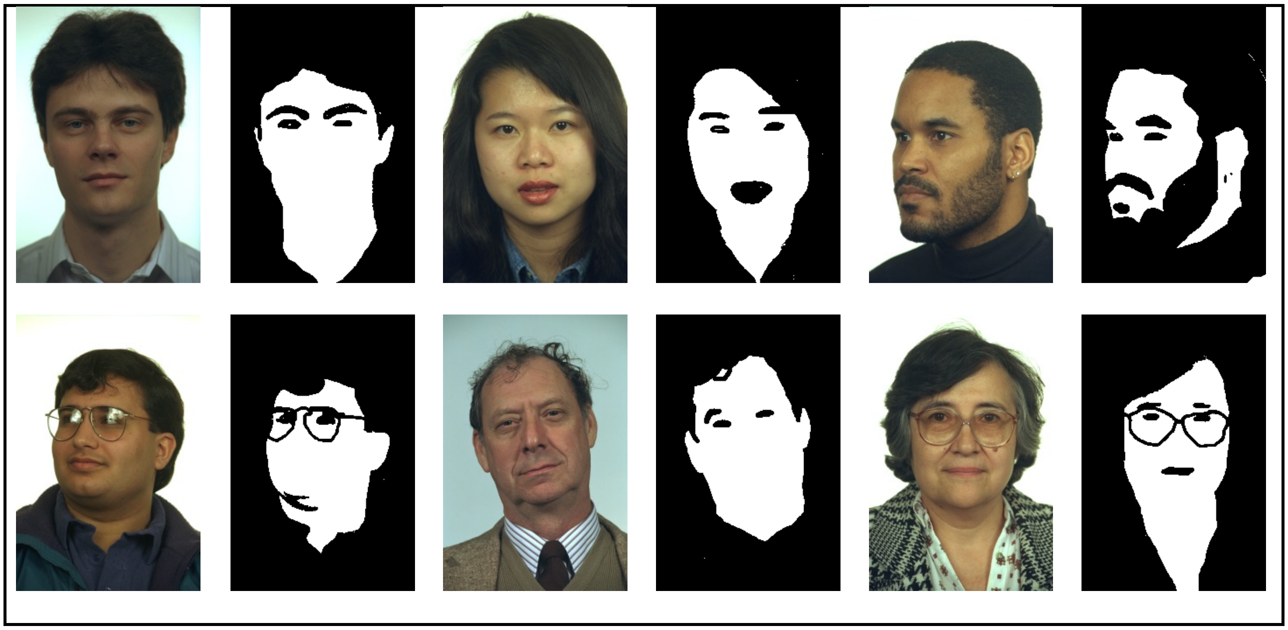


Figure 13. Images. Sample images from the SFA dataset: This image set comprises face images with different level of pigmentation, gender, poses, and variation of facial hair, use of glasses, and different hairstyles.

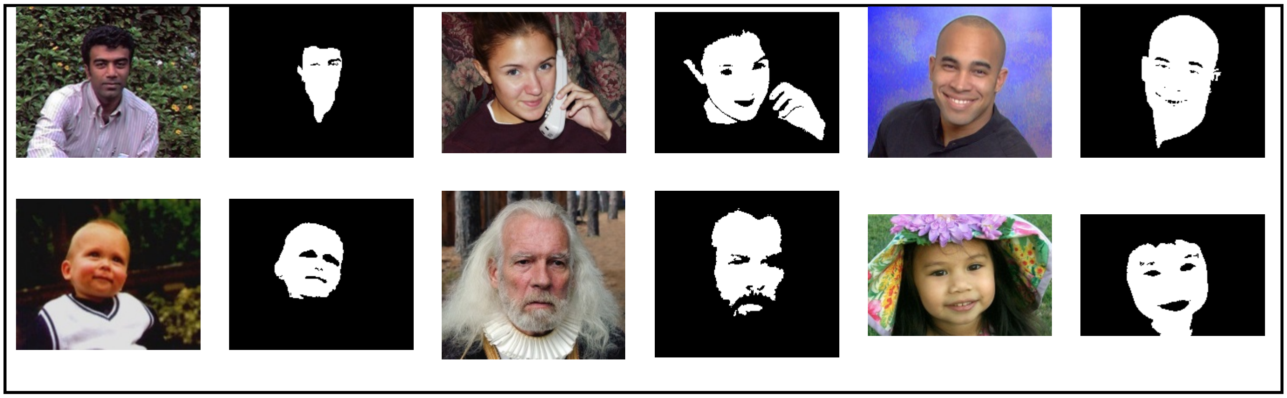


Figure 14. Images. Sample images with skin ground truth from the ECU dataset. Images from this dataset were taken with different indoor and outdoor backgrounds and more challenging situations.

The skin detection algorithm developed for this project is illustrated in Figure 15. We tested the algorithm on a color image, but the image was converted to grayscale and no subsequent operations, including the face detection algorithm used any color information. From the knowledge of the fiducial points on the face, we selected a skin mask, a non-skin mask, and a distance-based weight map, $W_{dist}(I)$, for image "I." This weight map was proportional to the prior probability of each pixel in the image being skin. The basic assumption here was that the closer the pixel is to the face, the higher its probability of being skin. As the pixel goes away from the center of the face, this probability decreases. The distance weight map also included the pose information. Depending on the person's head pose, the weight map changed its probability of the pixel being skin. The non-skin map selected the eyes, nostrils, and any other near black pixels (intensity < 5). The skin map comprised part of the forehead, nose, and cheeks. Our visual inspection showed that the chosen mask areas were mostly uncovered by facial hair and clothing/accessories. We tested over 1,200 images from both databases, and a comparison with the ground truth showed that 97% of the time the mask was a skin pixel.

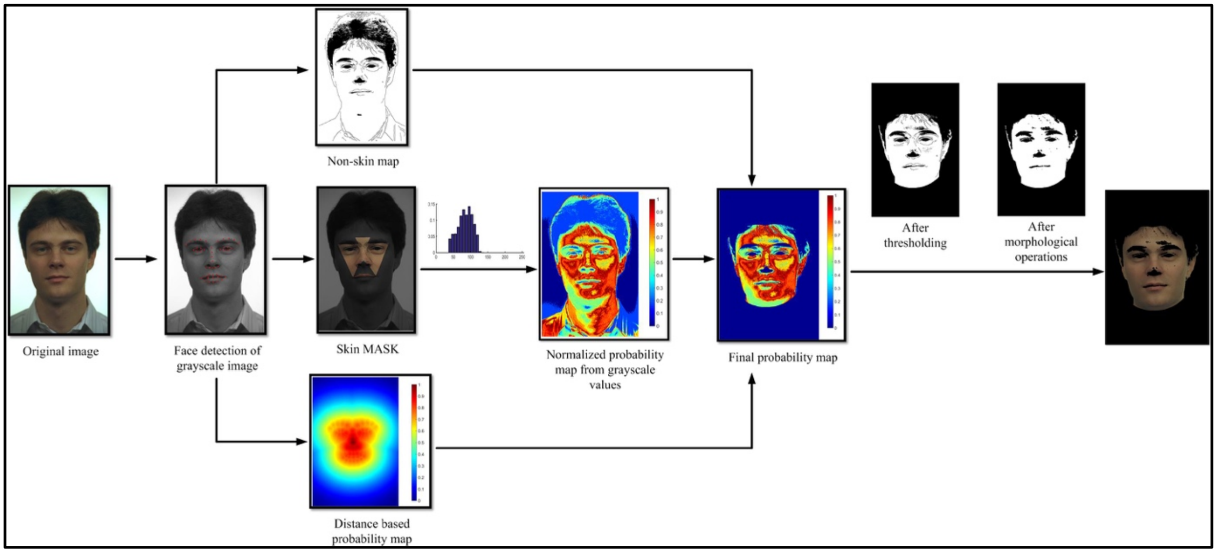

Non-skin map
Original image
Face detection of grayscale image
Skin MASK
Normalized probability map from grayscale values
Final probability map
After thresholding
After morphological operations
Distance based probability map

Figure 15. Diagram. The skin detection algorithm.

Once the skin mask was obtained, it was used as our reference for the full skin area. We computed the histogram-based probability of each grayscale value and assigned the probability to each of the pixels. The weighted value of each pixel (p) shown in Figure 15 is given by:

$$w_{gs}(p) \propto P(skin|c)$$

where, $c \in \{0 \dots 255\}$ are the grayscale values. The corresponding weight map of the image I is $W_{gs}(I) \in \{w_{gs}(p)\}$. The final weight map was calculated from these two weight maps, $W_{gs}(I)$, $W_{dist}(I)$, and from the knowledge of the non-skin map. Applying a threshold for each pixel, we calculated the binary image. Finally, after using morphological operators like erosion and dilation, we obtained the final skin map. We used two threshold values in our algorithm. The first was for the distance map, t_d. This threshold determined how far from a face a pixel should be considered a skin pixel. Any pixel beyond that point was given a weight of zero. The second threshold was for the grayscale map (t_{gs}). As discussed before, every pixel had a weight proportional to the probability derived from the grayscale histogram of the mask. We varied both thresholds to find the best set of thresholds to provide acceptable values for precision and recall. Figure 16 shows the average precision-recall curve for different threshold combinations. For our application, we were more interested in precision than recall. Upon careful inspection of the detected skins in the images, we found that the recall of the algorithm was mostly affected by misidentification of the neck area skin. However, for our application, we are not specifically focusing on the neck area skin. Therefore, a recall of 75% was assumed to be acceptable for the areas where we selected thresholds, resulting in an average precision of at least 90%. This corresponds to $t_d = 0.39, t_{gs} = 0.29$.

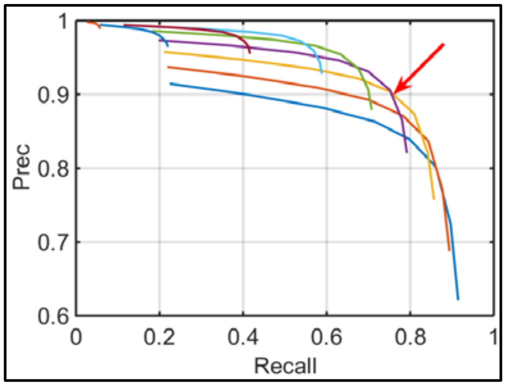


Figure 16. Graph. Precision and recall curve for threshold selection.

Results from the SFA dataset are shown in and Table 2 and Table 3. Table 2 shows success cases and Table 3 shows some of the failure cases. The results showed that this algorithm, although simple in nature, could successfully detect skin for people of different gender, age, and ethnicity. It often overcame challenges imposed by wearable devices (e.g., glasses), facial hair, and occlusion by hair style. One of the major drawbacks of this method was that if any part of the face or surrounding area matched the grayscale values in the mask, it blindly selected that area as a skin pixel. Also, there were cases where the probability distribution of the ground truth and the rest of the image were very similar. In other words, in theory, there would be a very low correlation between the distribution of skin and non-skin pixels as well as a low entropy distribution for the skin pixels. In, practice, this does not always hold true. An example can be seen in the top row of Table 3, where the hair or shirt of the person was confused for a skin pixel. In the bottom-right case, the same confusion can be seen over the change in an individual’s skin tone. Illumination and other reflections also created ambiguous situations that confused the algorithm, as they modified the appearance of the skin totally, as shown in the bottom left example in Table 3. Finally, as our algorithm primarily depended on a face and fiducial point detector, any fault in that detector propagated to the outcome of the whole algorithm. The bottom-middle example in Table 3 is one such case—the face detection algorithm fired at the wrong area of the image, leading to inaccurate selection of the skin-mask, throwing off the subsequent skin pixel detection process.

Table 2. Example of successful cases of skin detection from the SFA database.

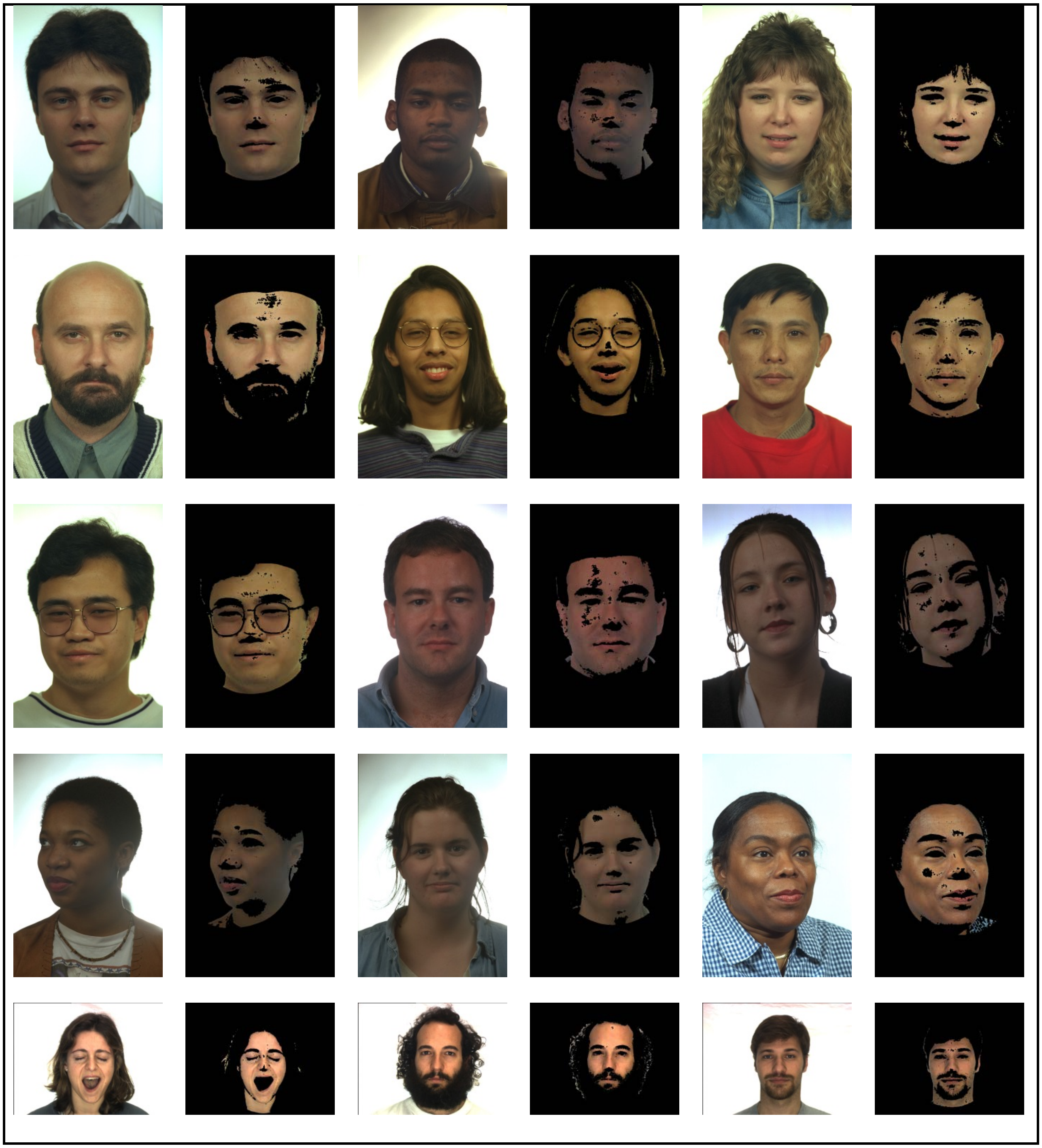
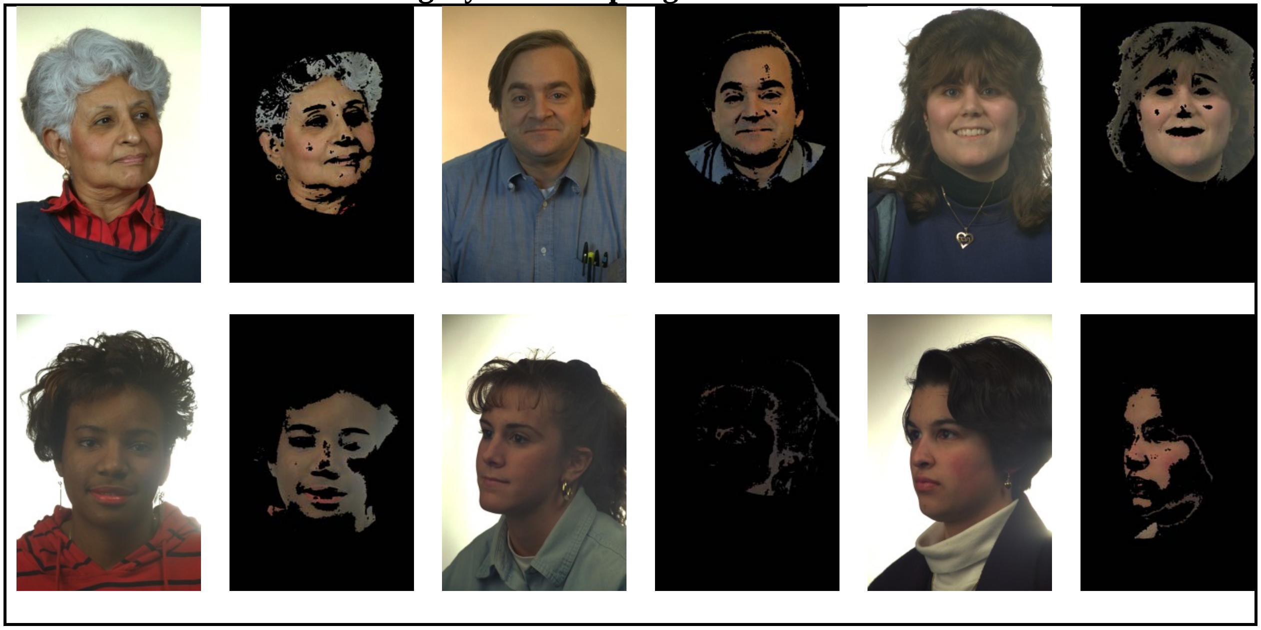


Table 3. Failure cases from the SFA dataset – Top-left and top-right: hair is confused as skin; top-middle: part of shirt has been selected; bottom-left: illumination and background confusion; bottom-middle: effect of wrong-face detection; bottom-right: inadequate grayscale sampling from mask.



Given the aforementioned pros and cons, using other facial features, especially texture features, is recommended for guaranteeing better skin segmentation. In some cases, researchers have used texture (Cula, Dana, Murphy, & Rao, 2005; Fotouhi, Rohban, & Kasaei, 2009; Garcia & Tziritas, 1999) or shape (Drimbarean, Corcoran, Cuic, & Buzuloiu, 2001) information in combination with color-based methods in order to refine the resulting skin segmentation. Wavelets, contourlets, and textons are among the textural features that have been previously used in conjunction with color or grayscale cues. Fractals, lacunarity, GLCM, and local binary patterns are some of the other methods that have been used for detecting texture in images. The particular method presented here has the major advantage of having a very low processing time. As face detection is part of the actual algorithm, this method only needs to find the grayscale histogram and assign a probability to every pixel. As such, using these other features was not explored further due to the anticipated additional processing requirements.

VIDEO MAGNIFICATION: PEAK DETECTION AND RESAMPLING

Once the face detection, registration, and skin patch selection were complete, we sent the video frames through the VidMag algorithm. In most real-life applications of interest, video data with a very high frame rate are not available. Recordings from standard mobile devices and surveillance cameras often range from 10 fps to 30 fps. The frequencies of interest for heart rate lie within the range of 40 to 200 beats per minutes (bpm), which is equivalent to 0.667 to 3.33 Hz. The rate of 10 fps satisfies the Nyquist criterion for this frequency range, but imposes limitations on the accuracy and resolution of the estimated pulse rate signal.

The basic VidMag work flow is summarized in Figure 17. Each registered video sequence was fed to the VidMag algorithm, which attempted to extract the temporal BVP signal. The BVP signal extracted from the magnified video was noisy, partially due to artifacts from motion and illumination, as well as from the low frame rate. Figure 18 shows one such sample, which includes spurious peaks. Additionally, the mean intensity varied over time. Even after smoothing and peak detection, the resulting sample suffered from low frame rate. For a 30-fps video, the heart rate estimator could not detect a heart rate that was known to be between 90 and 94.7 bpm. To address these problems, the BVP signal was resampled by a factor of 20 before the peak detection algorithm was applied. This helped increase the resolution of consecutive heart rate estimates, and heart rate could be calculated with a resolution less than 0.2 bpm.

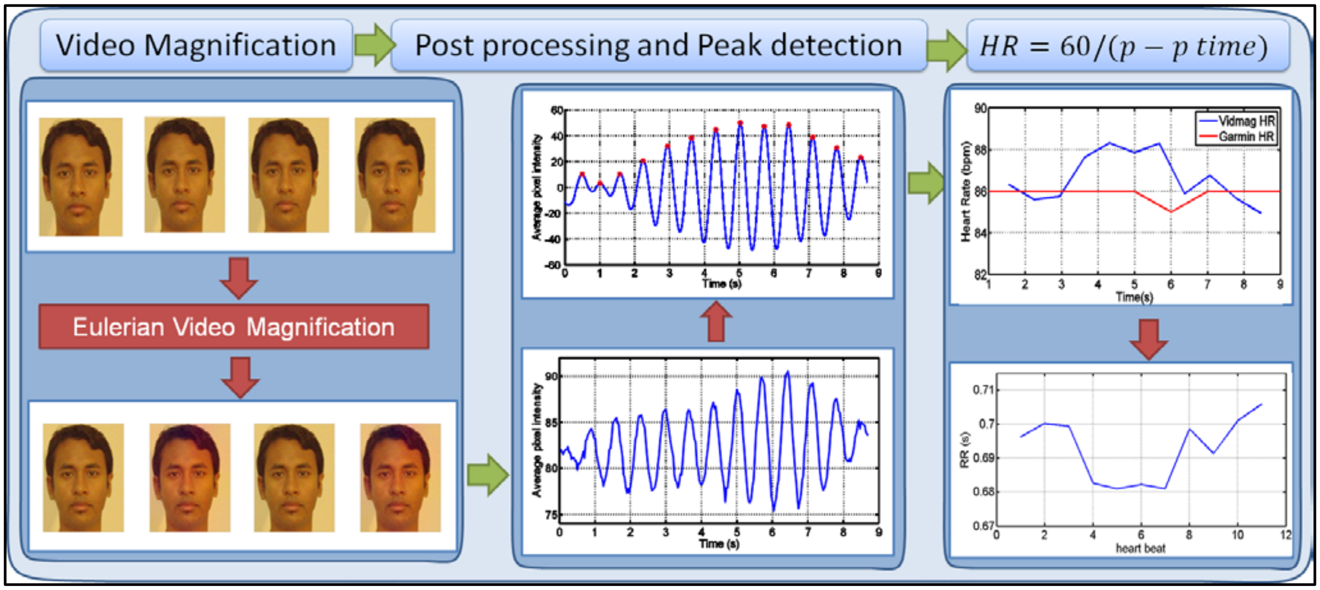


Figure 17. Diagram. Heart-rate extraction flowchart of the new system. The VidMag algorithm is used to extract a raw signal that is analyzed further for noise reduction, interpolation to address low frame rate, peak detection, and heart rate extraction.

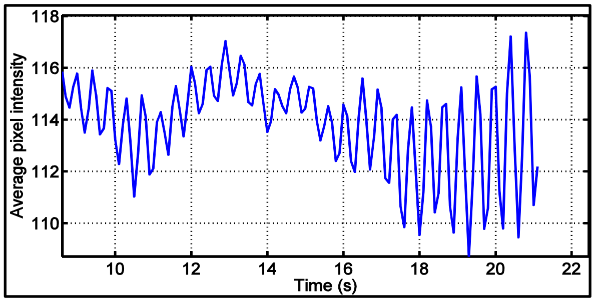


Figure 18. Graph. Raw output of BVP.

HEART RATE VARIABILITY (HRV)

Before we describe the methods for post processing of the extracted heart rate, we will discuss heart rate variability (HRV). Consideration and conservation of HRV was the biggest motivation for our post-processing algorithm development. In the following sections, we will describe HRV and the associated experimental setup for verification of HRV using VidMag. This discussion will lead us to successfully demonstrate the post-processing strategies for VidMag.

Overview of HRV

HRV represents the variability of time between two consecutive pulses (or R-R interval) as shown in Figure 19. HRV reflects the state of the autonomous nervous system and the balancing of the sympathetic and parasympathetic nervous systems. A detailed review by Acharya, Joseph, Kannathal, Lim, and Suri (2006) shows examples of HRV time domain analysis and frequency domain analysis. Healey and Picard (2005) used the Lomb-Scargle periodogram to study the stress level of a driver on the road, showing that the power spectral density in different frequency ranges is correlated with the stress level of the driver. Their work also shows that frequency analysis can provide information concerning the respiration of the driver. Rodriguez-Ibañez, Garcia-Gonzalez, de la Cruz, Fernandez-Chimeno, and Ramons-Castro (2012) showed the potential of using HRV for drowsiness detection using R-R intervals of electrocardiogram (ECG) data. Their technique calculates different time domain and frequency domain properties, and they show that specific HRV properties indicate distinct evidence of drowsiness. Acharya et al. (2006) further state that alcohol induces sympathetic activation and/or parasympathetic withdrawal. This in turn affects the measured HRV parameters. These characteristics of HRV could be crucial in assessing a person's psychophysiological conditions, particularly those that might lead to a crash.

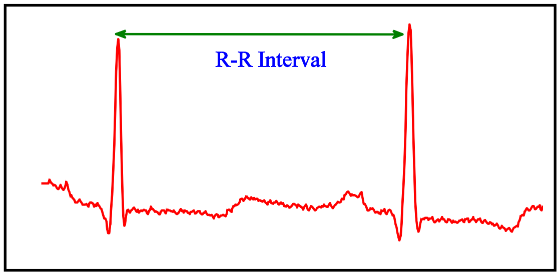


Figure 19. Graph. R-R interval. A sample ECG signal from the MIT-BIH arrhythmia dataset Goldberger et al., 2000).

Data Collection for HRV Verification

As we discussed above, HRV plays a crucial role in defining a driver's physical and psychological condition. Therefore, it is necessary to evaluate how precisely the VidMag algorithm represents the beat-to-beat accuracy. This requires an accurate measurement of the ground truth physiological signals with high resolution. We used a GE S/5 medical instrument to measure the ECG and pulse signal at a resolution of 100 Hz while recording video at a rate of 30 fps using a Canon Powershot on a tripod. Figure 20(a) shows the experimental set up. The S5 uses a three-point ECG measurement where the pulse is measured from the PPG pulse meter attached to the finger. Figure 20(b) shows sample signals from the ECG and pulse meter reading. As discussed in previous sections, we were interested in measuring the accuracy of the beat-to-beat interval as well as the actual traces of the signals. Figure 21 shows an example of the face video with a manually selected skin area.

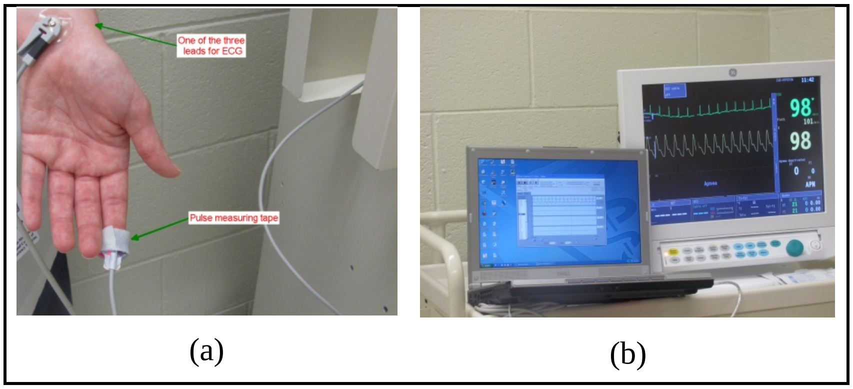


Figure 20. Photos. Experimental setup for data collection phase III. (a) PPG pulse meter is worn on the fingertip and one of the three leads of the ECG measurement kit is attached to the participant's left wrist. (b) The S5 machine with the data recording software.

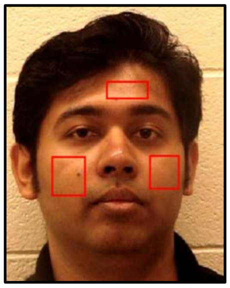


Figure 21. Screenshot. Manual selection of skin area of interest.

POST PROCESSING OF MAGNIFIED VIDEO

Pulse Signal Extraction

This section explains post processing of the raw heart rate signal extracted from the magnified face video. Estimation of actual heart rate can be challenging, particularly when the range of heart rate in a given time sequence is unknown. Human heart rates can vary from 40 bpm (0.667 Hz) to 240 bpm (4 Hz) depending on the subject's physical and cognitive condition, age, gender and various other factors. As a 240-bpm (4 Hz) heart rate, which is considered very high, is unlikely to occur in most real-life scenarios, we first used a wide bandpass filter in the range of 0.667–2.5 Hz and magnified any temporal signal in this range in the evaluation test. Due to the use of a wide bandpass signal, the extracted heart rate was often corrupted by noise from frequencies other than the true embedded heart rate. This noise was eliminated by learning the passband frequencies, or, in other words, using a narrow-band filter. At this point, we assumed that the signal-to-noise ratio in the initial magnified video was high and that most of the true heart rate signals had been captured. Therefore, performing a histogram analysis of the extracted heart rate returned an estimate of the actual range of heart rate for that specific time sequence. The 5% and 95% frequency values were assigned as the lower and upper cutoff frequencies for the updated bandpass filter. An analysis of the MIT-BIH arrhythmia dataset (Goldberger et al., 2000) indicates that instantaneous human heart rate may vary within a range of 20–50 bpm, which justified our new frequency band choice. Once the passband was defined, we ran the

VidMag algorithm again with the new bandpass filter. In a more challenging scenario, this could be performed iteratively until convergence. Table 4 shows the algorithm used to find the final narrow bandpass frequencies.

Table 4. Algorithm 1 – Select narrow bandpass frequencies.

Step	Action
1	Select a face video.
2	Initialize passband frequencies, $f_l = \frac{2}{3}\ Hz, f_h = 4\ Hz$.
3	Perform VidMag frequencies, f_l, f_h.
4	Calculate instantaneous heart rate.
5	Perform cumulative histogram analysis.
6	Assign new passband frequencies, f_l, f_h, from 5th and 95th percentile value from Step 5.
7	Repeat Step 3–Step 6 until convergence.

Savitzky-Golay (S-G) Filter Analysis

Once the frequency range of the passband filter was determined, the raw signal from the magnified video was extracted and heart rate was estimated for every single beat. The blue line in Figure 22 shows the raw output of the extracted heart rate. Clearly, there is a considerable mismatch between the output and the ground truth measurements (green line) from the pulse meter. This is often the result of noise from sampling and peak detection error. To minimize this error, we used an S-G filter (Press, 2007). Human heartbeats do not appear at a constant interval, and therefore heartbeat data are a classic example of non-uniformly sampled data. To address this, the S-G filter, which uses a locally concentrated least square method, was applied to smooth the raw data. Figure 23 shows the filtered result (HR_{vm}) being very close to the ground truth (HR_{gt}). To get the average heartrate (HR_{avg}), we used a moving average filter with a window of 5 seconds.

$$HR_{avg}(i) = \frac{1}{2n_i+1}\sum_{k=-n_i}^{n_i} HR_{vm}(i+k)$$

Figure 24 shows the ground truth average and the HR_{avg}.

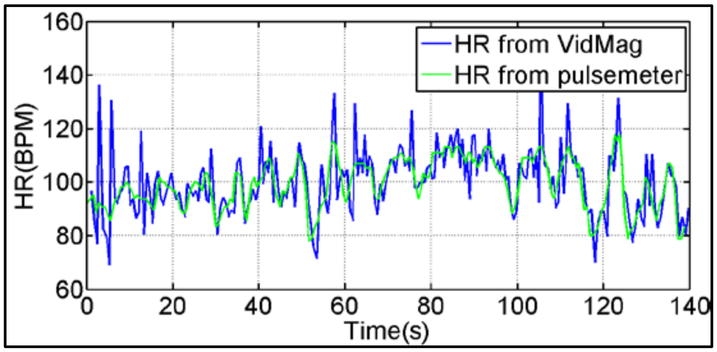


Figure 22. Graph. Raw heart rate estimation from VidMag.

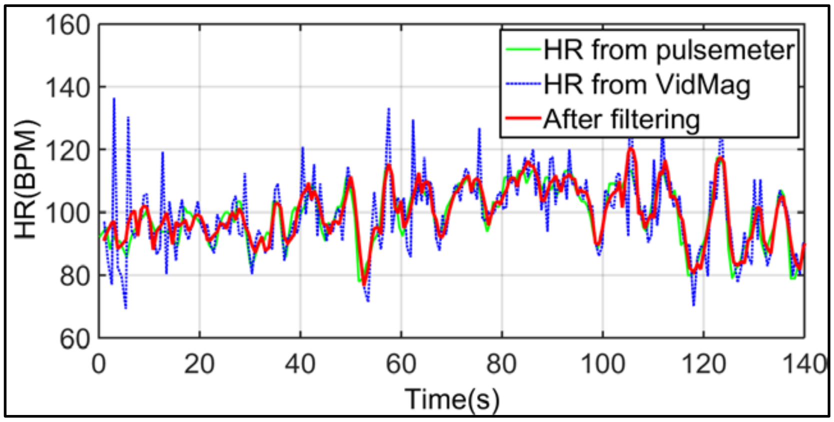
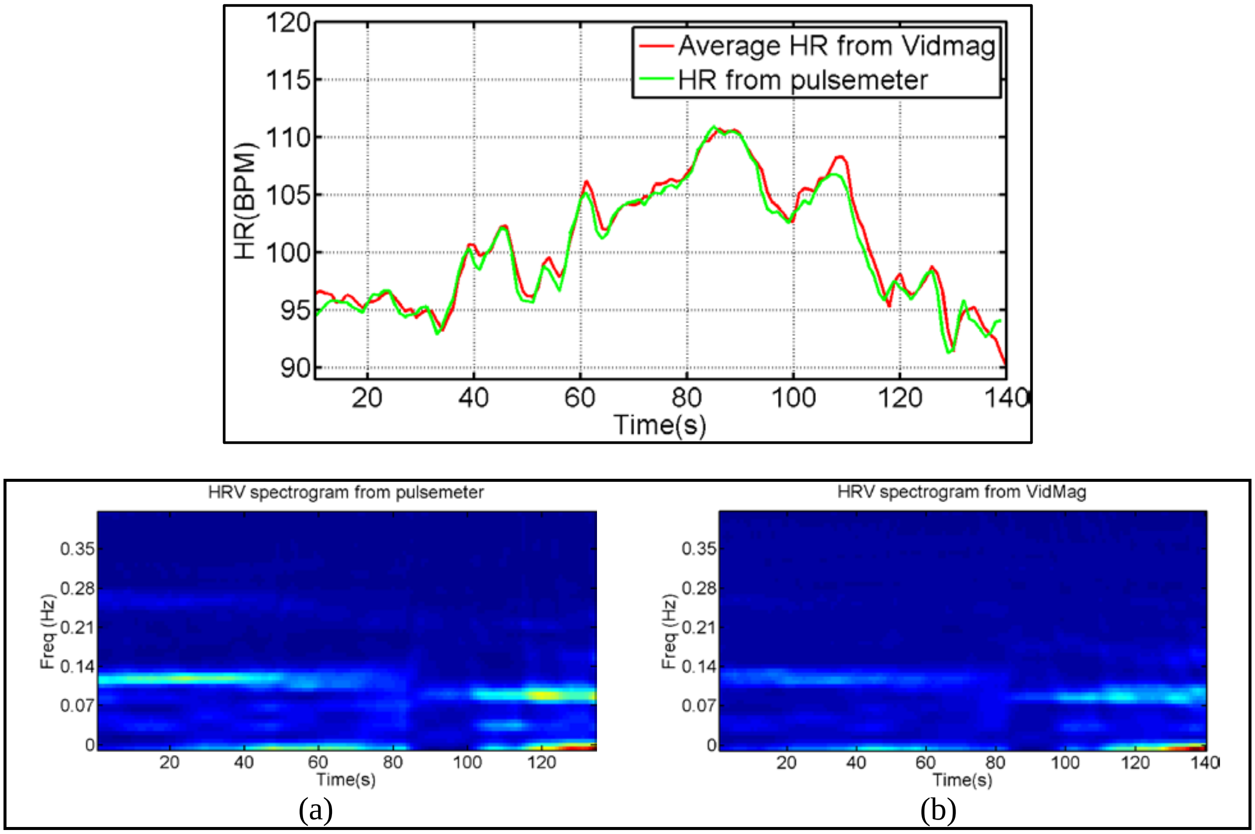


Figure 23. Graph. Heart rate after filtering.



(a)

(b)

Figure 24. Graphs. Comparison for average of heart rate. The heart rate has been averaged for every 5 seconds.

Lomb-Scargle Periodogram

Finally, we compared our results in the frequency domain. The sampling rate of the heart rate signal and the R-R intervals are non-uniform in nature. Therefore, standard frequency analysis methods using a standard periodogram often fail to provide a correct analysis. If there is any discrete time series data

$$\boldsymbol{x}(\boldsymbol{t}) = \{x(t_i)\}, i = 1,2, \dots N$$

then the classical periodogram calculates the spectral power as

$$P_{\boldsymbol{x}}(\omega) = \frac{1}{N}\left|\sum_{i=1}^{N} x(t_i)e^{-i\omega t_i}\right|^2, \qquad \omega = 2\pi f$$

When the samples are equispaced, then the fast Fourier transform (FFT) algorithm can easily calculate the spectral power at each frequency component. But, when the samples are taken at non-uniform time intervals, this algorithm does not apply. Resampling the data to uniform sampling is a simplified solution to this problem, but has the accompanying risk of losing data as well incorporating interpolation error. Lomb (1976) and Scargle (1982) address this problem by introducing a least square spectral analysis method:

$$P_{Lomb}(\omega) = \frac{1}{2\sigma^2}\left[\frac{\left(\sum_i (x_i - \bar{x})cos\left(\omega(t_i - \tau)\right)\right)^2}{\sum_i \cos^2\left(\omega(t_i - \tau)\right)} + \frac{\left(\sum_i (x_i - \bar{x})sin\left(\omega(t_i - \tau)\right)\right)^2}{\sum_i \sin^2\left(\omega(t_i - \tau)\right)}\right]$$

where

$$tan(2\omega\tau) = \frac{\sum_i sin(2\omega t_i)}{\sum_i cos(2\omega t_i)},$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \qquad \sigma = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2, \quad x_i = x(t_i).$$

The parameter τ is a time delay that Scargle introduced. He proved that the τ is selected by the algorithm in such a way that it makes the expression of $P_{Lomb}(\omega)$ equivalent to estimating the harmonic content at any angular frequency ω by least square fit to the model:

$$f(t, \omega) = Acos\omega t + Bsin\omega t$$

This makes the algorithm a better choice than conventional FFT methods for any non-uniformly sampled data. There is an emphasis on each point, so all observation points are used for the periodogram. Laguna, Moody, and Mark (1998) have shown the usefulness of the Lomb-Scargle method for analysis of the heart rate. We used the Lomb-Scargle periodogram to compare our results to the ground truth, and Figure 25 shows that our results closely matched the reference heart rate. Although there was some noise in the higher range of the periodogram, most of the high-power frequency components were present. We further analyzed the whole sequence of heartbeats with a spectrogram analysis. We selected a small window (T_w) in the given time and performed the Lomb-Scargle periodogram, then shifted the window by T_{sh} and performed it again. We used $T_w = 30$ sec and $T_{sh} = 1$ sec as the measurements. Figure 26 illustrates the comparison in the frequency domain when we used the Lomb-Scargle periodogram to compare power spectral density of the instantaneous heart rate values from VidMag and the ground truth measurement, indicating that the two show similar behavior.

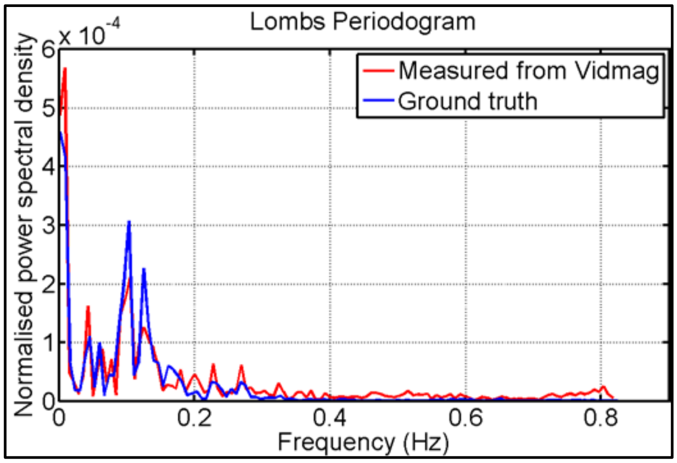


Figure 25. Graph. Comparison of instantaneous heart rate in frequency domain. The power spectral density is computed using the Lomb-Scargle periodogram.

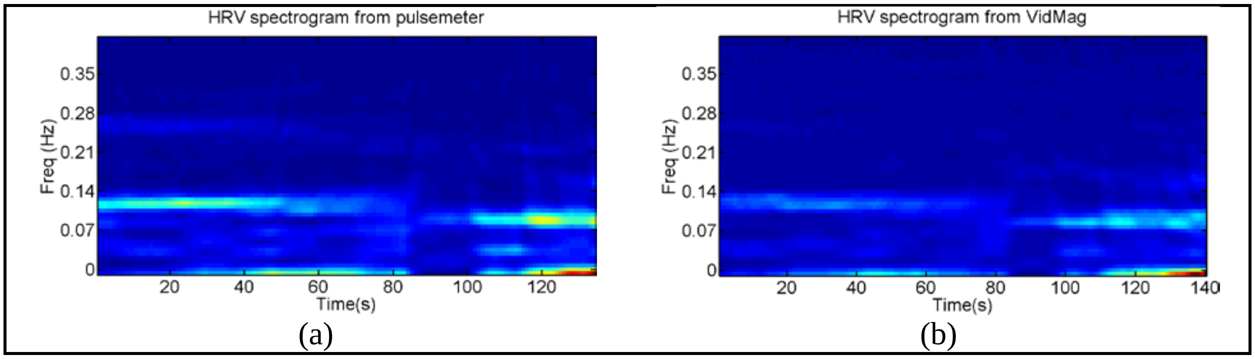


Figure 26. Graph. Comparison of VidMag algorithm by spectrogram analysis: (a) shows the spectrogram from the pulsemeter and (b) shows the results from the VidMag algorithm.

Error Analysis

In this section, we compare the ground truth measurement of the pulse rate with the final estimation of the pulse rate. We performed our analysis in both the time domain and the frequency domain. Figure 27 shows a scatter plot of the measured heart rate (HR_{vm}) against the reference ground truth (HR_{gt}) with a good correlation at 88.4%. The graphs in Figure 28 show that we can achieve an error rate of less than 2.84 bpm for the average value of the heart rate and an average error rate of less than 9.07 bpm for the instantaneous heart rate with a success rate of 95%.

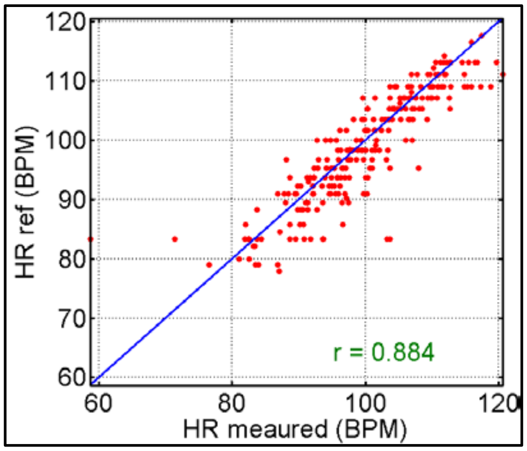


Figure 27. Graph. Comparison of instantaneous heart rate.

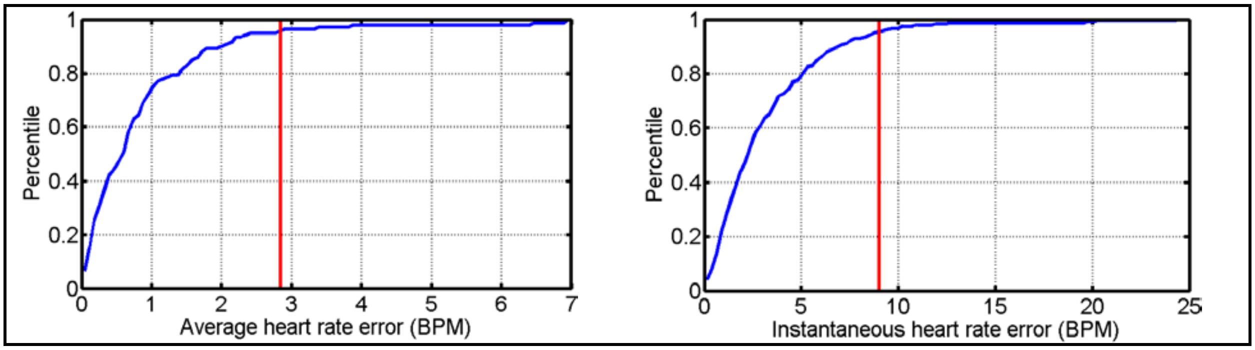


Figure 28. Graphs. Error analysis of heart rate.

CHAPTER 3. DATA COLLECTION

We collected data both for static subjects in indoor conditions as well as for subjects outside in vehicles. The data were used to tune and evaluate the video-based heart rate measurement methods described previously. In this section, we describe the collection methods and scenarios.

PHASE I – STATIC CONDITIONS

For initial validation of the VidMag algorithm, we ran an experiment in static indoor conditions. Reference heart rate data were collected using a Garmin Forerunner 620 wrist watch with a Garmin heart rate monitor (HRM; Figure 29). The HRM was chest wearable and transmitted heart rate information wirelessly to the Forerunner wrist watch. It provided a time-averaged heart rate for the wearer, and recorded a value once every second. In this case, three participants were asked to stand with minimal head movement, and videos were captured using a commercially available Canon PowerShot SX 120 IS that recorded RGB videos at 30 fps. Each participant was asked to rest before data collection, so that cognitive loads would be low. Each session was recorded for 60 seconds. No special control was imposed on the illumination condition.

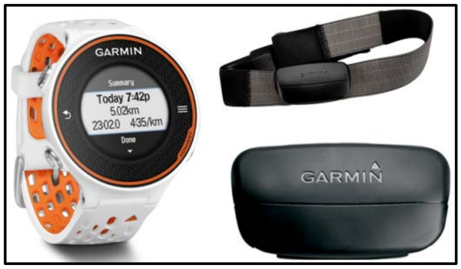


Figure 29. Photo. Garmin wrist watch and HRM. (Image source: http://www.garmin.com/en-US).

PHASE II – DYNAMIC CONDITIONS

The second set of data was recorded from the MiniDAS, a data acquisition system that was developed in-house at VTTI, which was specially designed for collecting in-vehicle data. Figure 30 shows a forward-view camera screenshot and a face-view camera screenshot of the collected video. The face-view camera was rigidly mounted near the rearview mirror. The system collected RGB face video data at a rate of 10 fps. A total of 2.5 hours of video was recorded for two participants. The Garmin heart rate data and video data were synchronized through Global Positioning System (GPS) time. Figure 31 shows cropped screenshots of the face-view video. The collected data included diverse situations, including low and high illumination, varying head poses, and normal and near-infrared videos.

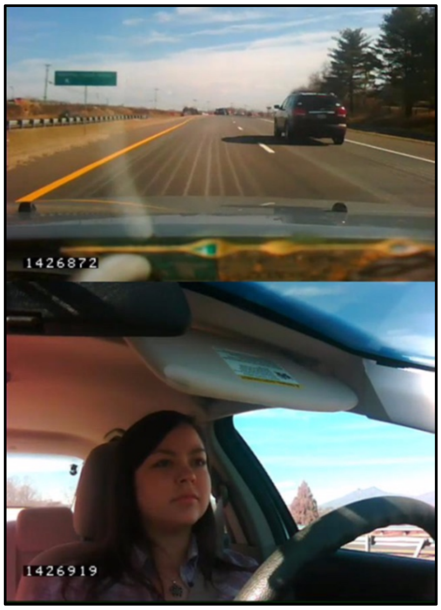


Figure 30. Screenshots. Example of collected video.

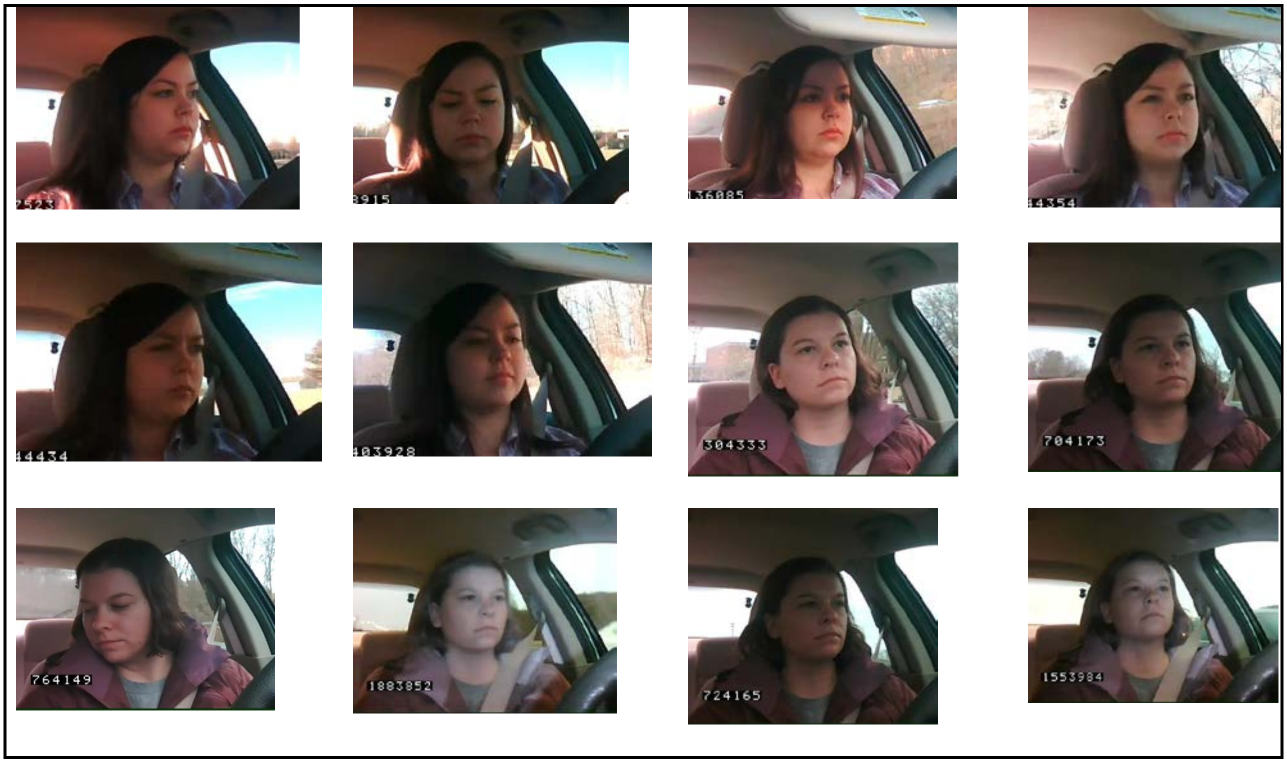


Figure 31. Screenshots. Sample screens from the collected video data using MiniDAS.

RAW-DATA ANALYSIS

Manual video analysis was performed to study drivers' behavior in naturalistic conditions. We manually examined each of the videos to see how drivers behaved on different road types, including city and interstate. Figure 32 shows the heart rate data reading for a session of driving, which can easily be divided into three segments.

1. In Segments A1, A2, and B, the participant was driving in a metropolitan area.
2. During Segment C, the participant was driving on an open highway with less traffic.
3. Segment B shows a sudden increase in the heart rate during the performance of a secondary task (searching for an item inside the vehicle) while continuing to drive.

A similar boost in the heart rate data occurred for the second participant when they suddenly had to brake hard at a traffic signal (Figure 33). These examples indicate that heart-rate analysis may have the potential to produce valuable information about driver state in naturalistic conditions. From the overall dataset, we selected some particular samples from the trip to test with the VidMag algorithm.

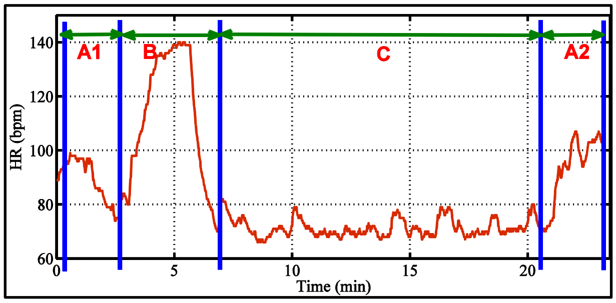


Figure 32. Graph. Sample run using Garmin HRM: Participant 1 – (A1 and A2 – inside city, B – inside city secondary task and extra cognitive load, C – interstate driving – less cognitive load).

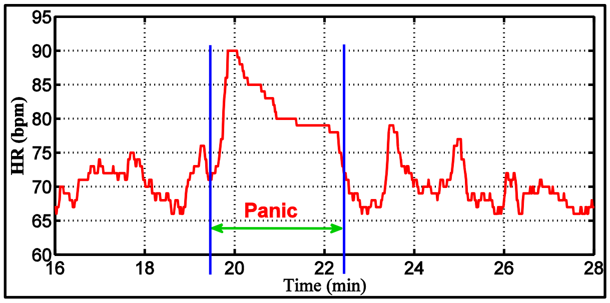


Figure 33. Graph. Sample run using Garmin HRM: Participant 2 – an instance of panic during hard braking.

EXPERIMENTS AND RESULTS

Figure 34 shows the result for a 1-minute video of a stationary subject indicating that the estimated heart rate matches the measured average heart rate value from the Garmin HRM reasonably well. The estimated heart rate has some outliers due to multimodal peaks in several situations.

Next, we chose a video sequence from our naturalistic video dataset, so that the head movement was within specified limits. Figure 35 shows the result for a participant driving on an interstate highway (Segment C from Figure 32). Figure 36 shows heart rate estimation for Segment B, where the driver had diverted attention due to the in-vehicle secondary task, as well as a high cognitive load due to driving in the city. In all cases, the results show good matches between the estimation and the ground truth. The results also indicate that the human heart rate can vary within a band of ±10 bpm. This variability corresponds well with results that we obtained with sample ECG signals from MIT-BIH (Goldberger et al., 2000).

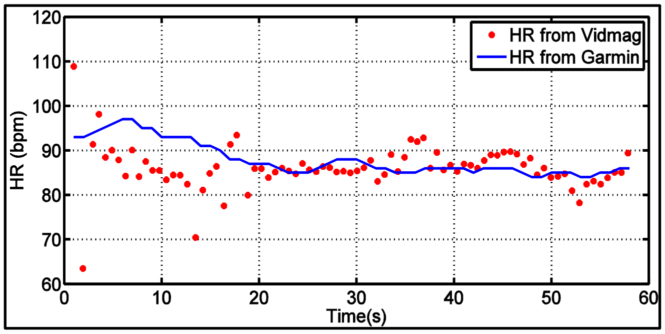


Figure 34. Graph. Heart rate estimation using the new procedure for a subject indoors.

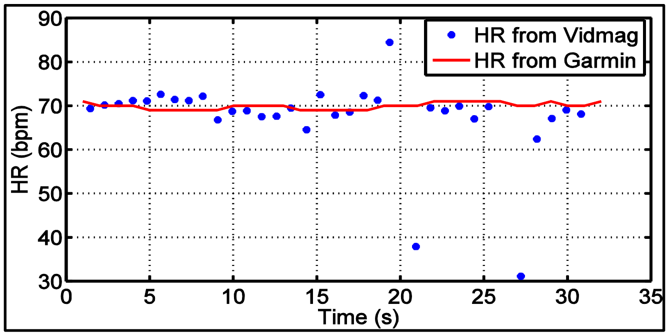


Figure 35. Graph. Heart rate estimation using the new procedure for video obtained while the subject was driving.

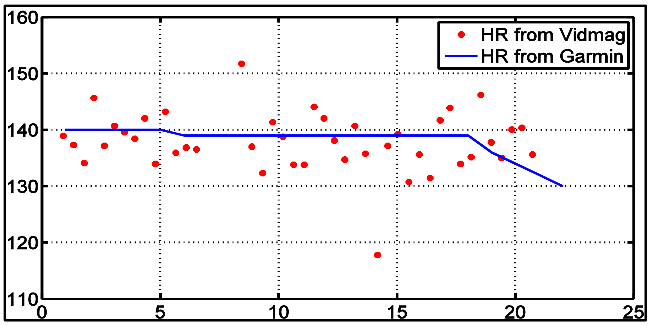


Figure 36. Graph. Heart rate estimation using the new procedure. The subject was performing a secondary task while driving in city conditions. This corresponds to Segment B in Figure 32.

HEART RATE VARIABILITY (HRV) ANALYSIS

The last step of the process was to analyze HRV in different situations in order to determine different psychophysiological conditions. HRV is generally computed from consecutive R-R intervals of a cardiac signal. Acharya et al. (2006) summarized different time domain, frequency domain, and nonlinear analyses of HRV from R-R intervals. We analyzed Segments B and C from Figure 32, first extracting the heart rate from the face video using the VidMag method as described in the previous sections, and then analyzing those segments using Kubios, a Matlab-based software for HRV analysis (Tarvainen, Niskanen, Lipponen, Ranta-Aho, & Karjalainen, 2009). As mentioned earlier, Segment B corresponds to a high cognitive load for the driver, and Segment C corresponds to interstate driving, where the cognitive load was comparatively low. Table 5 and Figure 37, respectively, show time domain analysis and frequency domain analysis using Welch's periodogram for Segment C. We compared this to results from Segment B in Table 6 and Figure 38. The comparison of mean heart rate (or R-R) values from the time domain analysis shows that the driver had a higher heart rate than normal during Segment B, as expected. The ratio of the areas under the power spectral density curves relates to the stress level (Healey & Picard, 2005). The red, blue, and yellow areas in Figure 37 correspond to very low frequency (VLF), low frequency (LF), and high frequency (HF) components of the spectrum, respectively. The higher the LF/HF ratio, the higher the cognitive load. Our experiment shows that Segment B had an LF/HF = 0.789, which is higher compared to Segment C (LF/HF = 0.688). This agrees with Healey and Picard's (2005) findings and validates our experiment.

Table 5. Time domain analysis of HRV for Segment C.

Variable	Units	Value
Mean RR	ms	849.0
STD RR (SDNN)	ms	53.3
Mean HR	min^{-1}	70.97
STD HR	min^{-1}	4.88
RMSSD	ms	57.5
NN50	$count$	7
pNN50	%	25
RR Triangular index		7.250
TINN	ms	175

Table 6. Time domain analysis of HRV for Segment B.

Variable	Units	Value
Mean RR	ms	436.4
STD RR (SDNN)	ms	17.8
Mean heart rate	min^{-1}	137.71
STD heart rate	min^{-1}	5.42
RMSSD	ms	26.6
NN50	$count$	9
pNN50	%	7.2
RR Triangular index		5.040
TINN	ms	90

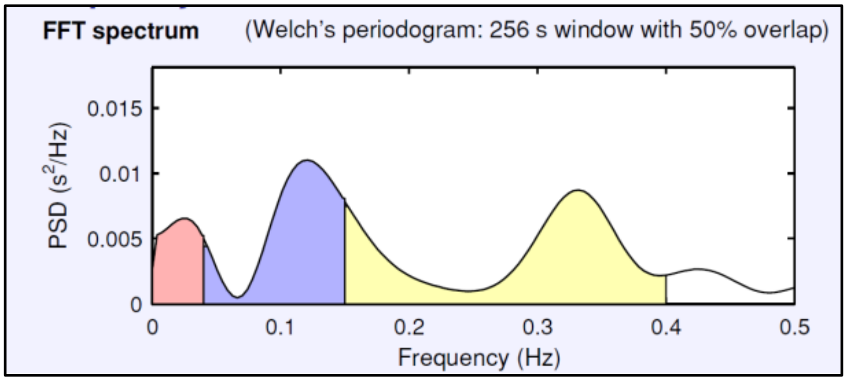


Figure 37. Graph. Frequency domain analysis of Segment C.

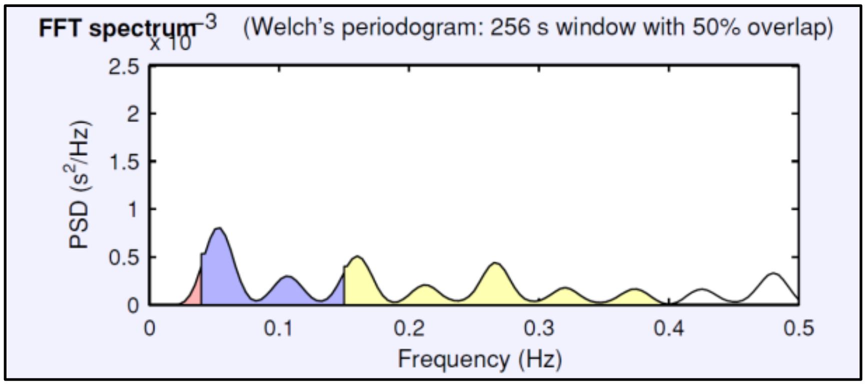


Figure 38. Graph. Frequency domain analysis of Segment B.

CHAPTER 4. DISCUSSION – CHALLENGES AND CONSIDERATIONS FOR FUTURE WORKS

This study demonstrated the feasibility of non-contact heart rate monitoring in naturalistic scenarios. The method used in this study was the VidMag algorithm, which measures BVP from face video data by exaggerating small, nearly imperceptible changes over time in pixel intensities. The use of the VidMag algorithm in real-life situations is challenging due to head movement, changes in illumination, low spatial resolution, low frame rate, and image noise. This report introduced a novel approach to addressing several of these limitations for the purpose of continuous heart-rate monitoring. Small movements of the head, for example, can be compensated for by careful registration of images over time.

One of the main contributions of this work is the introduction of an adaptive filtering technique that extracts heart rate from the video at a low frame rate. This novel approach suppresses noise and performs peak detection from magnified video, which relates to the R-R intervals in an ECG. The technique was tested for subjects in controlled conditions as well as for subjects driving under naturalistic conditions. This approach has the potential to extract information related to the psychophysiological state of a person in real time using HRV, and could perhaps be used for biometric authentication. The work done for this study demonstrates the feasibility of stress detection in real-life situations using time-domain and frequency-domain analysis. We believe that the use of advanced vision algorithms and signal processing methods and better quality video can further improve performance.

The preceding results included discussion and some concluding remarks; however, the work herein was largely exploratory in nature and resulted in a proof of concept. Along with these contributions and advancements, this work has given rise to a number of concerns that must be addressed before the proposed method can be successfully implemented in real-time cases. In the following section we will discuss some of the cases where our investigation did not reach a strong feasibility conclusion and raised questions for future work.

HRV ANALYSIS FROM BVP SIGNAL: AN ANALYSIS

In this work, we analyzed HRV, where the heart rate is measured from the BVP. Historically, HRV has been defined for the heart rate as extracted from an ECG signal. Therefore, a question still remains as to whether the HRV analysis from the BVP signal analyzed herein is equivalent to an HRV analysis from the conventional (and more thoroughly studied) ECG signal. An ECG signal is the reflection of the electrical behavior of the heart, whereas any BVP measurement, such as PPG, is an indication of the flow of blood volume in the periphery as a result of the heart's mechanical action. If we define the peak position from the ECG and PPG to be RR_{ECG} and RR_{PPG} and the time instant of the peaks to be $T_{RR_{ECG}}$ and $T_{RR_{PPG}}$ then for a particular beat occurrence, $T_{RR_{ECG}} - T_{RR_{PPG}} = T_{delay}$, where T_{delay} primarily indicates the total time delay for the blood to reach the measuring part of the PPG (often the fingertip) after the electrical signal for the beat is initialized. Even if we ignore the delay induced by the measuring instruments and other noise sources for simplicity, the mechanical delay may not be constant. Therefore, the distance between two consecutive heart rates are not always the same for an ECG and a PPG. Various studies (Selvaraj, Jaryal, Santhosh, Deepak, & Anand, 2008; Schäfer & Vagedes, 2013) over the last three decades indicate that most researchers agree that the behavior of HVR from both sources shows good correlation for time domain, frequency domain, and nonlinear analysis. Only a limited number of researchers have shown that, except from the resting position measurements, the short-term HRV may vary from one instance to the next (Khandoker, Karmakar, & Palaniswami, 2011).

FACIAL MAP FOR HEART RATE MEASUREMENTS

Physiologically, not every skin patch on the face provides the same amount of visual information for peripheral blood flow. Due to the variation of the density of the peripheral blood vessels, the amplitude of the extracted PPG signal will vary, and given a constant noise level, lower amplitude PPG signals are more prone to be corrupted by noise. In addition, blood flow due to the cardiovascular pulse does not reach each part of the face at the same time, which may introduce delay in the extracted PPG signals from each of the patches and influence the resulting heart rate measures in unknown ways. To have a deeper understanding of these variables and create a reliable weight map for refining the algorithms, future research projects should employ very high resolution video data.

ILLUMINATION

Challenges

This work made clear that one of the major problems with using naturalistic videos in conjunction with machine vision are varying illumination conditions. The problem is difficult to overcome with existing machine vision techniques, and is important for several reasons:

1. During the daytime, most illumination is from the sunlight. Because no road is exactly straight, the incident angle of the sun changes every moment. As a result, the amount of light appearing on the face varies.

2. The differences in the interior design and color of the vehicle pose another challenge. Ambient light generally enters though windows or the windshield and gets reflected from the interior parts of the car. This reflection affects the light rays that appear on the driver’s face.

3. A similar effect may be observed from the light reflected from the surroundings. For instance, if a red car passes by, its reflection will affect the total volume flow of the light field inside the vehicle and thus on the driver’s face.

4. The topography of the road plays a major role in the illumination condition. If the car is going through a forest or road with trees, then the perforated nature of the trees make the light field highly variable.

5. Lastly, shadow plays a vital role in illumination, as it results in the various points on the face not being illuminated uniformly. Hard shadows or an extremely bright background cause problems for the face detector as well as relative weight for every face patch, which determines which part of the face and the skin patches should get more priority. These need to be monitored and adjust constantly.

Figure 39 shows the luminance map of a frame where the gradient and variability is clearly visible. Given the challenges discussed above, we propose several mitigation strategies.

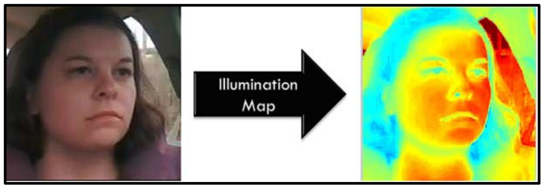


Figure 39. Diagram. Illumination gradient on the face.

Use of Vehicle's Stationary Parts for Illumination Correction

The interior body of the car visible to the camera is mostly stationary. The vibration of the camera due to speed and road conditions may lead to slight changes in the 2D location of a fixed point in the automobile, but for the purposes of this discussion, we will discard the effect of vibration. The movement of the driver may also occlude certain portions of the interior from time to time. Figure 40 shows an example of the interior of the car cropped from the actual image. Although it is generally assumed that the color and appearance of the interior will be uniform, the image depicts the variability discussed in the previous section. To provide an example of a means to address this, we took the driver side part of the interior and the parts closer to the driver (driver side window, back of mirror, etc.) and divided them into small patches. Next, the luminance part of each of the patches was tracked by converting the RGB image to the CIELUV color space. Generally, the "L" channel of the CIELUV color space captures the luminance. The luminance map from each point is mapped in Figure 41, where it is plotted against time in the *x*-axis and patch number in the *y*-axis. Ideally, there should be a correlation both spatially and temporally. The figure shows a pattern in the signal, although the patterns are not consistent for all the patches. Tracking background pixels from a fixed part of the car may be a useful way to eliminate the effect of illumination to a certain extent. This should be further explored in future work.

Figure 40. Screenshot. Stationary part of car interior.

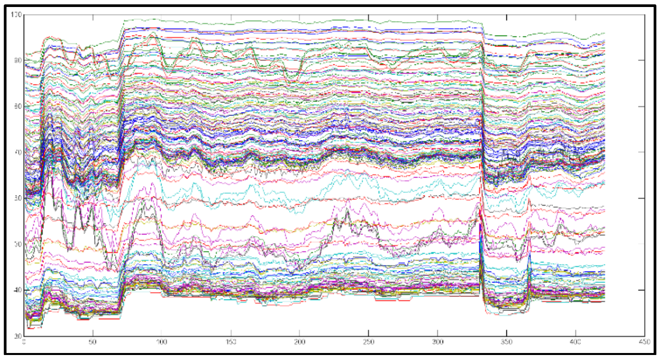


Figure 41. Graph. Interior illumination variation with time.

USE OF LAMBERTIAN SURFACE

The appearance of any surface is largely dependent on the ambient illumination, the reflectance of the surface, and the camera parameters. Recovering all of these parameters has been an active area of research in the field of face recognition and photo-realistic face generation. In addition, practical applications often require understanding the ambient lighting condition to optimally model the face parameters. We propose assuming that the human face is a Lambertian surface and modeling the problem as a Markov Random Field (MRF) to simultaneously find the texture, albedo, and the ambient illumination.

There have been a number of works related to relighting, shape, and texture in determining faces under various lighting conditions (Shan, Gao, Cao, & Zhao, 2003; Shim & Chen, 2008; Wang et al., 2007; Wang et al., 2009; Atick, Griffin, & Reclich, 1996), but the research by Wang et al. (2007; 2009) provides a systematic method for modeling the face albedo, texture, and illumination in a spatially connected model. Using spherical harmonic representation (Basri & Jacobs, 2003), it has been shown that the set of images of a convex Lambertian object obtained under a wide variety of lighting conditions can be approximated by a low-dimensional (nine or four) linear subspace. The model divides the face image into smaller regions and uses a different set of face model parameters for each region. This reduces the overall estimation error more than a single holistic approximation, and introduces neighboring coherence constraints to the albedo estimation for the divided regions. The energy minimization problem for a single image is formed as:

$$\begin{aligned}\arg\min_{\rho,\lambda,\beta,l} \quad & \sum_{q=1}^{Q} \sum_{(u,v)\in\Omega_q} \{W_{u,v}(I_{u,v} - \rho_{u,v}\sum_{i=1}^{9} h_i(\vec{n}^{M}_{u,v})l_i)^2 + W_{MM}(\rho_{u,v} - \rho^{q}_{u,v})^2\} \\ & + W_{SM}N_{sr} \sum_{(i,j)\in\mathcal{N}} \sum_{k=1}^{m-1} (\frac{\beta^i_k - \beta^j_k}{\sigma^{ij}_k})^2\end{aligned}$$

where ρ is the output albedo, (u, v) is the pixel index, l is the illumination, q denotes the q^{th} region, $\mathrm{N} = \{(i, j)|\Omega_{\mathrm{i}}\ and\ \Omega_j\ are\ neighbors\}$ is the set of all pairs of neighboring regions, $\vec{n}^{M}$ is constrained by the shape subspace normal, and ρ^{q} is constrained by the texture subspace (Wang, 2007). The objective function is an energy function of an MRF. The first two terms in the equation are the first-order potentials corresponding to the likelihood of the observation data given the model parameters, and the third term is the second-order potential, which models the spatial dependence between neighboring regions. Therefore, the problem of jointly recovering the shape, texture, and lighting of an input face image is formulated as an MRF-based energy minimization (or maximum *a posteriori*) problem. In the case of multiple frames, this can be extended with a minimization in the variance of the albedo and texture, given that the face belongs to the same person. Wang et al. (2009) approached the optimization using an iterative process. We propose future work to investigate the problem from the perspective of a graphical model and exploring relevant algorithms for optimization of similar problems and then making a comparison in relation to the given scenario. This will include the 3D structure of the face in terms of the learned surface normal and counter the effect of illumination through the training on the subject's natural albedo and ambient lighting. Most remote measurements of BVP assume the face to be a Lambertian surface (de Haan & Jeanne, 2013; van Gastel, Stuijk, & de Haan, 2015), and therefore any standard model can be merged with this lighting model.

PROBLEMS RELATED TO NATURALISTIC DRIVING VIDEOS AND SHRP 2

In addition to the data collections discussed thus far, we also manually reduced more than 100 videos of crash and near-crash scenarios from the SHRP 2 NDS (Campbell, 2012). A summary of the reduction is shown in Appendix B. During this explorative process, we noticed several practical problems that may limit the scope of the VidMag algorithm. Some of the SHRP 2 face video data uses automatic zooming on the face, causing a loss of continuity in the vision-based detection process. Additionally, sometimes the camera's field of view is directed to the sun, rendering the face barely visible due to the camera's automatic gain control and white balance.

Another major problem was the quality of nighttime video. Due to the low light conditions, the video frames have high spatial and temporal noise. Sometimes face detection is beyond human capabilities. The signal-to-noise ratio is also quite low in some particular video frames. Figure 42 shows one such example. If we crop everything out but the face, it is very difficult to recognize the resulting image as a person, and impossible to point out different parts of the face. The available face detectors often fail for these kind of video frames. The current quality of nighttime videos along with their low resolution makes it highly challenging to apply any remote measurement of physiological signal.

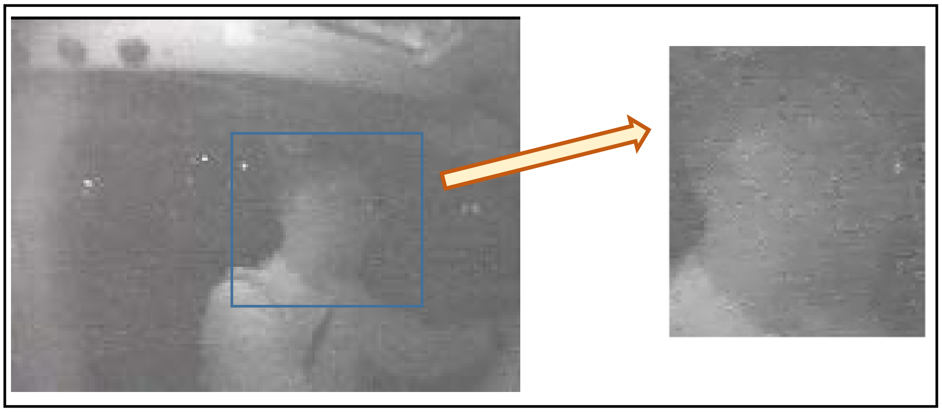


Figure 42. Screenshot. Face detection problem in SHRP 2 video.

We also noticed that in some cases the camera angle does not record the face, but is rather oriented in some other direction. Additionally, occlusions sometimes hide faces or the presence of the sun in the background creates problems with the white balance of the frame. The current algorithm is insufficient to overcome all of these practical issues with true naturalistic data.

As discussed previously, we observed a case study where a rising heart rate indicated interesting behavior (hard braking; Figure 33). In reviewing other video data from SHRP 2, we observed that, often, after a critical event, the driver moves out of the vehicle or moves in a way that occludes the face (e.g., puts the hands over the face area). This may limit the algorithm's utility for certain segments of data that are of particular interest and require additional research to understand their full implications.

AUTHENTIC DATA FOR BEHAVIOR ANALYSIS

We have discussed how physiological measurements are capable of indicating a person's behavioral characteristics. A substantial amount of work has been done in the last few decades to demonstrate these measures' effectiveness under controlled conditions. The study of a driver's behavioral patterns is not new either. However, at this point in time, the driving research community does not have a good set of reliable data that may validate these theories and allow for testing of the algorithms described herein. We believe that if a set of data, with simultaneous recording of face video and physiological variables (i.e., ground truth), were collected for drivers with events like drowsiness or alcohol consumption, it may be possible to come to some conclusions about the relationship among these variables and behaviors that would help justify moving toward increasingly refined and deployment-ready, non-obtrusive cardiovascular measurement systems within vehicles.

RECENT DEVELOPMENTS IN THE FIELD OF REMOTE MEASUREMENTS OF HEART RATE

Recently, Hernandez, McDuff, and Picard (2015) showed that it is possible, using a wrist wearable watch, to measure cardiac and respiratory parameters with a motion sensor and gyroscope. Similar measurement is possible via wearable glasses or any head-mounted motion sensor (Hernandez & Picard, 2014). It has also been reported by McDuff, Gontarek, and Picard

(2014b) that characteristics of systolic and diastolic peaks of pulse signal can be detected from face video. Another recent work showed that the microvascular blood flow can be monitored using a vision system (Zhang et al., 2016). This can show the structure of the vessels and the flow characteristics of the blood inside the vessels. All of these works have shown substantial promise over recent years, and we believe that with good quality video, these measurements are possible in the future and widely implementable in real-life conditions. Future research should consider the availability and applicability of these potentially overlapping methods for capturing biometric information.

CONCLUSION

In this work, we have demonstrated the feasibility of non-contact monitoring of heart rate in naturalistic situations using RGB face video data. We have used a “video magnification” approach, which exaggerates small, nearly imperceptible changes of blood volume over time reflected in pixel intensities. The use of the video magnification algorithm in real-life situations is challenging due to head movement, changes in illumination, low spatial resolution, low frame rate, and image noise. This work has shown that a step-by-step approach to addressing several of these limitations for the purpose of heart-rate monitoring is effective. With our algorithm, we can achieve very good estimation of instantaneous heart rate and average heart rate. Both of these measurements are useful to understand the psychophysiological condition of a driver. This includes stress, panic, fatigue, and drowsiness. We used two sets of data to demonstrate our results. First, we used an indoor dataset with very restricted conditions. Second, we used a naturalistic driving dataset. In both cases the heart rate estimation results were promising. The estimation of heart rate information is more challenging in real-life driving scenarios, but, if we carefully select the frame sequence where the change in illumination is restricted, we can predict panic using frequency domain analysis of heart rate variability. Despite several practical limitations, this work shows the possibility of nonintrusive heart rate measurements for developing driver state monitoring tools in naturalistic scenarios.

APPENDIX A. AVAILABLE FACE DETECTORS

Online	Freeware	Commercial
BioID	Flandmark	BetaFace
BetaFace	Semantic	CogniTec
LambdaLab	fdlib	KeyLemon
Face ++	OpenCV – numerous	Luxand
	Visage	NeuroTechnology
	FraunHofer	Tastenkunst
	SkyBio	Visage
	ImageVision	Facebook
	And more including recent work	

APPENDIX B. CRASH – SHRP 2

Event ID	File ID	Start time	End time	Event status ID	Event type ID	Note
30881809	3917717	363677	393677	3	181	Minor
35257227	5008191	61365	91365	3	181	forward crash but face not visible
61426936	5821251	1423976	1453976	3	181	Nothing
142007898	8076762	946369	976369	3	181	gets excited, moves out and came back, too much zoom out
135946520	12900554	201208	231208	3	181	Quickly moved out, too much zoom control
116163145	14647680	899985	929985	3	181	too much zoom control
132522489	18518468	272848	302848	3	181	minor event
29714720	32043038	2086804	2116804	3	181	not sure what happened, harsh light
151086604	37083609	113285	143285	3	181	moves out, car off
31281062	37366857	797471	827471	3	181	Harsh light
30879203	38179311	1077327	1107327	3	181	moves out, car off
142030120	40843113	950163	980163	3	181	moves out, car off
60633156	59011982	1648483	1678483	3	181	minor event

Event ID	File ID	Start time	End time	Event status ID	Event type ID	Note
33575087	62878789	554451	584451	3	181	Nothing
29877166	62938051	3415976	3445976	3	181	Nothing
151545522	65698175	1586094	1616094	3	181	Nothing
138605443	69649499	6605255	6635255	3	181	???
2934487	14219686	1649177	1679177	3	181	get out

REFERENCES

Acharya, U. R., Joseph, K. P., Kannathal, N., Lim, C. M., & Suri, J. S. (2006). Heart rate variability: A review. *Medical and Biological Engineering and Computing, 44*(12), 1031-1051.

Atick, J. J., Griffin, P. A., & Redlich, A. N. (1996). Statistical approach to shape from shading: Reconstruction of three-dimensional face surfaces from single two-dimensional images. *Neural Computation, 8*(6), 1321-1340.

Balakrishnan, G., Durand, F., & Guttag, J. (2013). Detecting pulse from head motions in video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3430-3437.

Basri, R., & Jacobs, D. W. (2003). Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(2), 218-233.

Campbell, K. L. (2012). The SHRP 2 naturalistic driving study: Addressing driver performance and behavior in traffic safety. *TR News*, *282*, 30-35.

Casati, J. P. B., Moraes, D. R., & Rodrigues, E. L. L. (2013). *SFA: A human skin image database based on FERET and AR facial images.* Paper presented at the IX workshop de Visao Computational, Rio de Janeiro.

Chai, D., & Ngan, K. N. (1999). Face segmentation using skin-color map in videophone applications. *IEEE Transactions on Circuits and Systems for Video Technology, 9*(4), 551-564.

Cula, O. G., Dana, K. J., Murphy, F. P., & Rao, B. K. (2005). Skin texture modeling. *International Journal of Computer Vision*, *62*(1-2), 97-119.

de Haan, G., & Jeanne, V. (2013). Robust pulse rate from chrominance-based rPPG. *IEEE Transactions on Biomedical Engineering, 60*(10), 2878-2886.

Drimbarean, A. F., Corcoran, P. M., Cuic, M., & Buzuloiu, V. (2001). *Image processing techniques to detect and filter objectionable images based on skin tone and shape recognition.* In *Proceedings of the International Conference on Consumer Electronics (ICCE),* 278-279.

Fleck, M. M., Forsyth, D. A., & Bregler, C. (1996). Finding naked people. In *European Conference on Computer Vision (ECCV'96*), 593-602.

Fotouhi, M., Rohban, M. H., & Kasaei, S. (2009). *Skin detection using contourlet-based texture analysis.* Paper presented at the Fourth International Conference on Digital Telecommunications.

Garcia, C., & Tziritas, G. (1999). Face detection using quantized skin color regions merging and wavelet packet analysis. *IEEE Transactions on Multimedia, 1*(3), 264-277.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., . . . & Stanley, H. E. (2000). Physiobank, physiotoolkit, and physionet components of a new research resource for complex physiologic signals. *Circulation, 101*(23), e215-e220.

Healey, J. A., & Picard, R. W. (2005). Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems, 6*(2), 156-166.

Hernandez, J., McDuff, D., & Picard, R. W. (2015). BioWatch: Estimation of heart and breathing rates from wrist motions. In *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare* (pp. 169-176). ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering).

Hernandez, J., & Picard, R. W. (2014). SenseGlass: Using Google Glass to sense daily emotions. In *Proceedings of the Adjunct Publication of the 27th Annual ACM Symposium on User Interface Software and Technology* (pp. 77-78).

Kakumanu, P., Makrogiannis, S., & Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern recognition, 40*(3), 1106-1122.

Khandoker, A. H., Karmakar, C. K., & Palaniswami, M. (2011). Comparison of pulse rate variability with heart rate variability during obstructive sleep apnea. *Medical Engineering & Physics, 33*(2), 204-209.

Laguna, P., Moody, G. B., & Mark, R. G. (1998). Power spectral density of unevenly sampled data by least-square analysis: Performance and application to heart rate signals. *IEEE Transactions on Biomedical Engineering, 45*(6), 698-715.

Lin, Y., Leng, H., Yang, G., & Cai, H. (2007). An intelligent noninvasive sensor for driver pulse wave measurement. *IEEE Sensors Journal, 7*(5), 790-799.

Liu, C. (2009). *Beyond pixels: Exploring new representations and applications for motion analysis* (Doctoral dissertation). Massachusetts Institute of Technology, Cambridge, MA.

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science, 39*(2), 447-462.

McDuff, D., Gontarek, S., & Picard, R. W. (2014a). Improvements in remote cardiopulmonary measurement using a five band digital camera. *IEEE Transactions on Biomedical Engineering, 61*(10), 2593-2601.

McDuff, D., Gontarek, S., & Picard, R. W. (2014b). Remote detection of photoplethysmographic systolic and diastolic peaks using a digital camera. *IEEE Transactions on Biomedical Engineering, 61*(12), 2948-2954.

National Highway Traffic Safety Administration. (2016). *2015 motor vehicle crashes: Overview* [Traffic Safety Facts Research Note] (DOT HS 812 318). Washington, DC: Author. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812318

Phung, S. L., Bouzerdoum, A., & Chai Sr, D. (2005). Skin segmentation using color pixel classification: Analysis and comparison. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(1), 148-154.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(10), 1175-1191.

Poh, M.-Z., McDuff, D. J., & Picard, R. W. (2011). Advancements in noncontact, multiparameter physiological measurements using a webcam. *IEEE Transactions on Biomedical Engineering, 58*(1), 7-11.

Poh, M.-Z., Swenson, N. C., & Picard, R. W. (2010). Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photoplethysmography. *IEEE Transactions on Information Technology in Biomedicine, 14*(3), 786-794.

Press, W. H. (2007). *Numerical recipes 3rd edition: The art of scientific computing*. New York: Cambridge University Press.

Rodriguez-Ibañez, N., García-Gonzalez, M. A., de la Cruz, M. A. F., Fernández-Chimeno, M., & Ramos-Castro, J. (2012). *Changes in heart rate variability indexes due to drowsiness in professional drivers measured in a real environment.* Paper presented at the Computing in Cardiology (CinC) 2012.

Scargle, J. D. (1982). Studies in astronomical time series analysis. II-Statistical aspects of spectral analysis of unevenly spaced data. *The Astrophysical Journal, 263,* 835-853.

Schäfer, A., & Vagedes, J. (2013). How accurate is pulse rate variability as an estimate of heart rate variability? A review on studies comparing photoplethysmographic technology with an electrocardiogram. *International Journal of Cardiology, 166*(1), 15-29.

Selvaraj, N., Jaryal, A., Santhosh, J., Deepak, K. K., & Anand, S. (2008). Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography. *Journal of Medical Engineering & Technology, 32*(6), 479-484.

Shan, S., Gao, W., Cao, B., & Zhao, D. (2003). Illumination normalization for robust face recognition against varying lighting conditions. *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2003)* (pp. 157-164).

Shim, H., Luo, J., & Chen, T. (2008). A subspace model-based approach to face relighting under unknown lighting and poses. *IEEE Transactions on Image Processing, 17*(8), 1331-1341.

Szeliski, R. (2006). Image alignment and stitching: A tutorial. *Foundations and Trends® in Computer Graphics and Vision, 2*(1), 1-104.

Tarvainen, M. P., Niskanen, J.-P., Lipponen, J., Ranta-Aho, P., & Karjalainen, P. (2009). *Kubios HRV—a software for advanced heart rate variability analysis.* Paper presented at the 4th

European Conference of the International Federation for Medical and Biological Engineering.

Tarvainen, M. P., Niskanen, J.-P., Lipponen, J. A., Ranta-Aho, P. O., & Karjalainen, P. A. (2014). Kubios HRV–heart rate variability analysis software. *Computer Methods and Programs in Biomedicine, 113*(1), 210-220.

Terrillon, J.-C., Shirazi, M. N., Fukamachi, H., & Akamatsu, S. (2000). Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. *Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, 2000* (pp. 54-61).

van Gastel, M., Stuijk, S., & de Haan, G. (2015). Motion robust remote-PPG in infrared. *IEEE Transactions on Biomedical Engineering, 62*(5), 1425-1433.

Vezhnevets, V., Sazonov, V., & Andreeva, A. (2003). A survey on pixel-based skin color detection techniques. *Proc. Graphicon, 3,* 85-92.

Wang, Y., Liu, Z., Hua, G., Wen, Z., Zhang, Z., & Samaras, D. (2007). *Face re-lighting from a single image under harsh lighting conditions.* Paper presented at the IEEE Conference on Computer Vision and Pattern Recognition, 2007 (CVPR'07).

Wang, Y., Zhang, L., Liu, Z., Hua, G., Wen, Z., Zhang, Z., & Samaras, D. (2009). Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(11), 1968-1984.

Wu, H.-Y. (2012). *Eulerian video processing and medical applications.* Cambridge, MA: Massachusetts Institute of Technology.

Wu, H.-Y., Rubinstein, M., Shih, E., Guttag, J., Durand, F., & Freeman, W. (2012). Eulerian video magnification for revealing subtle changes in the world. *ACM Transactions on Graphics (TOG), 31*(4), 65.

Xiong, X., & De la Torre, F. (2013). *Supervised descent method and its applications to face alignment.* Paper presented at the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Zhang, X. D., Wang, H., Antaris, A. L., Li, L., Diao, S., Ma, R., . . . Wang, J. (2016). Traumatic brain injury imaging in the second near-infrared window with a molecular fluorophore. *Advanced Materials, 28*(32), 6872-6879.

Zhu, X., & Ramanan, D. (2012). *Face detection, pose estimation, and landmark localization in the wild.* Paper presented at the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).