AN INVESTIGATION OF THERMAL IMAGING
TO DETECT PHYSIOLOGICAL INDICATORS
OF STRESS IN HUMANS

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Engineering

By

CARL BRADY CROSS
B.S., Wright State University, 2011

2013
Wright State University

WRIGHT STATE UNIVERSITY
SCHOOL OF GRADUATE STUDIES

April 25, 2012

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Carl Brady Cross ENTITLED An Investigation of Thermal Imaging to Detect Physiological Indicators of Stress in Humans BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Engineering.

Julie A. Skipper, Ph.D.
Thesis Advisor

Ping He, Ph.D., P.E.
Co-Thesis Advisor

Thomas N. Hangartner, Ph.D.
Chair, Department of Biomedical,
Industrial and Human Factors Engineering

Committee on Final Examination

Julie A. Skipper, Ph.D.

Ping He, Ph.D., P.E.

Doug T. Petkie, Ph.D.

R. William Ayres, Ph.D.
Interim Dean, Graduate School

Dedication

In memory of my brother, Chad Andrew Cross. For always being supportive, uplifting, and a personal inspiration.

Acknowledgements

I would like to thank my advisor, Dr. Julie Skipper, for her support, mentorship, and guidance. Her continuous strive for perfection drove me to put forth my full effort in every aspect of my work. I would also like to thank Dr. He and Dr. Petkie for kindly serving on my thesis committee.

I thank all of those who worked with me in this research for the provided advice, assistance, and passing of knowledge. I truly benefitted from having others around me sharing in the endeavor. For the financial support on this project, I owe thanks to the Center for Surveillance Research and all of the supporting members.

Finally, I thank my friends for their endless encouragement, and I thank my family for their loving support and praise.

Abstract

Cross, Carl Brady, M.S.Egr., Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, Dayton, OH, 2013. An Investigation of Thermal Imaging to Detect Physiological Indicators of Stress in Humans.

Real-time, stand-off sensing of humans to detect emotional state would be valuable in many defense, security and medical scenarios. Using a multimodal sensor platform that incorporates high-resolution visible-wavelength and mid-wave infrared cameras and a millimeter-wave (mmW) radar system, the detection of physiological indicators of psychological stress is tested through laboratory experiments. Our approach focuses on thermal imaging to measure temperature patterns in distinct facial regions representative of underlying hemodynamic patterns. Experiments were designed to: 1) determine the ability of thermal imaging to detect high levels of psychological stress and assess responses to physical versus psychological stressors; 2) evaluate the fidelity of vital signs extracted from thermal imagery and radar signatures; and 3) investigate the stability of thermal imaging under various confounding factors and real-world limitations.

To achieve the first objective, registered image and sensor data were collected as subjects (*n*=32) performed mental and physical tasks. In each image, the face was segmented into 29 non-overlapping segments based on fiducial points automatically output by our facial feature tracker. Image features were defined that facilitated discrimination between psychological and physical stress states. Four classifiers (artificial neural network, naïve Bayes, linear discriminant analysis, and support vector machine) were trained and tested, using a down-selected set of salient features, to evaluate efficacy under several classification paradigms. Very successful results were obtained. We were 100% accurate in classifying high mental stress, and nearly 99% accurate in the classification of mental versus physical stress.

The performance of two non-contact techniques to detect respiration and heart rate were evaluated: chest displacement extracted from the mmW radar signal and temperature fluctuations at the nose tip and regions near superficial arteries, extracted from the MWIR imagery, to detect respiration and heart rates, respectively. Results of the two techniques at stand-off distances of approximately six feet were similar. Estimated respiration rates from the MWIR imagery were accurate within one breath per minute for 72% of the samples and within two breaths per minute for 87% of the samples during baseline, and estimated heart rates were within two beats per minute for 27% of the samples and within six beats per minute for 53% of the samples.

The stability of the human thermal signatures was tested in additional experiments to investigate the effects of subject-to-imager distance, subject pose, normal day-to-day variability, facial muscle activation and topical skin products. Quantitative results for the effects of each of these factors on thermal signatures are provided. Autonomous face tracking was found to be successful when the subject-to-imager distance is 15 feet or less and pose angle is 30° or less. The perioptic region of the face exhibited the most stable thermal signatures, and the nose region was least stable. The confounding factors of local, forceful muscle activation and one skin product (liquid makeup) were determined to have a significant impact on measured values, each raising the skin temperature of specific facial regions up to 1.5°C. By characterizing the sensor suite under challenging conditions, its use for assessing human state in operationally feasible settings has been validated.

Table of Contents

List of Figures

List of Tables

1. Introduction

Over the past decade, security in many facilities has been dramatically increased for the protection of the people, information, and contents that may be at risk to acts of terrorism. Skilled professionals and screening systems have been implemented in an attempt to identify individuals who are potential threats. Observable cues, such as suspicious activity, behavior, body language, and gait, have been studied with an aim of detecting these individuals, but the efficacy of this approach can be thwarted by individuals who actively aim to conceal these indicators, allowing them to pass through undetected. Our aim is to provide a complementary approach to the screening process that keys on the autonomic nervous system's response to the expected feelings of stress and/or fear that are assumed to accompany individuals who are partaking in, or intending to perform illegal or terrorist activities. Such responses are much more difficult to intentionally manipulate, and they can be measured via physiologic sensing to accurately characterize emotional state. This is the basis for polygraph testing, where a number of physiological signals are recorded via contact sensors to evaluate a subject's truthfulness. In a surveillance setting which may exhibit high-volume traffic and a short amount of observation time, a polygraph-like approach is impractical. Therefore, a method that employs non-contact sensing to detect these physiological indicators at stand-off distances would be very beneficial.

Under conditions of stress or physical activity, the sympathetic division of the autonomic nervous system prepares the body for a rapid defense reaction by modulating hemodynamic patterns. Major responses include increased heart rate and contractile force, dilation of blood vessels in skeletal and cardiac muscles, and constriction of blood vessels that supply internal organs. The redistribution of blood flow in superficial vessels causes changes in skin temperature that can be detected by sensitive thermal cameras. We focus our thermal imaging efforts on the face and neck because this region is generally unobscured by clothing, facial blood vessels lie close to the skin surface, and information about heart and respiratory rates can be extracted from regions in the face and neck. The periodic fluctuations of temperature in regions of the skin above superficial arteries may be detected as a measure of the cardiovascular pulse. Likewise, respiratory patterns may be detected by observing the temperature fluctuations at the tip of the nose during inhalation and exhalation. Not only do these measures provide meaningful information for the detection of stress, but they make thermal imaging immensely useful in medical applications of non-invasive health monitoring.

A comprehensive body of work has been performed in our laboratory to evaluate thermal imaging techniques in the detection of stress. This thesis is organized around three human subjects studies designed to investigate the performance and feasibility of thermal imaging for the specific cases described below. In each experiment, ground truth data are provided by contact physiological sensors and our sensor suite records visible and infrared imagery and returned radar signals for analysis (described in detail in subsequent chapters).

1) Stress can generally be categorized as either mental or physical stress, each of which induces physiological activity that can be detected by thermal imaging. For security and surveillance purposes, we are primarily interested in detecting mental

stress. Therefore, it is important that an individual who is physically stressed (possibly due to moving at a fast pace or carrying a heavy object) is not falsely classified as being mentally stressed or that the physical stress does not hide the indicators of mental stress. For this investigation, subjects perform both mental and physical stress-inducing tasks. Data from the mental task are evaluated to determine if thermal imaging techniques can effectively detect high levels of mental stress, and the mental versus physical task data are compared to identify thermal features that distinguish mental from physical stress responses.

2) A second investigation evaluates the capability of thermal imaging to detect two commonly measured vital signs: respiratory rate (RR) and heart rate (HR). As a benchmark, the performance is compared to that of a millimeter-wave (mmW) radar system which has been validated for non-contact vital sign detection.

3) The final set of experiments is designed to investigate the stability of thermal imaging for stress detection under various confounding factors and real-world limitations, including subject-to-imager distance, pose (angle of view), normal intra-day and inter-day variability, and the impacts of ambient conditions, facial muscle activation and topical skin products on human thermal signatures.

The remaining chapters of this thesis are organized as follows. Chapter 2 summarizes the foundational concepts of the physiology of human thermal regulation as related to the responses of stress, and also the principles of thermal imaging that make it possible to detect these responses. A summary of the literature is provided that describes several methods of detecting stress, beginning with the use of contact physiologic sensors and transitioning to non-contact sensing via thermal imaging. In addition, previous works are reviewed that demonstrate the success of radar-based and thermal imaging techniques for vital signs detection. A description of the sensor suite is provided in Chapter 3, along with the general methods of data collection and pre-processing techniques. Chapters 4-6 describe the three separate studies above, with each chapter organized to provide methods, results, and discussion sections for the given investigation. Concluding remarks and suggestions for future work comprise Chapter 7.

2. Background

2.1 The Physiology of Human Thermal Regulation

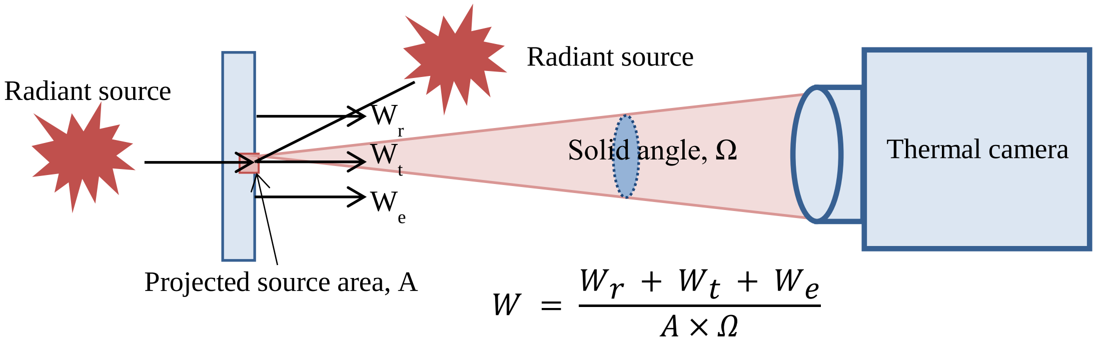
As a part of the autonomic nervous system, the hypothalamus has a major role in regulating core body temperature. Specialized sensory neurons that terminate in the hypothalamus directly control temperature by balancing heat generation and heat loss. When the hypothalamic temperature decreases, heat can be generated by shivering, which is the contraction of skeletal muscles that results in increased blood flow. In the case of an increase in hypothalamic temperature, increased sweat production allows the skin to cool and heat to be lost by evaporation. This regulation of a constant core body temperature is necessary to preserve homeostasis and allow the cells of the body to function normally. [1]

Under conditions of stress or physical activity, the sympathetic division of the autonomic nervous system prepares the body for a rapid defense reaction. Stimuli, such as emotional excitement, injury, stress, or exercise, can cause the hypothalamus to stimulate the adrenal medulla for an increase of epinephrine and norepinephrine secretion. These hormones are transported through the cardiovascular system and arrive at target tissues to enable the “fight-or-flight” response. Major responses include increased heart rate and contractile force, dilation of blood vessels in skeletal and cardiac muscles, and constriction of blood vessels in internal organs. The aim of these responses is to energize the muscles, brain, and heart for physical activity, while conserving energy by slowing the functions of internal organs and the gastro-intestinal system. [1]

As the metabolic activity of skeletal muscles increases, the core body temperature rises above the constant homeostatic range. The hypothalamus receives input from thermoreceptors to promote methods of heat loss. Dilation of the blood vessels in the skin allows heat to be transferred from the body core to its surface, which is then given off to the environment via three modes of heat transfer. Heat is transferred by conduction from the blood to the skin, and then by convection as air passes over or sweat evaporates from the skin. Along with convection, heat can be transferred from the skin to the environment by radiation.

2.2 Thermal Imaging Principles

Of the three modes of heat transfer, radiation is the process of most relevance to thermal imaging. The quantity of radiative heat flow is called radiance, which is measured as the amount of radiant heat flux at a point on the surface of the receptor divided by the solid angle and the projected area ($W \cdot sr^{-1} \cdot m^{-2}$) [2]. Radiance can be present in three forms: emitted from the object’s surface, reflected off of the object’s surface, or transmitted through the object’s surface; the total radiance is, therefore, the sum of the emitted, reflected, and transmitted components (Figure 2.1). The reflected and transmitted radiances originate from external sources of radiation, whereas the emitted radiance comes directly from the surface and is related to the surface temperature [3].



$$W = \frac{W_r + W_t + W_e}{A \times \Omega}$$

Figure 2.2. The total radiance (*W*) detected by the thermal camera is composed of the sum of the reflected, transmitted, and emitted components (W_r, W_t, and W_e), divided by the solid angle (*Ω*) and projected area (*A*).

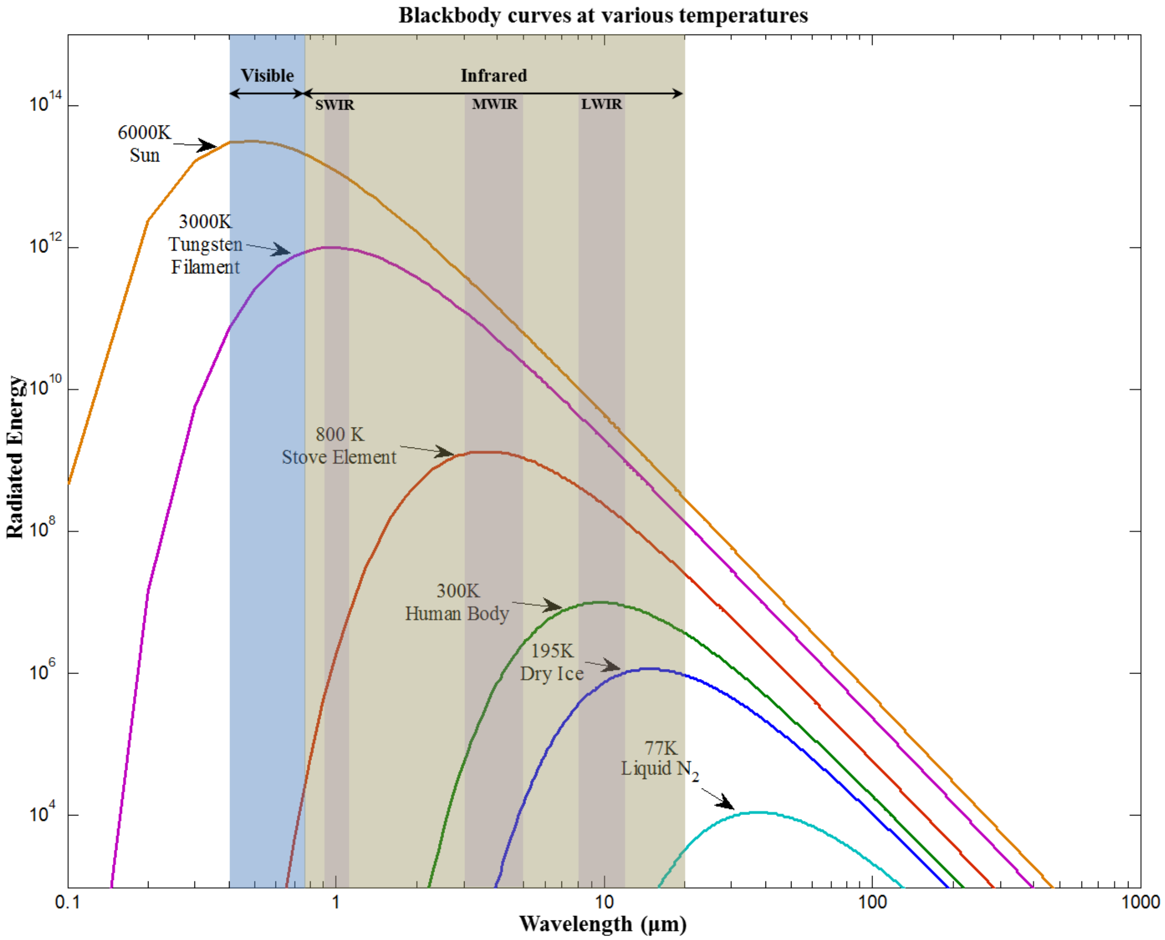


Figure 2.1. Spectral distribution of radiated energy, calculated from Planck's law, at several temperatures. Typical measurement ranges for short-wave infrared (SWIR), mid-wave infrared (MWIR), and long-wave infrared (LWIR) regions are given. As shown, a substantial portion of energy emitted from the human body exists in the MWIR region of the electromagnetic spectrum. Modified from [3].

The concept of a blackbody is important to understand for quantifying the radiance of real objects. A theoretical blackbody is a perfect absorber of all wavelengths of incident radiation, therefore having no energy reflected or transmitted. This makes a blackbody the ideal surface to measure by thermal imaging since the detector will only receive emitted energy and no other radiation that can confound the temperature measurement. Planck first determined the spectral intensity of a blackbody, $I_{\lambda,b}$ (W/m^3), which is given here as a function of wavelength, λ (m), and temperature, T (K):

$$I_{\lambda,b} = \frac{2hc^2}{\lambda^5[\exp\left(\frac{hc}{\lambda kT}\right)-1]}, \tag{1}$$

where the b subscript refers to *blackbody* intensity, h is Planck’s constant (6.626 x10^{-34} J·s), k is the Boltzmann constant (1.381 x10^{-23} J/K), and c is the speed of light in a vacuum (2.998 x10^8 m/s). This equation, known as Planck’s law, can be used to evaluate the spectral distribution of radiation emitted by a blackbody at a given temperature as shown in Figure 2.2. [4]

All objects radiate energy in the infrared (IR) portion of the electromagnetic spectrum. More energy is emitted by hotter objects, which can also radiate in the visible portion of the spectrum and be seen by our eyes. The visible spectrum extends from wavelengths of 0.4 μm to 0.75 μm, and the measurable IR spectrum begins at 0.75 μm and ends at about 20 μm. As objects cool, the amount of energy decreases and the wavelength increases to longer wavelengths of the IR region. This relationship is described by Wien’s displacement law which determines the wavelength of maximum radiation (μm) for a blackbody:

$$\lambda_m = \frac{b}{T}, \tag{2}$$

where b is Wien's displacement constant (2897 μm-K), and T is the absolute temperature of the object (K). [3]

The total radiant energy emitted from a blackbody over all wavelengths can be determined by the Stephan-Boltzmann law, which states:

$$E_b = \sigma T^4 \, , \tag{3}$$

where E_b is the total emissive power per unit area (W/m^2), and σ is the Stephan-Boltzmann constant (5.670 x10^{-8} W/m^2/K^4). [4]

The theory of a blackbody being a perfect emitter allows it to be used as a reference for measuring the emission of a real surface. Emissivity ε, which is a function of wavelength, is defined as the ratio of the radiant energy emitted by a surface to that of a blackbody. A blackbody, therefore, has an emissivity equal to one and represents the maximum radiation that can be emitted by any surface at the same temperature. Human skin is one of few surfaces to have an emissivity close to that of a blackbody; the emissivity of skin is about 0.98 and is fairly constant across wavelengths in the IR region (0.975 ± 0.05 for 3 to 15 μm). [2, 4]

Thermal imaging cameras use a focal plane array of IR detectors, on to which the lens focuses IR energy to be converted into an electrical signal. Detectors may be cooled or uncooled, and suited for specific spectral region measurements. Whereas uncooled cameras find use in consumer markets, scientific-grade devices incorporate cooling to mitigate the impact of random thermal noise on measured values. A thermal camera is usually classified by the spectral region it measures, e.g., short-wave, or near-, infrared (SWIR) (0.9-1.1 μm), mid-wave infrared (MWIR) (3-5 μm) and long-wave infrared (LWIR) (8-12 μm). Indium antimonide (InSb) is a narrow-gap semiconductor material used as an IR detector in cooled mid-wave infrared (MWIR) cameras; it is intrinsically sensitivity to wavelengths between 1-5 μm. To further select the spectral sensitivity bandwidth, an IR filter can be placed in front of the detector. [3]

Elements of the focal plane array collect IR photons over a period of time to compute a grey level by integrating the total counts. Pixel gray levels are stored as digital values that, with proper camera calibration, correlate to skin surface radiance; knowing the object's emissivity allows the digital value to be converted to temperature. The camera bit-depth, or dynamic range, typically ranges from 8-bit to 14-bit, which corresponds to 256 to 16,384 temperature level outputs, respectively. The sensitivity of a thermal imaging camera is described by its noise equivalent temperature difference (NETD). This parameter describes the necessary signal temperature to match the spatial and temporal noise in an image, obtaining a signal-to-noise ratio (SNR) equal to one. A camera with low NETD, typically given in units of millikelvin (mK), has good sensitivity and is able to distinguish between objects of very little temperature difference.

2.3 Stress Detection in Humans

With an understanding of the physiological processes controlled by the autonomic nervous system, stress recognition has been actively investigated with various sensing platforms. Psychophysiological measurements have been widely used to detect changes in stress levels [5]. Physiologic signals, such as electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), skin temperature (ST), and pupil dilation (PD), along with respiratory inductance plethysmography (RIP) and blood volume pulse (BVP) sensors, have been used in various trials as indicated in Table 2.1, to generate features that are most indicative of stress level changes [6-8]. Techniques of pattern recognition are often implemented to evaluate these features and label each data sample as belonging to one of multiple classes, e.g., stressed versus non-stressed (a two-class problem) or low versus medium versus high stress levels (a multi-class problem).

Healey and Picard [6] presented a method for analyzing physiological data to determine stress levels while driving. Subjects were measured during three real-world driving conditions: rest, highway driving, and city driving. Using video analysis and feedback from a questionnaire, 112 five-minute data samples were assigned a ranking of low, medium, or high stress level. Then 22 features from the measured physiological signals, including ECG, EMG, GSR, and RIP, were input into a linear discriminant analysis to classify the stress level during each of the sample periods. An overall accuracy of 97.4% classification accuracy in recognizing the three stress levels was reported.

Table 2.1 Summary of stress detection experiments using physiological contact sensor: electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR), respiratory inductance plethysmography (RIP), blood volume pulse (BVP), skin temperature (ST), and pupil dilation (PD).

Author	**Experimental Task(s)**	**Sensors**	**# Subjects**	**# Features**	**# Samples**	**Classification Performance**
Healey [6]	Driving	ECG, EMG, GSR, RIP	9	22	112	Accuracy: 97.4%
Barreto [7]	Stroop test	BVP, GSR, ST, PD	32	11	192	Accuracy: 90.1%
Shi [8]	Public speaking, arithmetic, cold water	ECG, GSR, RIP, ST	22	26	3,858	Precision: 68%

Barreto [7] used a computer-based Stroop test to elicit stress in subjects at pre-defined intervals while measuring BVP, GSR, ST, and PD. During the test, stressful and non-stressful periods were administered to generate 192 samples from 32 subjects. From the measured signals, 11 features were extracted and used for classification (two-class problem) in three learning algorithms. The Support Vector Machine (SVM) model was reported to have the highest accuracy of 90.10%.

Shi [8] built a personalized stress detection model to evaluate physiological measurements for the detection of stress during multiple tasks that represent social, mental, or physical challenges. Subjects were measured during four tasks and six rest periods, with self-report assessments collected after each to label the periods as stressed or non-stressed. ECG, GSR, RIP, and ST were used to extract 26 features in 60-second overlapping windows, forming 3,858 data samples. The model incorporated person-specific information with a SVM to classify each of the samples, achieving a reported precision (positive predictive value) as high as 68% at 80% recall.

These experimental studies [6-8] have found success in using contact physiological sensors to detect changes in stress levels in humans who are being mentally challenged. The collected signals represent the expected sympathetic responses to stress by the autonomic nervous system. With an increase of interest in the development of technologies to monitor or assess the physiological state of humans at stand-off distances, a different, *non-contact*, approach is necessary. By virtue of its ability to measure blood flow patterns during stress or physical activity, thermal imaging is a candidate modality for non-contact human state assessment.

Pavlidis [9] described a concept for detecting suspects involved in potentially harmful activities, primarily for defense and security scenarios. His aim was to remotely detect facial patterns of anxiety, alertness, or fearfulness that often accompany those preparing to or partaking in illegal activities. The main objective was to determine a unique correspondence between facial thermal patterns and particular psychosomatic activities, while explaining the observations with expected physiological responses. Thermal patterns occurring during other activities were observed to check if they could clutter the response to the anxiety event. Six subjects performed a series of tasks that included a startle stimulus, gum chewing, and leisure walking while being recorded with a LWIR camera. Selected image frames from the beginning, middle, and end of each activity were segmented into five facial areas. In response to the startle stimulus, Pavlidis observed increased temperatures in the perioptic area and over the carotid artery, and decreased temperatures in the cheeks. This response is expected to be associated with increased blood flow around the eyes, indicating the body's readiness for rapid eye movements. During the gum chewing activity, localized chin warming was observed; an obvious response to muscle activity. In the leisure walking activity, gradual cooling of the nose was observed while all other facial areas remained stable. Pavlidis describes this effect as a response to a more active breathing pattern. The results of this experiment demonstrate that a mild psychosomatic activity causes an appreciable thermal effect on the face that may be detected via thermal imaging.

Puri [10] presented an application of thermal imaging for measuring the emotional states of users completing a computerized version of the Stroop Color Word Test. Subjects were imaged during the computer tasks with an aim of extracting physiological variables from the facial thermal video that could indicate user stress or frustration. A region of interest (ROI) within the forehead area was selected and tracked through the experiment. Within the ROI, the mean of the ten percent hottest pixels, representing the temperature signal of the forehead frontal vessels, was computed for every frame. A bioheat model was then used to compute blood flow and volume of the frontal vessels. During a separate testing session, a ground truth measurement of energy expenditure was also obtained by analyzing respiratory activity; this measure has been documented as a reliable stress indicator [11]. A thermal stress indicator was computed from the difference in blood volume during baseline and Stroop Test for comparison to the ground truth. The Pearson correlation factor for 12 subjects, excluding one outlier, was 0.91, providing experimental evidence that thermal imaging is a viable method for measuring psychological stress.

Yuen [12] investigated the use of thermal imaging to classify emotional and physical stress by identifying the facial areas having increased temperatures and distinct patterns characteristic of stressor type. Subjects were measured with a LWIR camera during a series of tasks that include both emotional and physical stressors. The emotional stressor involved a quiz and mental arithmetic, during which all subjects were seen to have an increase in heart rate. The proportion of pixels in the facial region that exceed a randomly chosen threshold of 37.8 °C was calculated, and Yuen reported that, during the task, the number of hot pixels increased by an average of about 50%, with 90% of the increase found in the forehead, lip, and ear drum areas. The physical stressor involved running up stairs for two minutes then returning to be measured. The results of this stressor showed a predominant increase of hot pixels in the perioptic region, and less of an increase in the forehead in comparison to the emotional stressor. These findings were mentioned to be in contradiction to Pavlidis's work, which showed a substantial increase of temperature in the perioptic region during emotional stress and not during physical exercise. Yuen also claims to have found patches of elevated temperature regions in the forehead that are characteristic patterns to either physical or emotional stressors, although no specific description of these patterns are provided and the images of only one subject are used to display the distinction. The major contribution of this work is the finding of different facial regions that have an increased temperature as a response to mental or physical stress.

2.4 Non-Contact Vital Signs Detection

The work presented in the previous section has validated several techniques for detection of stress, either by contact physiological sensors [6-8] or non-contact thermal imaging [9, 10, 12]. In addition to the detection of stress, there is interest in the ability to detect and measure physiological vital signs, such as respiration and heart rate, to allow standoff-distanced acquisition of signals that are known to be directly associated with responses of the autonomic nervous system.

2.4.1 Radar for vital signs detection

Radar systems have many applications in measuring the position and displacement of an object by the nature of returned echo signals [13]. The measurement of vital signs via radar devices makes use of the Doppler shift, in that small physiological motions are detected by the associated phase shift in the signals reflected from the human torso. With the chest wall as the target, the returned signal indicates the displacement of the chest due to respiration and the mechanical movement of the heart during each beat. Millimeter wave radar signals, with frequencies in the range of 30 to 300 GHz, are capable of providing high sensitivity to such small displacements. Higher frequencies also provide the ability to maintain a collimated beam over long distances, thus reducing clutter and signal complexity.

A 228 GHz heterodyne radar system, described by Petkie [13], has been shown to be capable of measuring respiration at distances up to 50 meters and simultaneously measuring respiration and heartbeats at a distance of ten meters. The data presented were from a six-inch diameter collimated beam illuminating the subject's chest in the region of the heart apex. An advantage was found in using direction Doppler through the digitization of the in-phase (I) and quadrature (Q) signals from the intermediate frequency of the heterodyne receiver [14]. The measure of displacement is obtained by taking the arctangent of the quotient of the I and Q channels followed by phase unwrapping. The displacement signal is shown to represent respiration patterns in agreement to a ground truth signal obtained with a respiration belt that uses a piezoelectric pressure sensor to monitor chest motion. To obtain heartbeat signatures, the displacement due to respiration is considered clutter. Datasets collected while subjects held their breath clearly displayed the motion of the chest due to heartbeats and paired well with simultaneously-recorded ECG signals. When subjects breathed normally, the heartbeat signatures were more difficult to observe in the displacement signal. By taking the second derivative of displacement to obtain acceleration, the sharp features of the heartbeat signal were enhanced. Fourier transforms of the displacement and velocity waveforms show good agreement with the transforms of the respiration belt and ECG signals, validating the radar modality as a means of obtaining respiration and heart rates.

2.4.2 Thermal imaging for vital signs detection

Respiration can be measured via thermal imaging by detecting the temperature fluctuations around the nostril regions during inspiration and expiration phases. Inspired air is at or around room temperature, whereas expired air has a higher temperature since it is warmed by the respiratory passageways. A temperature waveform, representing the human breathing pattern, can be obtained by measuring the region of interest in thermal video.

Fei and Pavlidis have performed several experiments [15, 16] to measure human breathing via thermal imaging. To increase the contrast between expired air and the background, they apply an optical bandpass filter tuned to the CO_2 absorption zone (center frequency of 4.3 μm) between the lens and focal plane array of a MWIR camera. In an early experiment [15], subjects were positioned with one side of their face normal to the camera field-of-view. The tip of the nose was chosen as a tracking region of interest and used to reference a selected measurement region of interest, just in front of the mouth, where the expired air was expected to flow. The mean temperature in this region was computed in every frame to obtain the quasi-periodic temperature signal associated with respiration. Using a windowed fast Fourier transform (FFT) on the temperature signal and simultaneously collected respiratory belt signal, the dominant frequency was selected as the respiration rate from each modality. The Pearson correlation coefficient was computed to indicate the agreement between the imaged and ground truth respiration rates. The highest correlation, R = 0.9906, was found when using the largest window size for frequency analysis.

Fei and Pavlidis later improved their technique by implementing automatic tracking and localization of the nostril area, and eliminating the optical bandpass filter [16]. An initialization of the tracking region was made by interactively drawing a rectangle that encompasses the nostrils on a single frame. The nostrils are able to be located within the region due to the contrast created by the colder boundaries. Wavelet, rather than Fourier, analysis was used to determine the mean temperature and ground truth thermistor nasal signals and solve for the respiration rate. The two modalities provided nearly identical measurements, given by the complement of the absolute normalized difference (CAND), calculated as

$$CAND = 1 - \frac{BR_T - BR_G}{BR_G}\,, \tag{4}$$

where BR_T is the imaged breathing rate and BR_G is the ground truth breathing rate. The mean *CAND* for the experiment was computed as 98.27%. This technique provides a more feasible method of measuring respiration rate as it is less affected by head motion and the subject's face is normal to the camera field-of-view.

Heart rate can be measured via thermal imaging by detecting the cardiovascular pulse through skin temperature modulation [17]. Blood pulsations in superficial arteries cause skin temperature to increase through heat exchange between the vessel, overlying tissue, and skin. Thermal imaging has been utilized to obtain a measurement of the skin temperature profile in a region above major superficial arteries, often the external carotid or superficial temporal artery, which represents the cardiovascular pulse waveform.

Garbey [17] reports on a method to measure heart rate through Fourier analysis of skin temperature signals obtained by a MWIR camera at distances of three to ten feet. A line segment is interactively drawn along the center of a large superficial artery in the first thermal image. This line is expanded to a rectangular ROI with a width of 3-7 pixels which are averaged on subsequent frames, producing a temporal temperature signal for the ROI. A windowed FFT is applied to each of the signals and an average power spectrum is computed. An adaptive estimation function is used to compute the pulse frequency. This function involves convolving the current power spectrum and a weighted average of past power spectra, and assigning the dominant frequency as the heart rate. The ground truth measurement for the experiment is a piezoelectric pulse transducer worn on the fingertip during image acquisition. The performance of this method is reported as the complement of the absolute normalized difference to the ground truth, with values of 88.5% and 90.3% depending on the clarity of the vessel's thermal imprint.

Chekmenev [18] describes an alternative method to measuring heart rate through thermal imaging that overcomes several limitations of the Garbey's technique. These limitations include manual selection of the superficial blood vessel, a two-to-three minute delay in estimating heart rate, and the lack of a measured final arterial pulse waveform. In Chekmenev's method, automatic detection of the region of measurement is performed by using continuous wavelet analysis and periodicity detection. However, to achieve this automatic detection, the image field-of-view must be a zoomed-in region of the skin where the arterial pulse is expected to be observed. The region of measurement is tracked to obtain an arterial pulse waveform from the mean temperature measurement over time. For five subjects, 100% accuracy was obtained on the carotid artery area in comparison to a chest-strap heart monitor worn by the subject, and a high confidence level was obtained for two superficial temporal artery areas. The authors fail to specify how accuracy was defined.

2.5 Novel Aspects of the Current Research

Our investigations, described in the following sections, build upon previous work, yet we did not limit ourselves to sensor modalities or features described in the literature as useful in this domain. Instead, we implemented a systematic workflow in the extraction and fusion of features from our sensors to identify those patterns most indicative of the human states of interest. A foundational strength in our approach is the development of a novel segmentation pattern supported by advanced face tracking. Rather than user-selected ROIs that must often be manually identified, this segmentation allows for the analysis of distinctive facial regions that are delineated using anatomical landmarks. Several feature types are explored with pattern recognition techniques to aid in distinguishing the indicators of stress. Multiple classifiers are trained and tested to evaluate the effectiveness of these features for a given set of classification problems. These include two-class problems, such as stressed versus non-stressed, or mental versus physical stress, along with the multi-class problem of low versus medium versus high stress levels.

The detection of vital signs through non-contact sensing is demonstrated under more operationally feasible techniques than those of previous works. By utilizing the same face tracking-aided segmentation approach, no user input for ROI selection or initialization is required. Also, we demonstrate the capabilities of our techniques within more realistic settings by allowing subjects to move freely during the mental and physical task experiments.

3. Materials and Methods

3.1 Imaging System

The imaging system used for the experimental study incorporates a MWIR camera (ThermoVision SC6700, FLIR Systems Inc., Wilsonville, OR) and a high-resolution progressive scan electro-optical camera (A202k, Basler AG, Ahrensburg, Germany). Images from the two cameras are acquired synchronously through two Camera Link frame grabbers housed in a chassis equipped with a real-time system integration (RTSI) bus (NI PXI-1428 and NI PXI-1033, National Instruments Inc., Austin, TX). An MXI-Express cable connects the chassis to an acquisition PC with a NI PCIe-8360 host card installed.

The ThermoVision SC6700 camera features a cooled Indium Antimonide (InSb) 640 x 512 focal plane array (FPA) with 15μm pixel pitch that is filtered to provide a 3-5μm spectral response over the MWIR band [19]. A noise-equivalent temperature difference (NETD) of less than 25mK is quoted by the manufacturer. A fast pixel clock of 50 megapixels per second enables the camera to output 14-bit digital data at 126 frames per second at the full window size, and higher frame rates when the window is adjusted to a smaller size. The camera has fully adjustable integration time, frame rate, window size, triggering, and data transfer modes. Detailed technical specifications are provided in Table 3.1.

The Basler A202k camera uses a Kodak KAI-1020 interline transfer progressive scan charge-coupled device (CCD) sensor, providing a monochrome spectral response for wavelengths in the visible (VIS) light region [20]. The camera features high spatial resolution with a 1004 x 1004 array of 7.4 x 7.4 µm pixels. A pixel clock speed of 40 MHz enables a maximum frame rate of 48 frames per second at full resolution. Video output allows for either 8- or 10-bit data transferred via Camera Link.

Table 3.1. Technical specifications of the ThermoVision SC6700 and A202k cameras [19, 20].

Technical Specification	FLIR ThermoVision SC6700	Basler A202k
Detector Type	Cooled Indium Antimonide (InSb)	Interline Transfer Progressive Scan CCD
Spectral Range (µm)	3.0 – 5.0	0.4 – 0.7
Resolution	640 x 512	1004 x 1004
Pixel Size (µm)	15	7.4
Maximum Frame Rate (Hz)	125	48
Dynamic Range (bits)	14	8 or 10
Pixel Clock (MHz)	50	40
Video Output	GigE, Camera Link, Composite (BNC), Video (NTSC or PAL), S-video, Super VGA	Camera Link

3.1.1 MWIR camera calibration and non-uniformity correction

To ensure accurate and consistent thermal image data, the MWIR camera must be calibrated in order to convert the digital camera measurements to temperature values. FLIR Systems' RCal^{TM} utility of the RTools^{TM} radiometric software toolkit calculates polynomial equations for conversion of camera grayscale values to radiance values and, subsequently, radiance values to temperature values. The calibration is performed using a blackbody source (IR-2100/301, Infrared Systems Development Inc., Winter Park, FL), by taking measurements of a ROI within the source at several temperature settings across a desired temperature range. For these experiments, blackbody measurements were taken in one degree increments from 20 to 38 °C. Along with the blackbody measurements, data such as source emissivity, source area, camera to source distance, and background temperature are used by RCal^{TM} to generate the calibration equations. Both emitted and reflected radiance components from the blackbody source are assumed to be detected by the camera, making it necessary to estimate the reflected component and subtract it from the total radiance to obtain the emitted component only. The radiance is calculated by integrating the Planck's law equation over the spectral bandwidth.

To account for the differences in the pixel-to-pixel responses for each element of the detector array, a non-uniformity correction (NUC) is performed to obtain a uniform pixel response. A NUC table, consisting of a map of bad pixels, along with each pixel's necessary gain and offset values, is applied to each live camera image to mathematically correct for response variations. There are three different approaches for creating the NUC table: one-point, two-point, and offset update [19]. The two-point correction process is the most effective and performs better than the one-point correction for a wide range of temperatures. This correction requires two uniform sources, set to the upper and lower ends of the desired temperature range to be measured; for these experiments, these temperatures were set to 22 and 38 °C. One at a time, the sources are positioned so that they fill the camera's entire field-of-view; images of these two sources are then analyzed to compute the unique gain and offset for each pixel. Bad pixels are subsequently detected by checking if the gain coefficient exceeds the user defined limits, or if the pixel exhibits large changes over several frames. Pixels that are indicated as bad are replaced using either the nearest neighbor or two-point gradient algorithms. An offset update is applied prior to each subject session. This correction retains the current gain values of the NUC table and updates the offsets using the camera’s internal temperature flag as the NUC source.

Image uniformity is impacted by heating and cooling of the camera electronics during normal use, thus requiring modification of the offset value (the gain setting is typically constant). When using a two-point NUC table, an efficient offset update correction process can be performed periodically to maintain the existing bad pixel map and gain but compute new offset coefficients in the table. This correction is made using the internal temperature flag as the uniform source. [19]

3.1.2 MWIR and VIS image registration

Post-acquisition co-registration of the MWIR and VIS images geometrically aligns the images to allow for area-based analysis in corresponding image regions. Geometric correspondences are found by control-point matching in the images of a test pattern. Our 20 x 30 inch test pattern features a black-and-white checkerboard pattern with IR LEDs located at several intersections, providing distinct features to be selected as fiducial points in both images (Figure 3.1). In this study, affine geometry was assumed and transformation was achieved through bilinear interpolation. Using our in-house registration algorithm [21] that allows for both coarse and fine registration and assesses the resultant registration accuracy via a mutual information measure [22], a transformation coefficient matrix is produced that best scales, translates, and rotates the MWIR image to match the viewpoint of the VIS image.

Our custom MATLAB (The MathWorks Inc., Natick, MA) routine prompts the user to interactively select control points in the first pair of corresponding images (Figure

3.2), after which the MWIR image is transformed and displayed for user acceptance. If unsatisfactory, the user can re-select different control points and repeat the process. Once the user is satisfied with the quality of the registration, all remaining frame pairs in the collection sequence are transformed and the registered MWIR images are saved in a separate directory. All subsequent image processing uses the registered MWIR images.

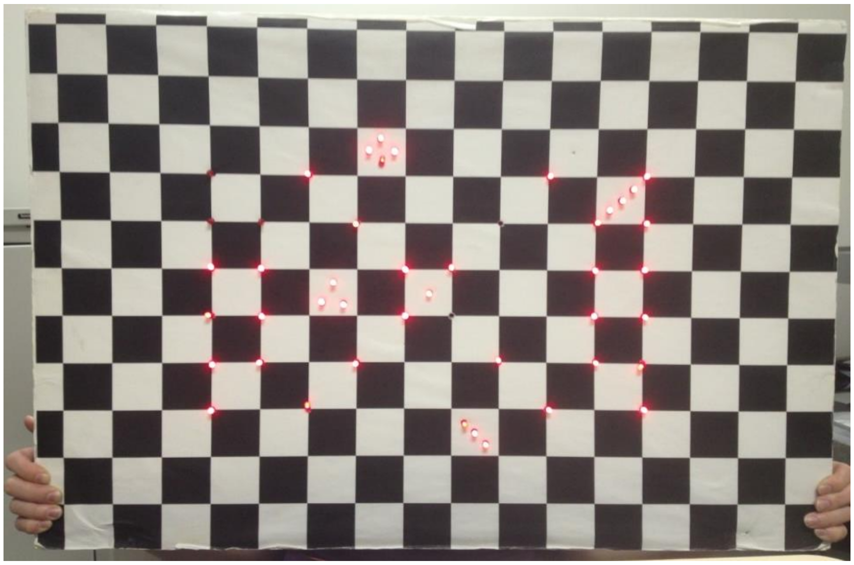


Figure 3.1. The test pattern for image registration is held by the subject in a frontal plane orientation at the position of his/her face. The board features a black-and-white checkerboard pattern with IR LEDs that provide distinct features to be selected in the VIS and MWIR images.

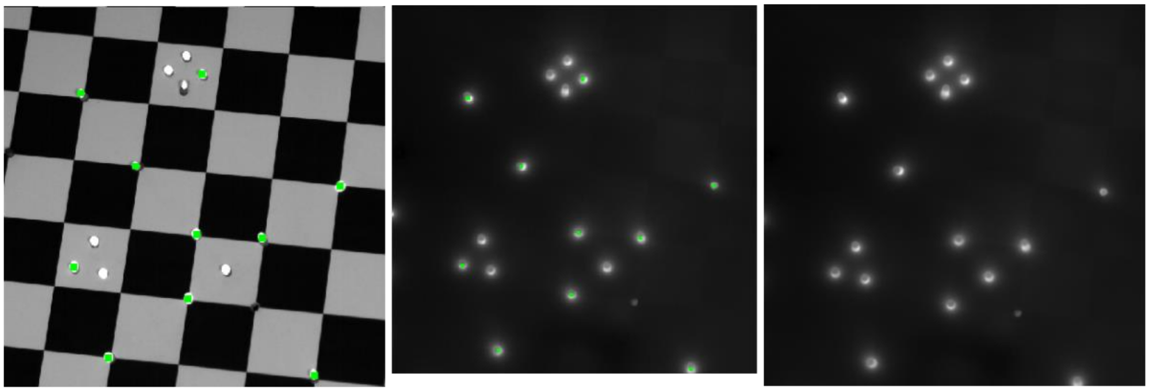


Figure 3.2. The green markers identify selected control points for registration of VIS (left) and MWIR (middle) images. The transformed MWIR image (right) is geometrically aligned with the VIS image and has the same dimensions. Pixels outside the common field-of-view are set to black.

3.1.3 Facial feature tracking

To perform facial segmentation, an automated facial feature tracker (visage|SDK^{TM} FaceTrack, Visage Technologies AB, Linkoping, Sweden) is implemented to return coordinates of anatomical landmarks and estimates of point positions where no landmarks are present. The tracker returns 84 fiducial points and head rotation parameters (yaw, pitch, and roll) for every frame of real-time video sequences. Tracking is performed using the VIS images because of the higher contrast in facial features as compared to the MWIR images. To initialize the tracking, the operator positions a subject-specific mask over the face in a *neutral* frame (i.e., the subject is generally forward-facing and has a neutral expression) by scaling, rotating, and dragging the vertices of a predefined mesh to match the subject's facial features. Once completed, the tracker iterates through each of the remaining frames to output coordinates of the points based on the fitted mask. Using 49 of the 84 points returned by the visage|SDK face tracker, and an additional 25 custom-derived points, we divide the face and neck into 29 non-overlapping segments (Figure 3.3).

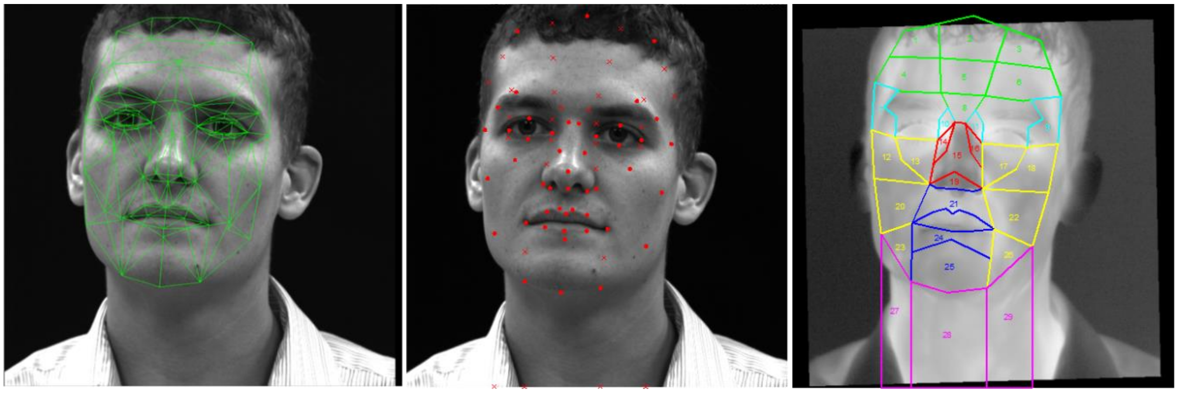


Figure 3.3. Subject-specific adjustments are made manually to a mesh overlaid on a neutral frame (left) prior to tracking. From the 84 control points returned by the Visage facial feature tracker, 49 points (circles) and an additional 25 derived points (x's) (middle) define the vertices of 29 non-overlapping facial segments (right).

Each frame is automatically analyzed so that an acceptable segmentation mask is developed from the output coordinates. At the six-foot subject-to-imager distance, the mask must meet the criteria of having a width of at least 30% of the image size, and a height of at least 50% of the image size, measured by the distances between the outermost points, to ensure that an appropriately sized mask is overlaid on the face. This provides an alternative to using the yaw, pitch, and roll parameters as rejection criteria for the degree of head rotation since the size of the mask decreases if the head is substantially rotated in any direction away from the frontal view. A frame will also be rejected if any of the output coordinates are out of the image bounds, which indicates that a portion of the subject's face is out of frame.

3.2 Radar System

A 35 GHz continuous wave, mmW radar system (SRR-35121010, Ducommun Inc., Carson, CA), illuminating a 10-inch diameter region on the heart apex, is used to detect the motion of the chest wall for non-contact vital signs sensing (Table 3.2). The system is constructed with an I/Q mixer, having a 90 degree phase difference, to obtain directional information of the Doppler radar signal. The two channels relate to the target's direction in that when the phase shift is positive the target is approaching; negative phase shifts indicate that the target is receding [23]. The signals of the I and Q channels are digitized by a DAQ board (National Instruments Inc., Austin, TX) at 600 Hz. The phase is calculated by taking the arctangent of the quotient of I and Q to provide a measure of chest wall motion.

Table 3.2. Technical specifications of the Ducommun 35 GHz radar system [24].

Technical Specification	Typical Value
RF Frequency	35.5 GHz
Transmitter Output Power	+10 dBm
Receiver Conversion Loss	10 dB
IF Bandwidth	DC to 100 MHz (minimum)
I/Q Channel Phase	90° ± 10°
Antenna 3 dB Beamwidth	12°
Antenna Side Lobe Level	-20 dB (maximum)
Polarization	Right hand circular
DC Bias	+5.5 V / 350 mA
Operation Temperature	-40 to +85 °C

3.3 Contact Physiological Sensors

Physiological signals are recorded during the experiment with a wireless physiological monitoring system (BioRadio 150, Cleveland Medical Devices Inc., Cleveland, OH). A 12-channel wireless ambulatory device is worn by the subject to acquire, amplify, sample and digitize up to 12 signals from sensors attached to the body. The signal data are wirelessly transmitted to the USB Receiver that is coupled to an acquisition PC. The BioRadio 150 is capable of recording, displaying, and analyzing all signals in real time.

For this project, we measured ECG, GSR, chest respiratory effort, blood oxygen saturation (SpO2) and heart rate. The ECG signal is recorded with snap electrodes in a three-lead configuration on the subject’s wrists and forearm (Figure 3.4). GSR is recorded with two snap electrodes measuring the skin conductance level between the palm and middle of the forearm on the subject’s non-dominant side (Figure 3.4).

Respiratory effort is measured with a piezoelectric respiratory belt worn around the subject’s chest that records the voltage generated by the sensor during chest expansion and contraction. SpO2 and heart rate are recorded by a pulse oximeter sensor that wraps around the subject’s thumb of the non-dominant hand (Figure 3.4). The ECG, GSR and respiration signals are sampled at 600 Hz; and the signals from the pulse oximeter sensor are sampled at 60 Hz.

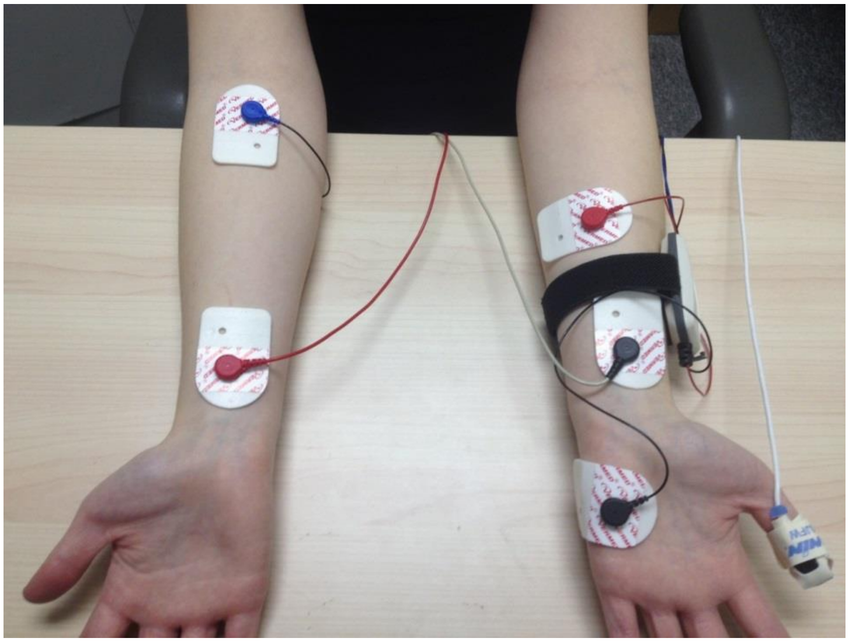


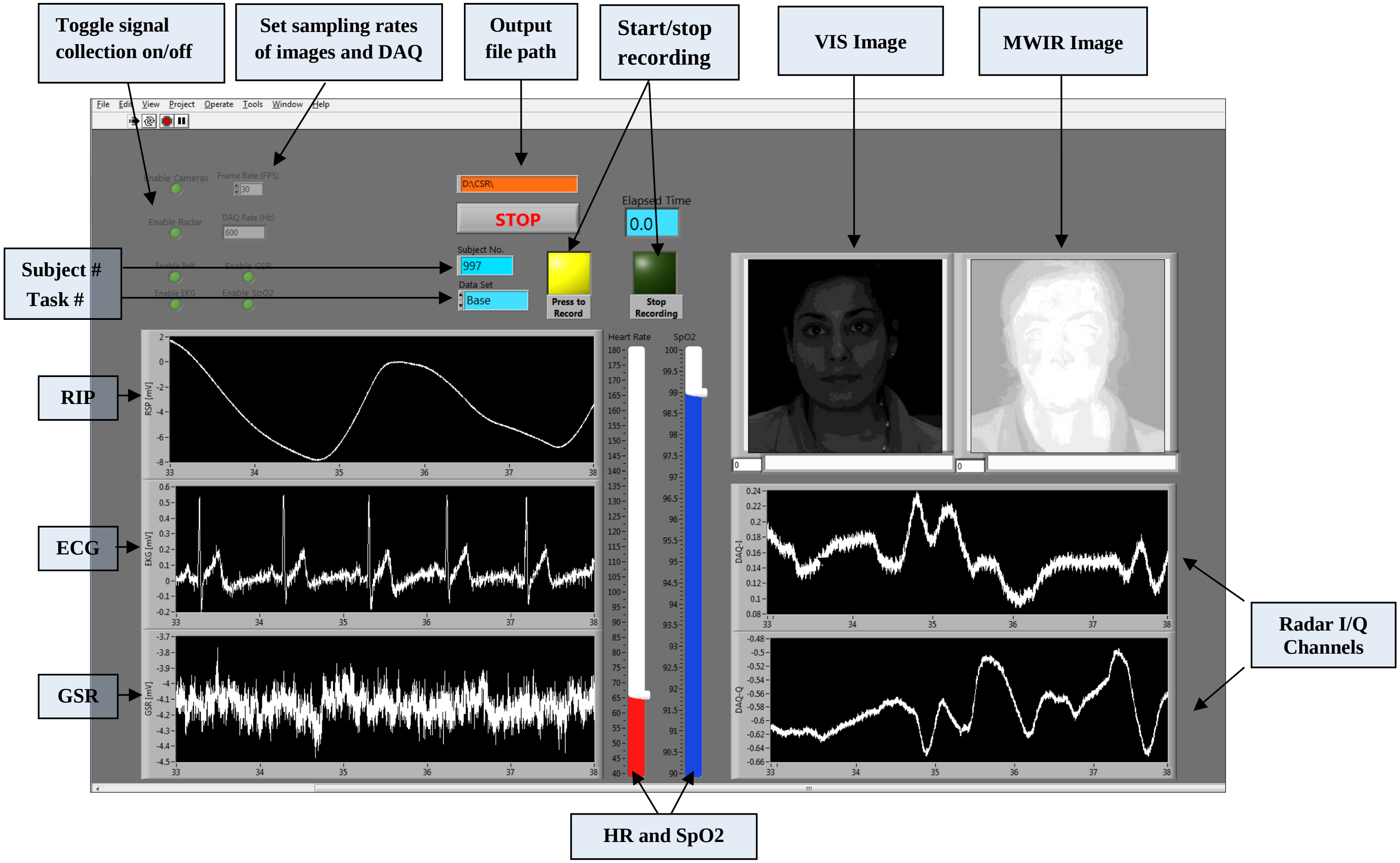
Figure 3.4. ECG electrodes attached to subject’s left and right wrist, and right forearm, GSR electrodes attached to subject’s left palm and forearm, and pulse oximeter sensor attached to subject’s left thumb.

3.4 Data Collection

The specifications described in this section apply to all of our human subject trials, aside from certain trials that do not include physiological monitoring. The general protocol is described here; details relevant to each experiment are provided in the respective sections.

A custom graphical user interface (GUI) was developed within LabVIEW (National Instruments Inc., Austin, TX) to collect and allow for real-time viewing of VIS, MWIR, radar, and BioRadio data. The interface allows the user to specify which signals to collect, the image sampling rate and the appropriate file output path (Figure 3.5). Time-synched VIS images (windowed to 600 x 600 pixels) and MWIR images (windowed to 512 x 512 pixels) are collected at 30 frames per second (fps). An integration time of 4 ms was chosen for the thermal camera to maximize the usable dynamic range without saturation. Although image collection is synchronized to within one clock cycle, or 33.3 ms, the slower transmission speed of the BioRadio data results in a timing offset between the physiologic data and the image data of 0.80 ± 0.05 s. Room temperature and humidity are recorded at the time of each experimental session using a data logger (OM-71, OMEGA Engineering Inc., Stamford, CT).

The human subjects studies followed a protocol approved by the Wright State University Institutional Review Board. Upon arrival to the experimental session, subjects complete a demographics survey and answer questions about physical activity immediately prior to the session. Subjects are seated in an adjustable-height chair with a distance of six feet between the MWIR FPA and the subject’s face. The VIS camera is positioned in front of the MWIR camera, and the millimeter wave radar is set on a height-adjustable stand in front of the imaging system, approximately four feet from the subject (Figure 3.6). Once the subject is in position and the chair height and cameras have been aligned so that the subject’s face is centered and fully within each camera’s field of view, images of the registration test pattern are acquired. The subject holds the test pattern normal to the camera and in the frontal plane at the position of their face for several



Frame Rate (FPS) 30

DAQ Rate (Hz) 600

D:\CSR\

Elapsed Time 0.0

Subject No. 997

Data Set Base

Figure 3.5. The LabVIEW graphical user interface provides a real-time view of the VIS, MWIR, radar, and BioRadio data as they are being collected. The operator is given options for which signals to collect, the sampling rates, and the file output path.

seconds as the camera operator collects images. The remaining data collections involve imaging the subject under baseline conditions for three minutes, and then while performing one or more tasks, which are described in the following chapters.

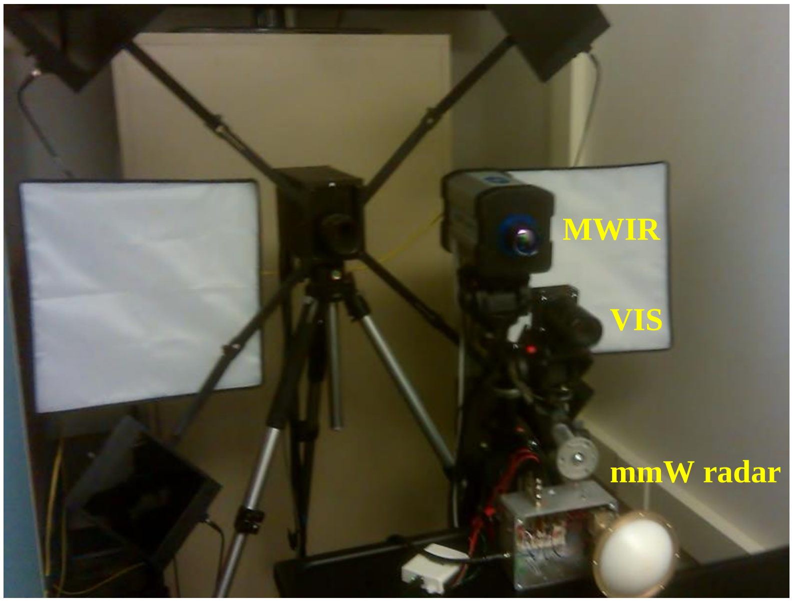


Figure 3.6. Subject’s view of the imaging system. The MWIR camera is positioned six feet from the subject and aligned with the VIS camera such that the face and neck fill the field of view of each imager. The mmW radar system sits on a stand four feet from, and aligned with the subject’s chest.

4. Mental and Physical Stress Detection

4.1 Experimental Procedure

Image and sensor data were collected from 32 study participants (16 males, 16 females) who were Wright State University students between the ages of 18 and 44 years, with a mean age of 22.2 years. 24 Caucasian, 3 Asian, 2 African-American, 1 Latino and 2 other students participated. Each subject underwent three sequential data collections: baseline, mental task and physical task. Prior to the data collections the subject answers a questionnaire to rank their current stress level, on a Likert scale (1 to 10), based on activities and experiences throughout the day. During the baseline measurement, the subject is seated and facing the imagers (i.e., frontal pose), remaining still as data are recorded for three minutes.

4.1.1 Mental task

Immediately following baseline measurement, the subject completes the Stroop Color Word Test mental task, which is facilitated by a wireless keyboard and a 55-inch television located above the imaging system approximately ten feet from the participant. Following task instructions that are displayed on the TV screen, there is an opportunity to ask the experimenter questions about the task. Next, 12 example items are presented without a response time limit. After completing the examples, the subject is again allowed to ask any questions before performing the task.

This version of the Stroop Color Word Test was programmed using the Inquisit (Millisecond Software, Seattle, WA) experimentation software package for designing and administering psychological experiments. The subject is presented with a word which is the name of a color, and then must respond quickly with the color of text in which the word is displayed. This task has been reported to create mental overstimulation due to cognitive conflict, as the word that is written may not agree with the color of the text, and there are time-pressure effects [25]. For example, the word “red” may be written in blue text, so that the correct response is “blue.” The words and colors consist of combinations of red, green, blue and black. Subjects enter their response on easy-to-find home-keys of the keyboard; typing “d” for red, “f” for green, “j” for blue, and “k” for black. These corresponding keys and colors are displayed on the screen throughout the test. If an incorrect response is entered, or if the subject does not respond within the allotted time, a buzzer sound is played. The test lasts for approximately three minutes with the allotted response time decreasing each minute from 1.5, to 1.0, to 0.5 seconds.

4.1.2 Physical task

After completing the mental task, the subject moves to a table to complete the stress analysis questionnaire. Similar to the pre-task questionnaire, the subject ranks his or her stress level, on a Likert scale (1 to 10) to indicate how he or she felt during the task. At this time, the experimenter replaces the subject’s chair with a recumbent exercise bicycle (Fusion 4545, Stamina Products Inc., Springfield, MO) that allows for physical exertion with minimal head motion. The seat is positioned such that the height and leg extension distances are appropriate and comfortable for the subject. The experimenter verbally instructs the subject to pedal the exercise bicycle at a moderate speed for five minutes. Data collection begins once the subject begins pedaling. The experimenter monitors the subject’s heart rate during the exercise period and instructs the subject to increase or decrease the pedaling speed as necessary to remain within the target heart rate range for aerobic exercise. This range is defined as 60% to 80% of the subject's maximum heart rate, which is calculated using the simple and traditional (albeit widely debated [26]) estimate of 220 beats per minute minus the subject's age. After five minutes of exercise, the subject is told to stop pedaling but to continue facing the cameras for an additional three minutes as data collection continues during this recovery period.

4.2 Data Processing

Data processing (Figure 4.1) includes registration, facial feature tracking and segmentation (all described in Chapter 3), followed by extraction and selection of thermal features under several different paradigms for considering mental and physical stress detection.

4.2.1 Feature extraction

Following human review of segmentation on registered MWIR images, MATLAB routines are used for segment-based feature extraction. Based on the findings of one of our parallel research projects which found that statistical features outperformed textural features for human state classification, we short-listed our feature set to include the mean pixel value (MEAN), maximum pixel value (MAX), and the mean of the top 10% hottest pixels (TTM). These three features are computed for each segment on every image frame by creating polygonal masks defined by the segments' vertices, overlying those masks on the image to extract a ROI and computing each feature of interest for the spatial ROI. At our collection frame rate of 30 fps for 14 minutes (baseline, mental, physical tasks), 29 segments per frame and three features per segment yields about 2.2

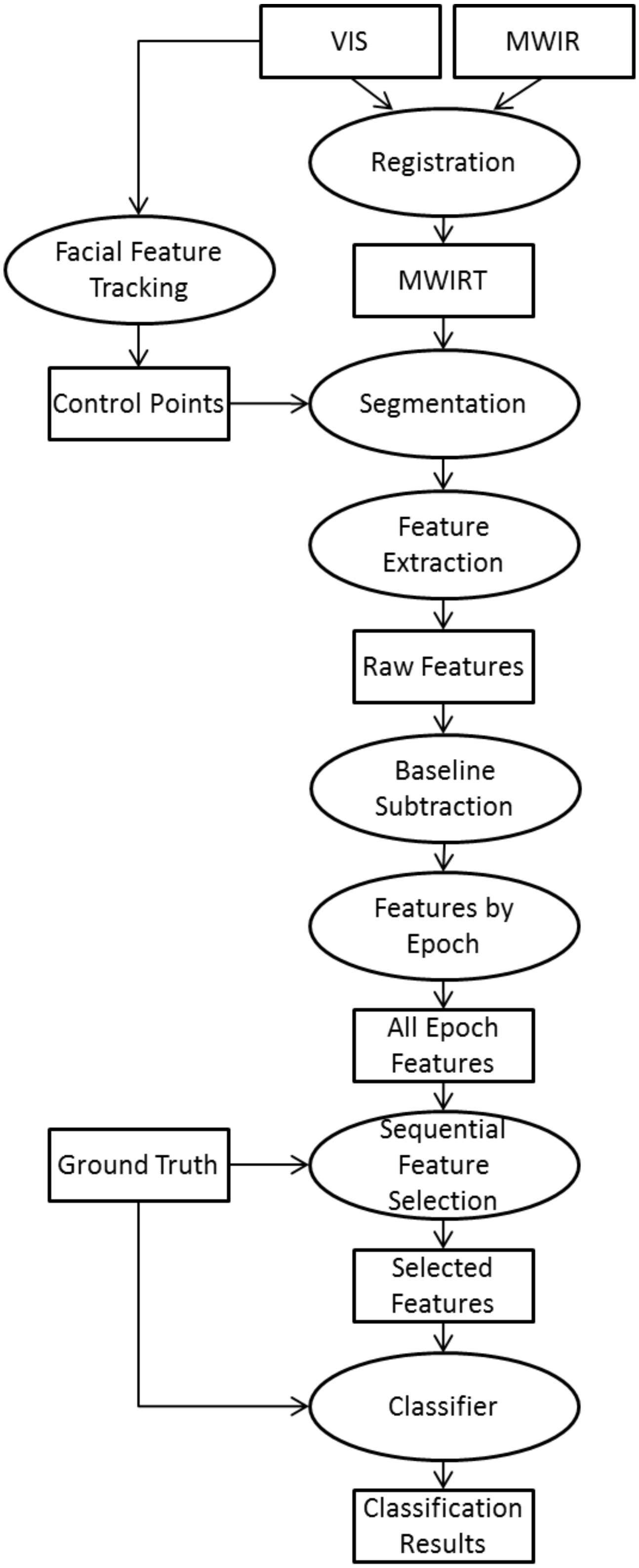


Figure 4.1. Data processing procedure for the classification of data from the mental and physical stress trials. Rectangles represent data; ellipses represent processes. MWIRT refers to the transformed (registered) MWIR images.

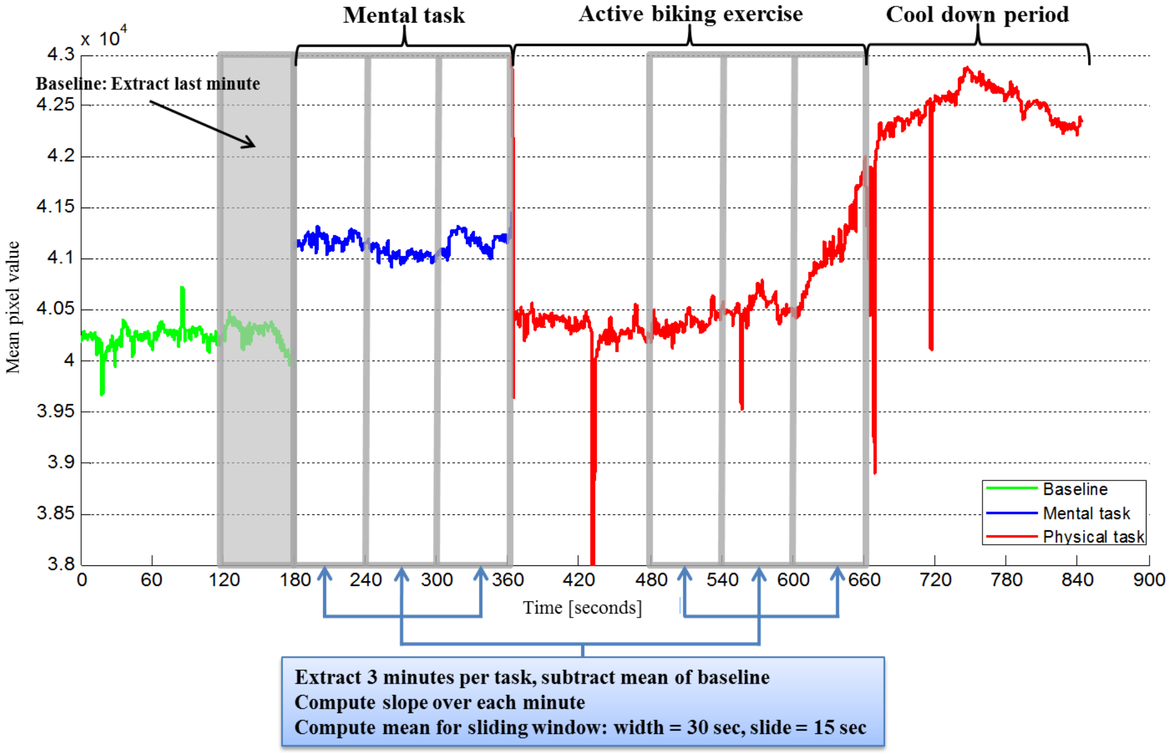
million features per subject. These *raw* features must obviously be reduced to a more manageable set size for input into the classification algorithms.

4.2.2 Features by epoch

Data reduction is accomplished in several stages. First, the average over the last minute of baseline is subtracted from all of the raw task features such that the feature data now represent deviation from baseline. Next, the number of features is reduced by defining time epochs of interest (Figure 4.2).

A three-minute time span is then selected from which the means of a sliding window (30-second width, 15-second slide) are computed, along with the slopes of each individual minute. Therefore, each time epoch yields 11 means and three slopes for each of the three statistical features for each of the 29 segments, resulting in 1,218 features per epoch. Selection of the three-minute analysis window varies by stress classification task. For the mental task, the time span covers the entire task collection so that these features may be used in classifying levels of mental stress. In the classification of mental versus physical stress, the mental task is compared to a selected three-minute time span of the physical task: either the first three minutes of exercise (minutes 1-2-3), the last three minutes of exercise (minutes 3-4-5), or the recovery period after exercise (minutes 6-7-8).

Studies have shown that a human’s stress response is generally greatest at stressor onset and then gradually subsides as habituation to the stressor occurs [27]. Therefore, we also investigated another time-segmentation approach that uses data from only the first task minutes for the classification of mental versus physical stress. Using the means over 10-second, non-overlapping windows, and slopes over 30-second windows (where appropriate), feature sets ranging from 10 to 60 seconds of data per task are created. By comparing the classification accuracies of these different feature sets, the trade-off between collection time and classification accuracy is investigated.



Extract 3 minutes per task, subtract mean of baseline
Compute slope over each minute
Compute mean for sliding window: width = 30 sec, slide = 15 sec

Figure 4.2. The data reduction approach to reduce raw features to features by epoch. This method is used to compare features from the three-minute mental task to one of the three-minute time spans of the physical task.

4.2.3 Ground truth estimation from physiological measurements

In the essence of machine learning or statistical modeling, ground truth refers to an objective data set by which the accuracy of the method can be analyzed. When dealing with the human state, the ability to establish objective ground truth becomes difficult due to the dynamic and unpredictable nature of human responses. For the application of stress detection, it is desired to use ground truth data to separate the observations into two or more classes that represent stress levels. Two common ways to collect ground truth, physiological measurements and self-report, are used in this experiment.

The stress analysis questionnaires ask subjects to rank their stress level before and after completing the Stroop Color Word Test. The differences in these values are calculated for each subject, resulting in average difference of 1.41, and a standard deviation of 1.92. Based on the distribution of the differences, there is not a clear separation that could be used to divide the subjects into classes based on stress level (Figure 4.3). Several subjects even reported a lower value of stress level after the task than before leading us to question the reliability of the self-report as a measure of ground truth in this study. Additionally, because combining these measures of ground truth is difficult and because the thermal imagery is expected to more closely represent the physiologic responses to the stressors, we ultimately chose to confine our ground truth data to only the physiological measurements.

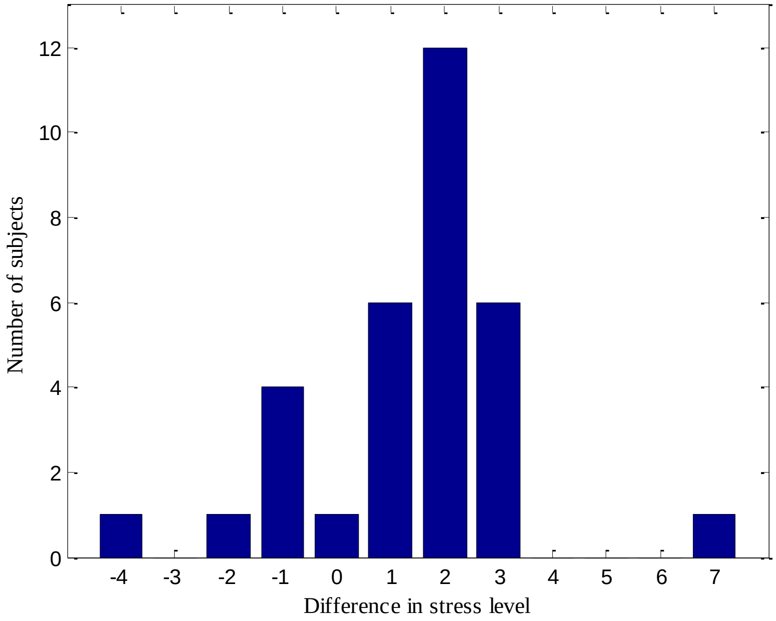


Figure 4.3. The distribution of the differences in pre- and post-mental task stress levels, as recorded via subject self-report. Most (~78%) subjects reported mild-to-moderate increases in stress as a result of the task, but given the mean difference of 1.41 with a standard deviation of 1.92, the use of these data as a measure of ground truth may lead to unreliable results.

Three physiological measures are used to estimate the level of mental stress for each subject: heart rate (HR), respiratory rate (RR), and galvanic skin response (GSR). Each of these measures is expected to increase in response to the stressor, representing the activity of the sympathetic nervous system [1]. Increased sweat in the skin increases the electrical conductivity, which is measured by the GSR sensor. HR is measured by the pulse oximetry sensor at a sampling rate of 60 Hz (Figure 4.4), RR through frequency analysis of the respiratory belt signal (Figure 4.5), and GSR (Figure 4.6) by two contact electrodes; the latter measures are digitized at 600 Hz; the plots depict the three measures for one study participant. For each physiological measure, the average value is calculated over each of the three minutes of the mental task. The averages of the baseline measures are subtracted from the averages of the task minutes such that the features represent change from baseline. With three minutes of mental task and three physiological measures, a total of nine physiological variables are used for ground truth estimation.

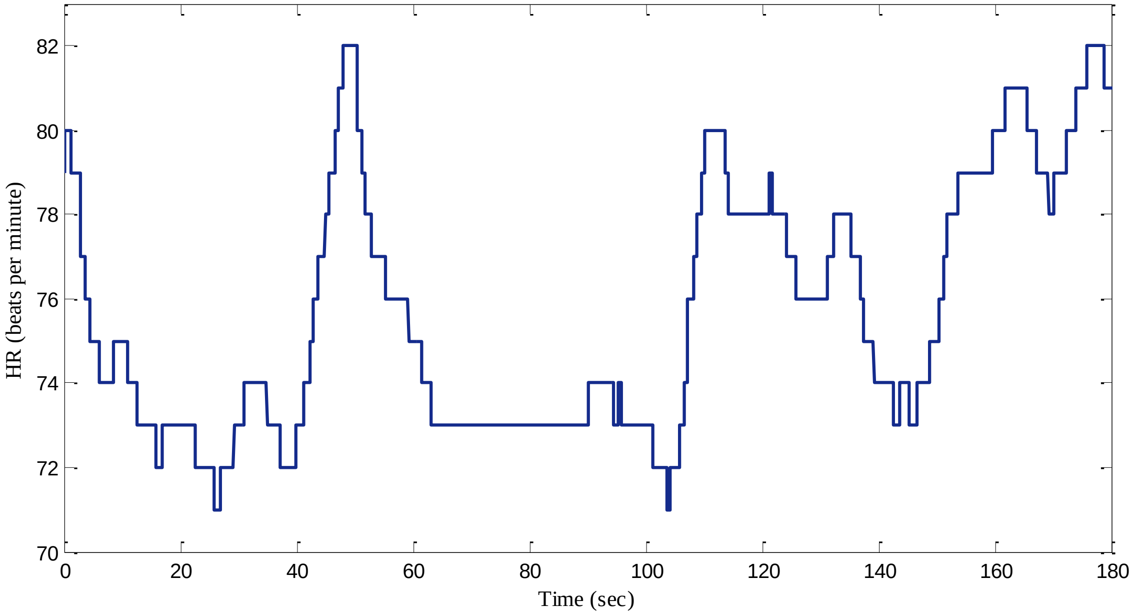


Figure 4.4. Heart rate sampled at 60 Hz with the pulse oximetry sensor over three minutes of the mental task for one subject.

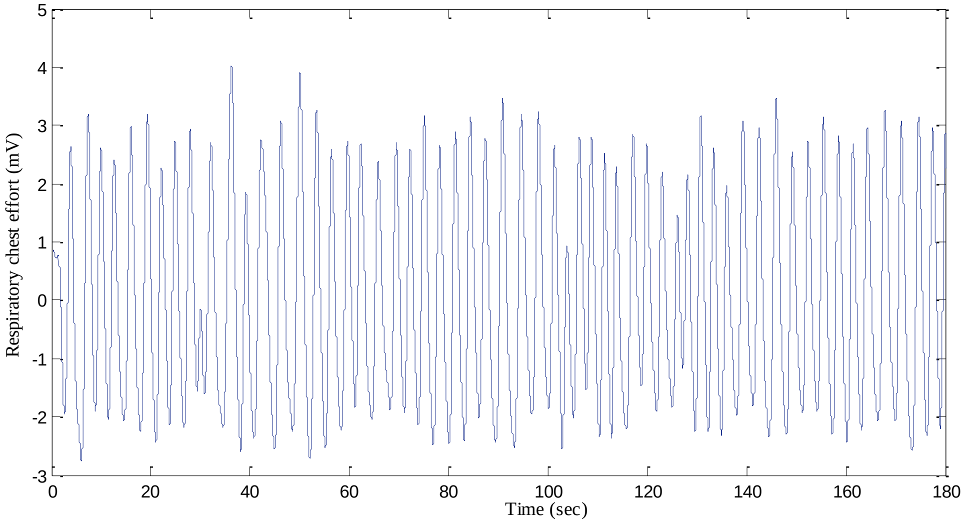


Figure 4.5. The respiratory belt signal digitized at 600 Hz over three minutes of the mental task for the same subject.

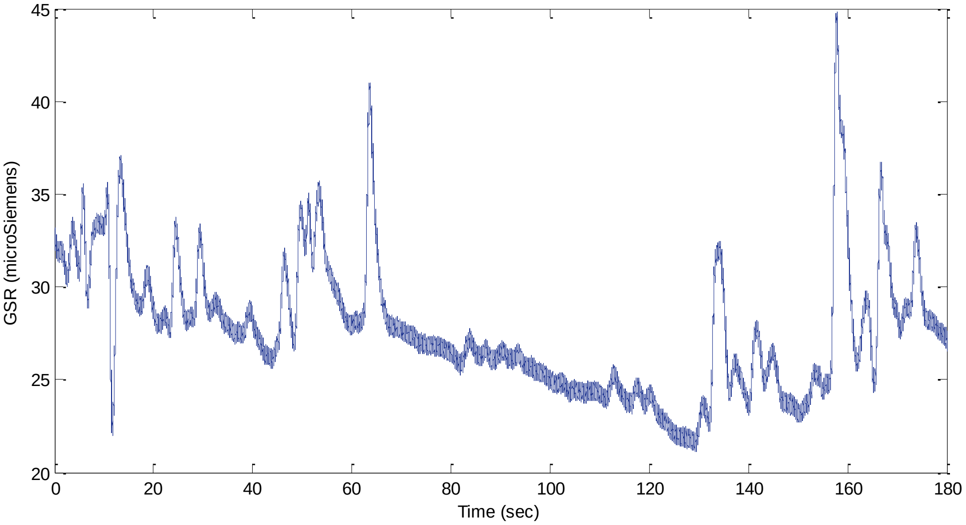


Figure 4.6. GSR digitized at 600 Hz over three minutes of the mental task for the same subject.

Frequency analysis is used to determine RR by performing a fast Fourier transform (FFT) on one minute of the recorded respiration signal. The frequency component with the highest magnitude is selected as the respiration frequency. Before performing the FFT, a bandpass filter with cutoff frequencies at 0.08 and 0.5 Hz is applied to the respiration signal; this frequency band corresponds to the expected respiratory rate range of 4.8 to 30 breaths per minute. As the respiratory signal may still be impacted by noise related to body motion, false dominant frequency peaks may be selected. Therefore, the respiration signals and frequency spectra are carefully manually analyzed to determine whether or not the appropriate frequency is returned by the algorithm.

Although positive correlation was expected between the physiological measures, the correlation between variables during the same minute ranged from R = -0.34 to 0.23. Perhaps this lack of correlation represents the unpredictable nature of human responses or results from the task that only induces mild-to-moderate stress in most subjects. Since the variables are found to be independent, they are combined to form a single ground truth measure as follows:

The nine physiological variables are converted to a z-score,

$$Z = \frac{X - \mu}{\sigma}, \tag{5}$$

where X is the raw value, and μ and σ are the population mean and standard deviation which are estimated from the sample mean and standard deviation, respectively. A composite score is obtained by summing the nine z-scores, and calculating the z-score of the sums.

Based on the composite z-scores, the subjects are divided into three classes of stress levels: high, neutral, and low. The highly stressed subjects have a composite z-score greater than +0.5, the low stressed subjects have a score less than -0.5, and the neutral subjects have scores between -0.5 and +0.5 (Figure 4.7). As a result of this stratification, the high and low stress level classes each contain eight subjects and the neutral class contains 12 subjects.

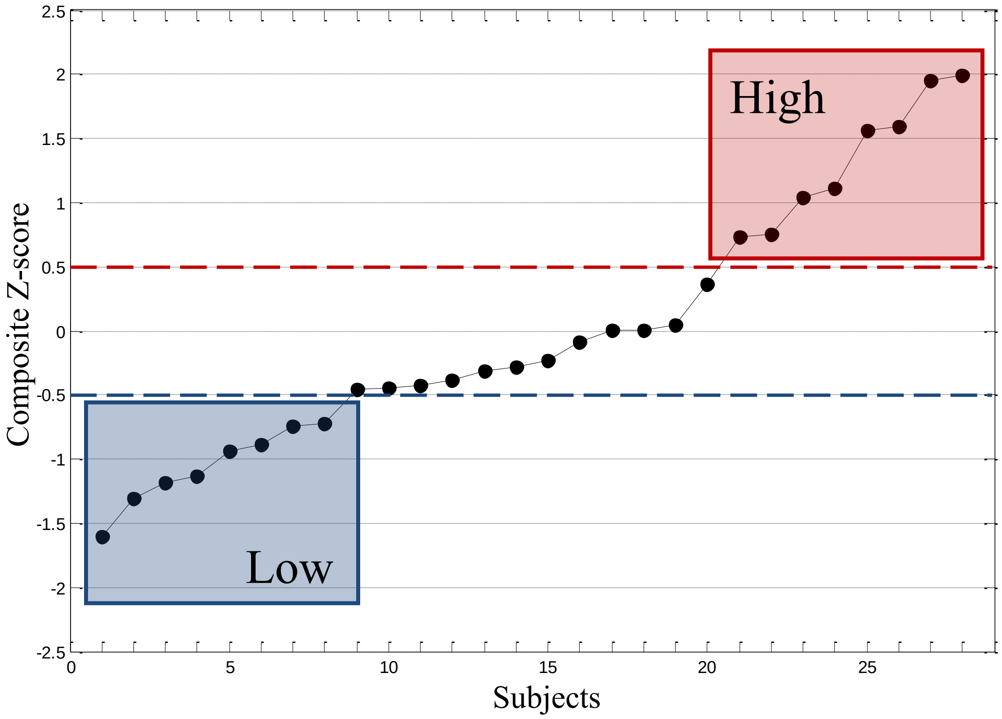


Figure 4.7. The composite z-score distribution is used to divide subjects into three stress-level classes based on ground truth physiological measures.

4.2.4 Sequential feature selection

Of the pool of 1,218 features per epoch, we aim to identify those features most relevant to the class states. The "curse of dimensionality" is described as a major problem in pattern recognition because, although more information about entities being classified is useful, this leads to high computational complexity and poor generalization of the classifier. Using feature selection, the goal is to greatly reduce the number of features and at the same time retain as much of the class discriminatory information as possible. This means that a subset of features should be selected that leads to large between-class distances and small within-class variances. [28]

A criterion measure is needed for feature selection that appropriately combines features to create the best feature vector for effective class discrimination. Scatter matrices are easily computed and provide information on feature separability. The within-class scatter matrix, as defined by Theodoridis [28], provides a measure of the variance of features in each class, and is calculated as

$$S_w = \sum_{i=1}^{M} P_i \Sigma_i, \tag{6}$$

where Σ_i is the covariance matrix for class i, and P_i is the *a priori probability* of class i, which is the number of samples in the class out of the total number of samples. The between-class scatter matrix provides a measure of the average distance of the mean of each class from the global mean value, calculated as

$$S_b = \sum_{i=1}^{M} P_i(\mu_i - \mu_0)(\mu_i - \mu_0)^T, \tag{7}$$

where μ_0 is the global mean vector. A criterion value J is computed based on these scatter matrices to minimize within-class variances and maximize between-class distances:

$$J = trace\{S_w^{-1} S_b\}. \tag{8}$$

This calculation multiplies the inverse within-class scatter matrix by the between-class scatter matrix and sums the elements of the main diagonal with the *trace* operation.

Sequential forward feature selection (SFS) is used to select a subset of features that maximize the criterion value J. Since SFS generally looks to minimize a criterion, we established that criterion to be $1/J$. The iterative procedure sequentially adds features that improve the class separability. In the first step, the criterion value is calculated for each feature individually and the feature with the smallest value is selected. Next, the previously-selected feature is combined with each remaining candidate feature to form two-dimensional feature vectors. The criterion value for each is computed and that with the smallest value is selected. This process continues, increasing the size of the feature vectors each time, until the improvement in *1/J* does not exceed an empirically defined threshold value of 0.01. This value was chosen by initially allowing SFS to run through 25 iterations so that the best feature vectors, ranging in length from one to 25, are selected. Each of these feature vectors is then input into the classifier to assess accuracy versus number of features. Generally, as features are added, classification accuracy increases to a maximum, after which a gradual decrease is observed. Improvements in accuracy tend to cease once the change in criterion due to the addition of a feature stabilizes to a value of about 0.01 or less.

4.2.5 Classification

A supervised learning classifier is designed to predict the class of a sample based on the observations of training data whose ground truth classes are made known. Four kinds of classifiers are used and compared in this study: an artificial neural network (ANN), a naïve Bayes classifier, linear discriminant analysis (LDA) and a support vector machine (SVM).

An artificial neural network with a two-layer perceptron architecture is able to solve nonlinearly separable problems. This architecture consists of an input layer, where the input feature data are applied and no processing takes place, a hidden layer consisting of a number of nodes that weight each of the input features, and an output layer that consists of the same number of processing nodes as classes. The ANN uses supervised learning to assign weights to each node in the network as it is trained to identify classes based on the input data. Training data are repeatedly tested in the network to gradually adjust the weights until a classification error criterion is minimized. Once trained, the network may be used to classify samples that were not included in the training data.

A naïve Bayes (NB) classifier is a simple method of classification that uses Bayesian probability rules to determine the likelihood that an unknown sample, represented by a feature vector *x*, belongs to a given class ω_i [28]. Bayes' Theorem states that the conditional probability, i.e. the probability that *x* belongs to class $\omega_{i,}$ is

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}, \tag{9}$$

where $p(x|\omega_i)$ and $p(x)$ are the probability density functions (PDF), and $P(\omega_i)$ is the *a priori probability* of the class. Obtaining accurate estimates of the PDF based on a number of training samples *N*, an *l*-dimensional feature space would require N^l points. When an assumption is made that individual features x_j, $j = 1, 2, \ldots, l$, are statistically independent, the PDF can be obtained as

$$p(x|\omega_i) = \prod_{j=1}^{l} p\left(x_j|\omega_i\right), \tag{10}$$

thus requiring only *lN* points for each class. Equation 5 can now be written as

$$P(\omega_i|x) = P(\omega_i)\prod_{j=1}^{l} p\left(x_j|\omega_i\right), \tag{11}$$

which is used by a naïve Bayes classifier to assign the feature vector *x* to the class with the largest probability, $P(\omega_i|x)$. As is the case here, the independence assumption made for a naïve Bayes classifier is generally not true for real data sets, but the technique has been shown to still perform well despite this invalid assumption.

Linear discriminant analysis is a classification technique that attempts to find a linear combination of features to separate classes. The criterion *c* of the feature vector *x* belonging to a class ω_i is computed as the projection of *x* onto the weight vector *w*, such that

$$c = wx. \tag{12}$$

For the two-class problem, the weight vector is calculated as

$$w = \Sigma^{-1}(\mu_1 - \mu_2), \tag{13}$$

where μ_1 and μ_2 are the mean vectors of the training data in each of the two classes. This technique is based on Bayes Decision Theory with the assumption that the covariance matrices of the two classes are equal. LDA essentially creates a decision hyperplane in the feature space, and the class to which *x* belongs is determined by the side of the hyperplane in which the value *c* lies. [28]

A support vector machine is designed to find the optimal decision hyperplane,

$$g(x) = w^T x + w_0 = 0, \tag{14}$$

having a direction *w* and position in space w_0, by maximizing the margin from the nearest training vectors of each class. These are called the support vectors and are critical in the optimization problem, which can be solved analytically using Lagrange multipliers λ under the Karush-Kuhn-Tucker conditions, such that

$$w = \sum_{i=1}^{N_s} \lambda_i y_i x_i, \tag{15}$$

where N_s is the number of support vectors. [28]

Cross-validation is a technique used during classifier testing to provide a non-biased estimator of classification accuracy. *K*-fold cross-validation involves randomly dividing the data into *k* folds, and then using all but one of the folds as training data whereas the remaining fold is used as testing data. This process is repeated so that all of the folds are tested individually and an average of the *k* accuracies is calculated. When *k* is equal to the number of samples, only one of the samples is tested and the rest are used for training. This *leave-out-out cross-validation* is used when there are a very limited number of samples.

Classifier performance may vary between repeated trials for few reasons. First, the random division of the data into folds during cross-validation yields different training sets to be used for each trial. Those training sets that best represent the entire data set result in higher classification success. Secondly, with the leave-one-out method, where the training data are identical for any given sample, an ANN classifier may still exhibit inconsistencies on repeated trials due to random weight initialization. As is common practice, we used one of the most effective weight initialization techniques, the Nguyen-Widrow randomization [29]. This technique uses a set of equations based on the number of nodes in the input and hidden layers to adjust the initial weights for decreased training time, but the weights still begin as random variables. Multiple executions of the ANN are performed when using the leave-one-out method, but only one execution is necessary for the other classification techniques.

4.3 Results

4.3.1 Sequential feature selection

SFS was performed using all features by epoch (1,218 for a three-minute time period) to select a subset that provides a high separability based on the assigned class labels. For mental stress detection, the class labels were assigned using physiological ground truth measurements. Three executions of SFS were performed to generate feature subsets specific to our three different classification problems: high versus remaining, high versus low, and high versus neutral versus low. For the detection of mental versus physical stress, feature data from the mental task were compared against feature sets from three different time epochs of the physical task: Set 1 is the first three minutes of exercise, Set 2 is the last three minutes of exercise, and Set 3 is the three minute recovery period immediately after exercise. Using data from each of these three sets, SFS selected different feature subsets to be used for classification.

To further illustrate how the SFS selection stopping criterion value threshold was selected, when the ANN accuracies for the classification of mental versus physical stress (Set 2) using feature subsets of lengths one to 25 are plotted versus the number of features n (Figure 4.8), accuracy is observed to gradually decrease for subsets having more than nine features. Plots of the criterion value $1/J$ and the change in criterion value $\Delta 1/J$ versus n show that at n = 9, $\Delta 1/J$ first falls below 0.01 (Figures 4.9 – 4.10). This criterion is used as a surrogate measure of expected classifier performance (since J is quickly calculated versus incrementally inputting feature vectors of increasing length into the classifier). In this example, the stopping criterion results in six to eight features selected for the subsets of mental stress detection, and nine to eleven features for mental versus physical stress detection (Table 4.1).

Table 4.1. Number of features selected by SFS for different mental stress detection and mental versus physical stress detection problems, where the three-minute mental task is compared to one of three selected time windows of the physical task.

Mental Stress Detection	**High v. Remaining**	**High v. Low**	**High v. Neutral v. Low**
Number of features	11	9	11
Mental v. Physical Stress Detection	**Minutes 1-2-3 of Physical Task**	**Minutes 3-4-5 of Physical Task**	**Minutes 6-7-8 of Physical Task**
Number of features	7	6	8

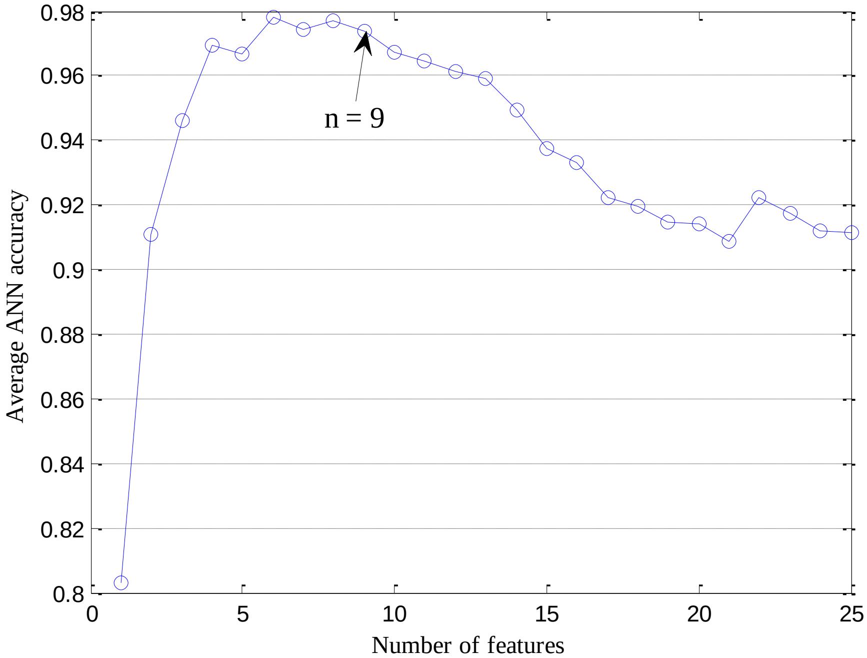


Figure 4.8. Generally, classification accuracy improves dramatically with the addition of the first several features to a maximum value. Accuracy then degrades as features are added due to the high-dimensional feature space and small number of test samples. To confirm the general pattern, the average ANN accuracy over 1,000 trials versus number of features selected by SFS is plotted here for the classification of mental versus physical stress.

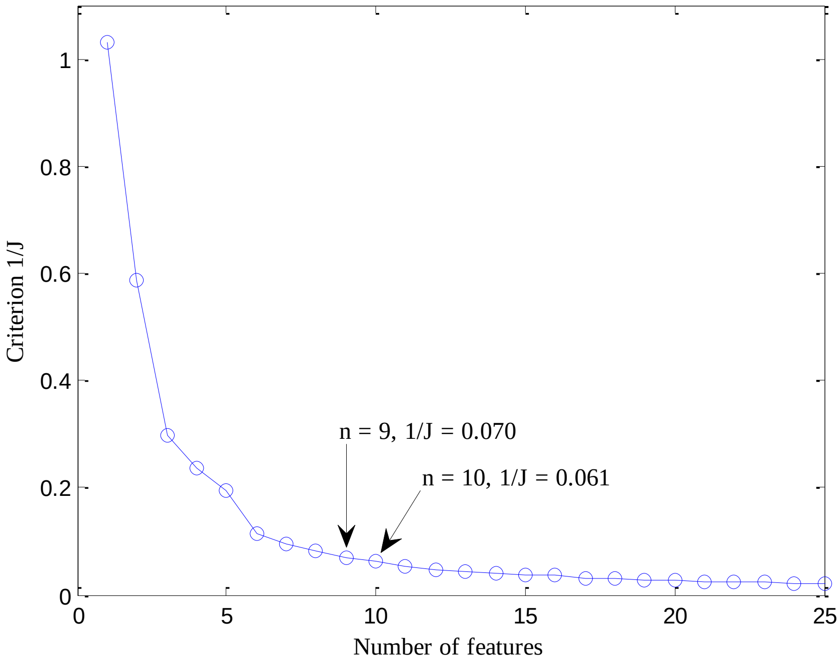


Figure 4.9. Rather than testing the classifier with feature vectors of iteratively increasing length, we define a more efficient procedure using a class separability measure criterion value, 1/*J*. For the same data shown in Figure 4, the criterion value versus number of features for each of the subsets selected by SFS is plotted. When the change in criterion value falls below 0.01 with the addition of a feature, the subset is considered optimal for classification (here, n = 9 features are selected).

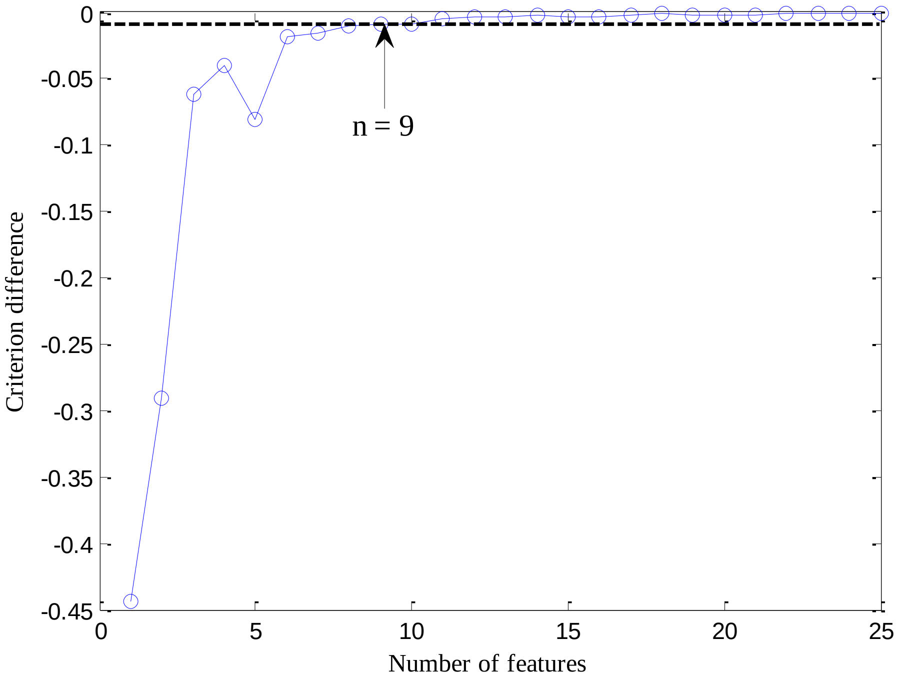


Figure 4.10. For the datasets described in Figures 4 and 5, features are added until the change in criterion value resulting from the addition of a feature through SFS falls below the stopping threshold of 0.01.

4.3.2 Mental stress level classification

Three classification trials were used to predict the level of mental stress in subjects using different combinations of the high, low, or neutral classes assigned by the ground truth physiology scores: high versus remaining, high versus low, and high versus neutral versus low. In each of the trials, a classifier was trained and tested using leave-one-out cross-validation. For the ANN, ten executions are performed, allowing each subject to be tested ten times while the remaining subjects are used as training data. The ANN is retrained in every trial with different initial weights being used. Classification accuracy is calculated as the number of correctly predicted test samples divided by the total number of test samples.

Confusion matrices provide the number of instances in which a sample was predicted to be in a particular class, versus the actual number of samples in the class. Therefore, the diagonals from the top-left to the bottom-right of the tables give the number of correct classifications and overall accuracy is in the lower right-hand cell. The far right columns provide the classification accuracies by class, and the predictive accuracy is summarized in the bottom row.

When classifying the highly stressed versus the remaining subjects, the ANN achieved an average accuracy of 96.4%, the SVM achieved 92.9% accuracy, and the NB classifier achieved 82.1% accuracy (Table 4.2). Using the LDA classifier and only the top two features, 26 out of the 28 subjects were correctly classified, for an accuracy of 92.9% (Figure 4.11). The features were the mean of the third 30-second window of the maximum pixel value in segment 28 (medial neck) and the slope of the first minute in the mean of the top 10% hottest pixels of segment 23 (lower right cheek). When using the top four features, 100% accuracy was achieved (Figure 4.12); when additional rank-ordered features were included iteratively, the classifier continued to perform perfectly. The third and fourth ranked features were the slope of the first minute in the maximum pixel value of segment 26 (lower left cheek) and the slope of the first minute in the mean of the top 10% hottest pixels of segment 3 (upper forehead).

Table 4.2. Confusion matrices from each of the four classifiers for the discrimination of subjects under high stress versus the remaining classes (neutral and low). Leave-one-out cross-validation was executed. Ten trials were performed for the ANN due to outcome variability caused by randomization in assignment of initial weights; the average of these trials is reported.

ANN		Predicted		
		High	Remaining	Accuracy
Actual	High	70	10	87.5
	Remaining	0	200	100.0
	Accuracy	100	95.2	**96.4**

LDA		Predicted		
		High	Remaining	Accuracy
Actual	High	8	0	100.0
	Remaining	0	20	100.0
	Accuracy	100.0	100.0	**100.0**

NB		Predicted		
		High	Remaining	Accuracy
Actual	High	5	3	62.5
	Remaining	2	18	90.0
	Accuracy	71.4	85.7	**82.1**

SVM		Predicted		
		High	Remaining	Accuracy
Actual	High	6	2	75.0
	Remaining	0	20	100.0
	Accuracy	100.0	90.9	**92.9**

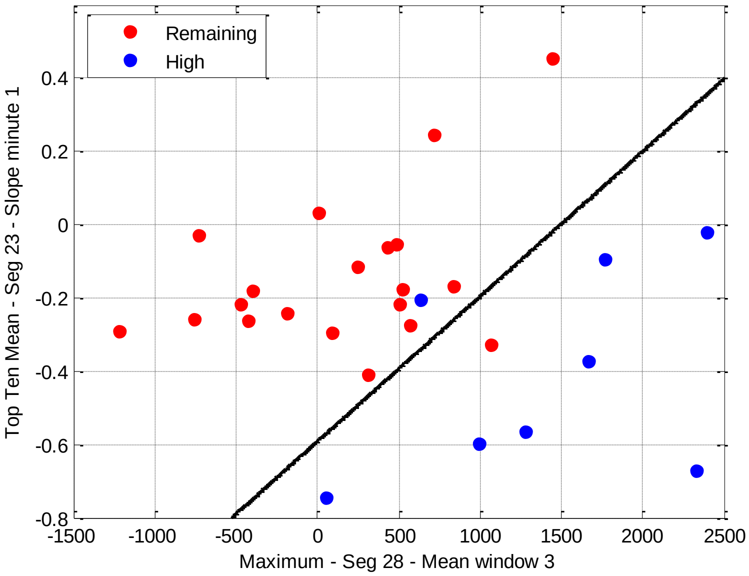


Figure 4.11. Classification of the *highly* stressed subjects (n = 8) versus *remaining* subjects (n = 20) using linear discriminant analysis with two features: the mean of the third 30-second window of the maximum pixel value of the medial neck segment, and the slope of the first minute in the mean of the top 10% hottest pixels of the lower right cheek segment.

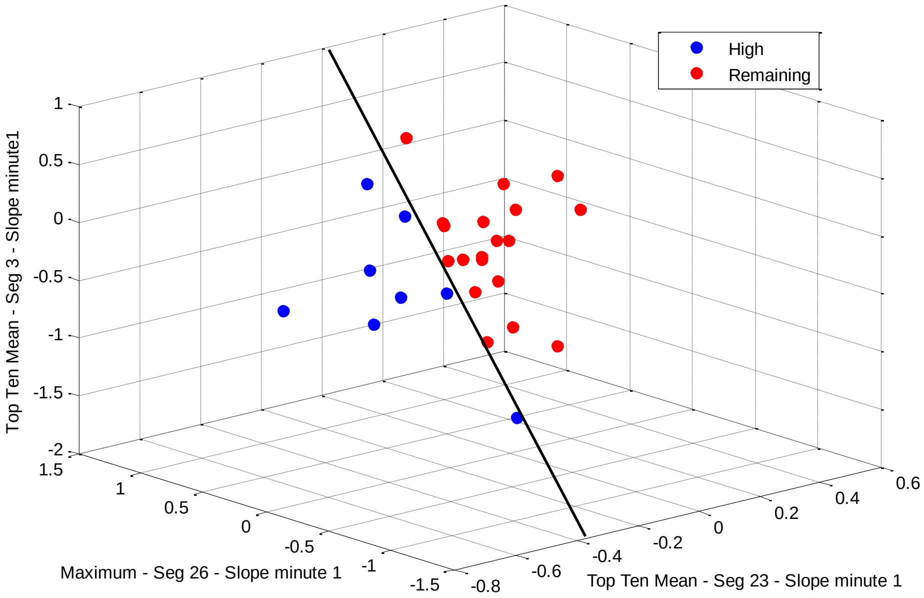


Figure 4.12. Classification of the *highly* stressed subjects (n = 8) versus *remaining* subjects (n = 20) using linear discriminant analysis with four features (only the second-, third- and fourth-best features are plotted here). 100% accuracy was achieved when adding the third and fourth features: the slope of the first minute in the maximum pixel value of the lower left cheek segment, and the slope of the first minute in the mean of the top 10% hottest pixels of the upper forehead segment.

For classifying the highly-stressed subjects versus the least-stressed subjects, the ANN achieved 88.8% average accuracy and the NB classifier achieved 87.5% accuracy (Table 4.3, upper left, lower left). The SVM and LDA classifiers correctly predicted 15 out of the 16 subjects, an accuracy of 93.4% (Table 4.3, upper right, lower right). For the three-class problem of high versus neutral versus low stress levels, the LDA classifier correctly predicted 25 out of the 28 subjects for an accuracy of 89.3% (Table 4.4). Importantly, all eight of the subjects in the high stress class were predicted correctly and no subject was misclassified into the high stress class. The ANN performed poorly for the three-class problem, achieving an average accuracy of 59.6% (Table 4.5).

Table 4.3. Confusion matrices from each of the four classifiers for the discrimination of subjects under high versus low stress levels. Leave-one-out cross-validation was executed ten times for the ANN, once for the other classifiers.

ANN		**Predicted**		
		High	**Low**	**Accuracy**
Actual	**High**	69	11	86.3
	Low	7	73	91.3
	Accuracy	90.8	86.9	**88.8**

LDA		**Predicted**		
		High	**Low**	**Accuracy**
Actual	**High**	8	0	100.0
	Low	1	7	87.5
	Accuracy	88.9	100.0	**93.4**

NB		**Predicted**		
		High	**Low**	**Accuracy**
Actual	**High**	7	1	87.5
	Low	1	7	87.5
	Accuracy	87.5	87.5	**87.5**

SVM		**Predicted**		
		High	**Low**	**Accuracy**
Actual	**High**	8	0	100.0
	Low	1	7	87.5
	Accuracy	88.9	100.0	**93.4**

Table 4.4. Confusion matrix LDA using leave-one-out cross-validation to classify the *highly* stressed subjects (n = 8) versus *neutral* subjects (n = 12) versus the *least* stressed subjects (n = 8).

LDA		Predicted			
		High	Neutral	Low	Accuracy
Actual	High	8	0	0	100.0
	Neutral	0	10	2	83.3
	Low	0	1	7	87.5
	Accuracy	100.0	90.9	77.8	**89.3**

Table 4.5. Confusion matrix for ten trials of ANN using leave-one-out cross-validation to classify the *highly* stressed subjects (n = 8) versus *neutral* subjects (n = 12) versus the *least* stressed subjects (n = 8).

ANN		Predicted			
		High	Neutral	Low	Accuracy
Actual	High	52	1	27	65.0
	Neutral	1	79	40	65.8
	Low	5	39	36	45.0
	Accuracy	89.7	66.4	35.0	**59.6**

4.3.3 Mental v. physical stress classification

The similar physiological responses to mental and physical stress present the possibility of falsely classifying as a threat an individual who is physically stressed. After demonstrating the use of thermal features to detect high mental stress levels, we now aim to determine if there are distinguishing features that can discriminate mental versus physical stress responses. For operationally practical purposes, we are interested in testing the thermal features at different time periods of the physical exercise, as well as after exercise.

Feature data from the mental task were compared against feature sets from three different time epochs of the physical task: Set 1 is the first three minutes of exercise, Set 2 is the last three minutes of exercise, and Set 3 is the three-minute recovery period immediately after exercise. There were 28 subjects with usable data for Sets 1 and 2, and 21 subjects for Set 3. Since all subjects performed both the mental and physical tasks, the number of samples per class is equal to the number of subjects. Each type of classifier (ANN, NB, LDA and SVM) was trained and tested for ten executions of four-fold cross-validation. Of the three classifiers, LDA achieved the highest accuracies (98.8%, 98.9%, and 97.9%) with the smallest variances (Table 4.6). The accuracies of the SVM and ANN were equal to those of LDA for the first and second time spans, respectively, but with higher variances, whereas NB accuracies were significantly lower. The difference inaccuracies between the first and second time spans were not found to be significant (*p* = 0.563).

The classification accuracies of the ANN are shown to increase by 18.1%, 17.6%, and 16.4% for the different time spans when using the subset of features selected with SFS compared to using all of the features by epoch (Table 4.7).

Table 4.6. Average accuracies of ten trials of four-fold cross-validation to classify mental stress versus physical stress using Sets 1, 2, and 3 of the physical task for each of the four classifiers. Standard deviation is indicated in parentheses.

Period of Physical Task	**ANN**	**NB**	**LDA**	**SVM**
Minutes 1-2-3 (Set 1)	96.1 (1.13)	91.1 (2.79)	98.8 (0.86)	98.8 (1.47)
Minutes 3-4-5 (Set 2)	98.9 (1.73)	93.8 (1.74)	98.9 (0.92)	98.6 (1.13)
Minutes 6-7-8 (Set 3)	92.8 (3.44)	82.9 (1.88)	97.9 (0.75)	97.4 (0.75)

Table 4.7. Average accuracies of ten trials of four-fold cross-validation to classify mental stress versus physical stress using Sets 1, 2, and 3 of the physical task. Standard deviation is indicated in parentheses.

	All features	**Selected features**
Minutes 1, 2, 3 of Physical Task (First Three Minutes of Exercise)		
Number of Features	1218	11
Average Accuracy	78.0 (3.37)	96.1 (1.13)
Minutes 3, 4, 5 of Physical Task (Last Three Minutes of Exercise)		
Number of Features	1218	9
Average Accuracy	81.3 (3.06)	98.9 (1.73)
Minutes 6, 7, 8 of Physical Task (Recovery Period After Exercise)		
Number of Features	1131	11
Average Accuracy	76.4 (4.28)	92.8 (3.44)

Via separate analyses, isolated facial regions were used for classification to determine if the discrimination between mental and physical stress is possible using a selected sub-region of the face. The 29 segments were divided into six regions: forehead (segments 1-6, 8), perioptic (segments 7, 9-11), cheeks (segments 12-13, 17-18, 20, 22-23, 26), nose (14-16, 19), mouth (21, 24-25), and neck (27-29). Using each region's features, SFS provided down-selected feature subsets. The classifiers were tested with each feature subset using ten trials of four-fold cross-validation. In each of the three time spans, the highest accuracies were consistently achieved using a particular segment region for three out of the four classifiers. For the first three minutes of the physical task this was the perioptic region, for the last three minutes it was the cheeks, and for the three minutes after exercise it was the mouth region (Table 4.8). High classification accuracies are achieved across all time spans using only the cheek features (> 97%) or the perioptic features (> 96%), proving that these regions alone provide very useful for discriminating mental versus physical stress.

Table 4.8. Average accuracies of ten trials of four-fold cross-validation using six general facial regions to classify mental stress versus physical stress. The maximum accuracies for each time period and classifier type are highlighted. Standard deviation is indicated in parentheses.

Region	Number of Features	ANN	LDA	NB	SVM
Minutes 1-2-3 of Physical Task (First Three Minutes of Exercise)					
Forehead	24	71.6 (3.09)	91.8 (3.49)	65.7 (2.77)	79.3 (3.69)
Perioptic	14	88.8 (3.67)	97.1 (1.62)	84.1 (1.97)	92.7 (1.32)
Cheeks	18	87.3 (2.45)	97.5 (1.51)	78.4 (3.62)	90.2 (2.82)
Nose	19	74.1 (2.56)	97.3 (1.38)	70.0 (1.99)	77.7 (2.27)
Mouth	13	81.1 (4.99)	91.9 (2.22)	66.3 (3.62)	84.3 (2.03)
Neck	19	82.5 (4.67)	93.8 (2.10)	72.9 (2.35)	86.8 (2.41)
Minutes 3-4-5 of Physical Task (Last Three Minutes of Exercise)					
Forehead	24	74.3 (4.39)	91.8 (3.07)	69.1 (2.07)	78.8 (4.49)
Perioptic	12	89.3 (2.23)	96.4 (2.08)	84.1 (1.78)	95.0 (1.13)
Cheeks	13	90.2 (3.06)	97.0 (0.84)	85.5 (1.97)	92.7 (2.30)
Nose	13	85.0 (4.14)	87.3 (2.93)	75.7 (2.55)	84.1 (2.30)
Mouth	13	80.7 (2.64)	86.6 (3.46)	76.4 (3.13)	85.2 (2.24)
Neck	25	74.8 (3.90)	91.1 (3.36)	72.7 (1.21)	74.3 (3.28)
Minutes 6-7-8 of Physical Task (Recovery Period After Exercise)					
Forehead	18	67.3 (4.45)	93.1 (3.83)	55.7 (2.79)	73.1 (3.73)
Perioptic	18	79.8 (4.64)	96.2 (2.01)	77.6 (2.79)	83.3 (2.75)
Cheeks	15	84.2 (4.50)	97.9 (0.86)	71.0 (2.70)	86.0 (2.62)
Nose	9	87.1 (2.30)	93.1 (2.27)	78.3 (2.37)	90.0 (1.51)
Mouth	10	90.0 (2.18)	93.0 (2.75)	87.9 (2.62)	93.1 (1.76)
Neck	15	72.0 (4.17)	91.6 (3.41)	64.0 (2.54)	78.8 (2.85)

To investigate the minimum amount of data collection time needed to achieve good discrimination between mental and physical stress states, a series of trials was performed with features from the first 10, 20, 30, 40, 50 and 60 seconds of data per task. Using data from 28 subjects, ten executions of the ANN with four-fold cross-validation were performed using each feature set. Accuracies increased from 79.1% when using the first 10 seconds of data, to 86.8% when using the first 50 seconds of data (Table 4.9). Analysis of variance (ANOVA) results were used to determine that a significant change in accuracy occurred as an incremental 10 seconds of input feature data were added from the 20-seconds to 30-seconds feature sets.

Table 4.9. Average accuracies of ten trials of four-fold cross-validation for feature sets ranging from 10 to 60 seconds of data per task, and the number of selected features after SFS. Statistical significance: * p < 0.05, significance reflects the difference from adjacent time epochs.

Amount of Time [s]	**Number of Features after SFS**	**Average Accuracy on Testing Set**
10	20	79.1 (6.13)
20	18	79.5 (3.60)
30	17	83.9 (3.95)*
40	18	83.9 (4.69)
50	16	86.8 (2.26)
60	16	85.9 (5.08)

4.4 Discussion

4.4.1 Mental stress classification

Among the various paradigms tested, the best results were achieved when discriminating the most highly stressed subjects from the remaining subjects. LDA was 100% accurate for this task, followed by ANN at 96.4% (Table 4.1). With only ten false negatives (FN) and no false positives (FP), the classifier is 87.5% sensitive and 100% specific. The SVM was 92.9% accurate, with 2 FN and no FP, making this classifier 75% sensitive and 100% specific. The NB classifier had the lowest accuracy of 82.14% (62.5% sensitive and 90% specific). In applications dealing with surveillance for potential threats, it would seem advantageous to favor a lower FN rate over a low FP rate, although if too many false alarms occurred, the screening would not be practical. The ten false negatives that were produced over ten executions of the ANN represented misclassifications of five out of the eight highly stressed subjects. Three subjects were misclassified only one time out of ten, one subject was misclassified twice, and one subject was misclassified five times. The subject who was misclassified in five out of the ten executions has the lowest composite *z*-score of the highly stressed subjects, Z_c = +0.734. This is acceptable since the cutoff of +0.5 between the neutral and high classes was chosen arbitrarily. With larger training data sets, the cutoff threshold could be adjusted, or dual thresholds could be identified to stratify subjects into "possible" and "probable" high stress states. Operationally, these identifications would be used in conjunction with other screening tools to identify threat.

The LDA classifier predicted 26 out of 28 subjects correctly using only the top two rank-based features (Figure 4.9). It appears that a parabolic boundary would provide a better decision boundary in this feature space, however, when classification was tested using quadratic discriminant analysis, the results were identical to those using LDA.

The classification of the high versus low stress classes using LDA and SVM resulted in 93.4% accuracy for both, with only one FP and no FN. The ANN gave an accuracy of 88.8% (Table 4.3) with 86.3% sensitivity and 91.3% specificity. Three highly stressed subjects were misclassified: one with a Z_c = +1.56 was misclassified once, and subjects with a Z_c = +0.75 and +1.11 were each misclassified in five out of the ten executions. These two subjects have composite z-scores in the lower half of the highly stressed class. Only two subjects in the low stress class were misclassified: one with a Z_c = -0.74 was misclassified once, and one with a Z_c = -1.19 was misclassified six times.

These misclassifications may be caused by anomalies in the physiological data resulting in the assignment of subjects to stress level classes in which they do not belong. The naïve Bayes classifier provided a comparable accuracy to the ANN in this case, 87.5%, with only one FP result and one FN result. Interestingly, LDA is considered the simplest of the four classifiers, requires the shortest computation time, yet consistently provided the highest classification accuracies. Although an unexpected finding, this is a fortunate advantage.

The LDA classifier was shown to be successful for the three class problem (high v. neutral v. low), achieving an accuracy of 89.3% (Table 4.4).When the ANN is used for the three class problem, the accuracy falls substantially to 59.6% (Table 4.5). The accuracies for the high and neutral classes were 65.0% and 65.8%, respectively, whereas the accuracy for the low class was 45.0%. Multiclass classifications can be problematic for the ANN as some algorithms are binary by nature and require special formulations to overcome this disadvantage [30]. These include extending the binary algorithms to handle the multiclass problem, or decomposing the multiclass problem into several binary classifications. Further, the decreased number of samples per class negatively impacts accuracy.

4.4.2 Mental v. physical stress classification

In the classification of mental versus physical stress, the greatest distinction between classes was observed using minutes 3-4-5 of the physical task, at a remarkable 98.9% using the ANN and LDA classifier, 98.6% for the SVM, and at 93.8% for the NB classifier (Table 4.6). Because Set 2 is taken from minutes 3-4-5 of the physical task, the response of physical stress is expected to be most apparent here. The first three minutes of the physical task was expected to provide lower discriminability between mental and physical stress responses, but the differences are still very observable within just a few minutes of exercise. In fact, using only 50 seconds of data per task, the classification accuracy reaches nearly 87%. Feature Set 3, taken from the three minutes after the physical task, represents a very practical challenge in human state assessment. The accuracy of the LDA classifier indicates that is it possible to classify a person who has just undergone a physical stressor, such as running or carrying a heavy object, versus a person that is mentally stressed, with nearly 98% accuracy. The persistence of the feature information post-physical task can be exploited in many security scenarios.

Table 4.7 shows consistent improvements in classification accuracies due to feature down-selection; ANN accuracies increased by 16-18% in each case. Although inputting more information to the classifier would be expected to increase accuracy, the poor generalization associated with very high dimensionality offsets the gain from additional information. Given that, at most, 11 features comprise the final feature set and these features can all be calculated with computational efficiency, the transition of the system to a near real-time, fieldable tool is quite reasonable.

Isolated facial regions can yield high accuracies during classification of mental versus physical stress (Table 4.8). For the first three minutes of exercise, the highest accuracies are observed using features from perioptic region within three out of the four classifiers. Features from the cheeks most commonly provided the highest accuracies for the last three minutes of exercise. The perioptic region was previously found to be active under mental and physical stress [9, 12] where an increased number of hot pixels was observed after two minutes of exercise. However, in these same studies, the cheeks were only mentioned as a region that undergoes cooling after a mental stressor. In contrast, we found evidence to support the use of the cheek data for task discrimination. Because the perioptic region may not be observable thermally if the subject is wearing glasses, we can focus on the cheek data and achieve comparable results. Further, unlike other regions of the face and neck, the cheeks are likely unobscured by clothing or hair. For the three minutes after exercise, the highest accuracies were achieved using features of the mouth region within three out four classifiers. This was an unexpected finding, but may be explained by changes in the breathing patterns, possibly due to reduction in RR during the recovery period, or due to the switch from mouth to nose breathing or vice versa.. Another explanation may be the gathering of sweat toward the chin and mouth region after exercise. Somewhat surprisingly, features in the forehead region yielded the lowest, or near-lowest, classification accuracy for each set. This is contradictory to previous findings [6] that described patches of the forehead that exhibit characteristic patterns of physical and emotional stressors.

The accuracy when using only the first 10 seconds of data was found to be 79.1% (Table 4.9). A statistically significant increase in accuracy is observed when using the first 30 seconds of data, which yields 83.9% overall accuracy. Extending the amount of time beyond this point did not significantly improve accuracy. These results demonstrate that only a short amount of data collection (ideally, about 30 seconds) is necessary to successfully discriminate between mental and physical stressors. This would be an attractive option for practical scenarios such as security checkpoint screening.

5. Vital Signs Detection

5.1 Procedure

The human subject pool and experimental procedure that are described in Chapter 4 were also used for the investigation of vital signs detection through non-contact sensors. Frequency analyses of the respiratory and cardiovascular pulse waveforms extracted separately from the thermal imagery and radar signal data provide estimations of respiratory rate (RR) and heart rate (HR). To validate these estimated parameters, ground truth values are obtained by frequency analyses of chest respiratory effort and ECG signals from the contact physiological sensors.

5.1.1 Discrete cosine transform filtering method

The experimentally measured non-contact vital sign signals contain corrupting noise from a variety of sources. A Discrete Cosine Transform (DCT) filtering method is applied to block the specific frequency ranges of the noise. The DCT converts the discrete time domain signal to a sum of cosine functions with different frequencies. This is similar to the commonly used Discrete Fourier Transform for frequency analysis, but differs in that the DCT-transformed signal contains real, rather than complex, values. The filtering operation is performed by assigning zeros to the indices of the DCT corresponding to the frequency range to remove. The inverse DCT is then performed to obtain the filtered time domain signal. This ideal frequency domain filtering method removes all unwanted frequencies but adds noise to the signal due to the discontinuities in the frequency domain created by multiplying with a rectangular function that lead to oscillations in the time domain signal, as is described by the Gibbs phenomenon. Despite this disadvantage, the DCT filtering method has shown far better performance for removing noise from biomedical signals in comparison to several practical filters [31].

Using the DCT filtering method, bandpass filters are applied to block all frequencies that are outside of the expected range for particular vital signs. The average respiratory rate for healthy adults is 12 to 18 breaths per minute (but may vary between individuals and due to medical conditions) [32], and so the respiratory signal filter cutoff frequencies were set to 0.08 and 0.7 Hz, equivalent to 4.8 and 42 breaths per minute. The typical resting heart rate for adults is 60 to 80 beats per minute (although variation between individuals and for medical conditions occurs) [32], and bandpass cutoff frequencies of 0.9 and 2.8 Hz, equivalent to 54 and 168 beats per minute, were applied to the cardiovascular pulse waveform.

5.1.2 Signal processing for thermal imagery

The respiration waveform is observed in thermal imagery as cyclic temperature fluctuations at the tip of the nose and nostril region (Segment 19, Figure 5.1). The mean pixel value of this segment in every frame represents the discrete-time signal for the estimation of RR. The cardiovascular pulse waveform is observed via temperature fluctuations of the skin over the external carotid artery region (Segments 27, 28, 29, Figure 5.1). Analysis of the performance of the temperature signals from the mean pixel values of these segments to estimate HR showed that Segment 29 achieved the best accuracy; this segment is therefore used in all further analyses.

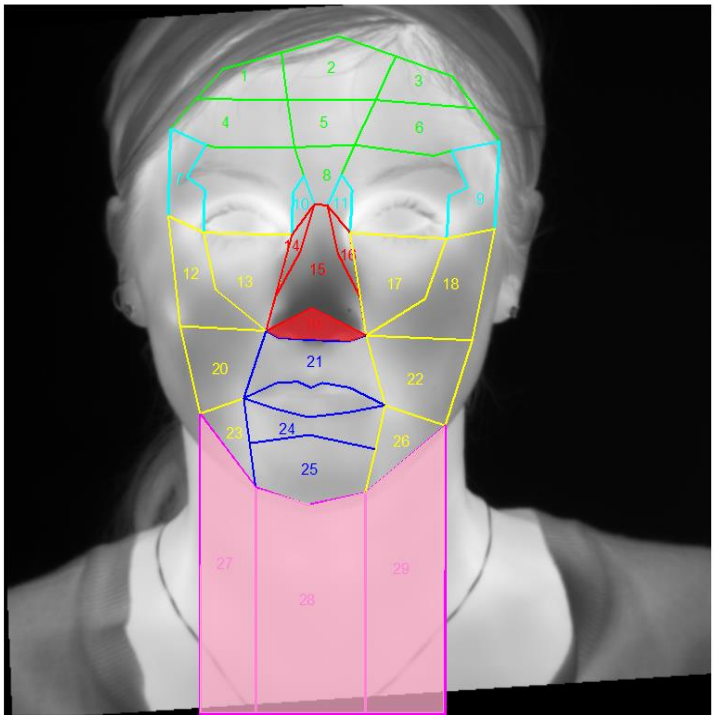


Figure 5.1. The mean pixel value of Segment 19 (shaded in red) is used to track the respiration waveform. Of the candidate segments (Segments 27, 28, and 29, shaded in pink) that can be used to track the cardiovascular pulse, Segment 29 was chosen for analysis.

The temperature signals from segments 19 and 29 are processed similarly for the estimations of RR and HR, respectively. First, interpolation is used to fill in missing data caused by poor tracking performance. For frames in which the subject's head is rotated to a large degree (>30°) away from the frontal pose, or in which the subject performs very quick and large head motions, the Visage software is unable to successfully track the facial feature coordinates. During feature extraction, coordinate values from these frames are set to NaN ("not a number") in the vector so that they may be interpolated at this step. Interpolation is applied using a least squares approach that does not modify any known values. Following interpolation, random noise is mitigated using a 10-point sliding-average filter.

Next, using the DCT filtering method, narrow bandpass filters are applied to the signals, with cutoff frequencies corresponding to the expected RR or HR range, 0.08 to 0.7 Hz and 0.9 to 2.8 Hz, respectively. As these are dynamic signals, the FFT is applied to the 30-second, non-overlapping windows of the temperature data and the frequency component with the highest magnitude is chosen as the RR or HR for that window (Figures 5.2-5.3).

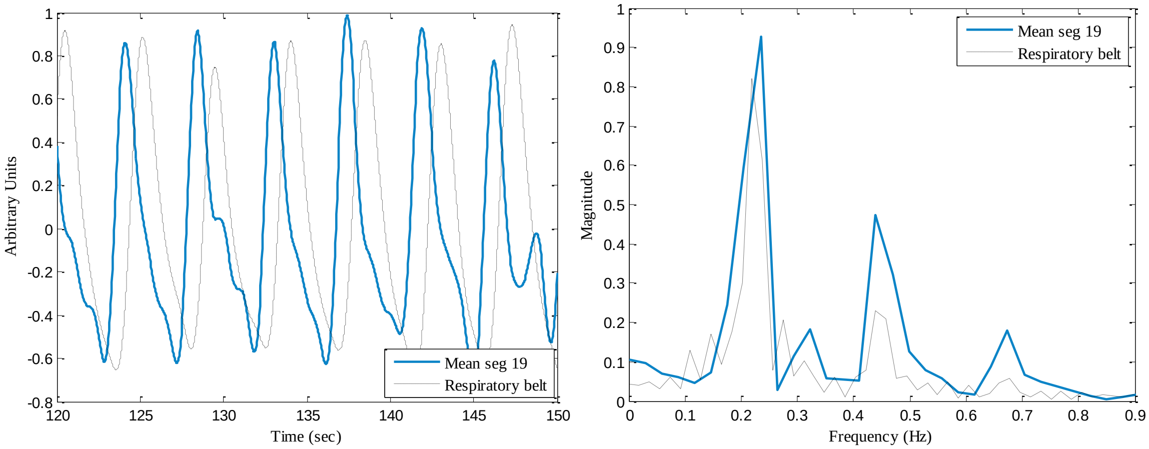


Figure 5.2. The respiration waveform extracted from thermal image Segment 19 and the respiratory belt signal for a 30-second window show good correlation aside from the delay in the belt signal (left). The dominant frequencies of the two signals (right) are selected and the difference between the assessed RR is calculated.

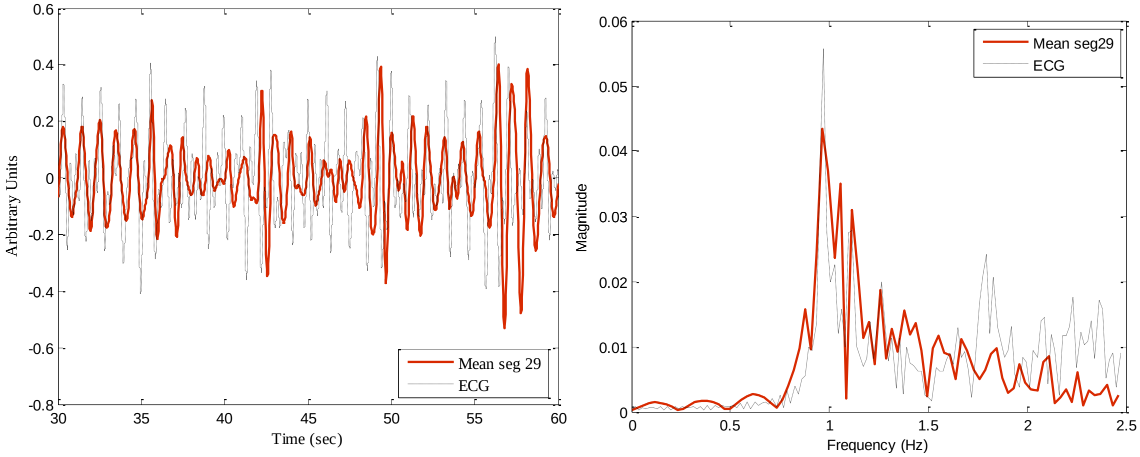


Figure 5.3. The cardiovascular pulse waveform extracted form thermal image Segment 29 and the ECG signal for a 30-second window show a matching number of peaks (left). The dominant frequencies of the two signals are used to calculate the difference between the assessed HR.

5.1.3 Radar signal processing

The in-phase (I) and quadrature (Q) radar signals are digitized and recorded during the experiment. Phase unwrapping is performed by taking the arctangent of the quotient of I and Q to provide a measurement of chest displacement. Both respiration and cardiovascular pulse waveforms are observed through the chest displacement signal, which is filtered with each of the appropriate sets of cutoff frequencies to obtain these signals of interest. Analysis proceeds as for the temperature signals from MWIR imagery,

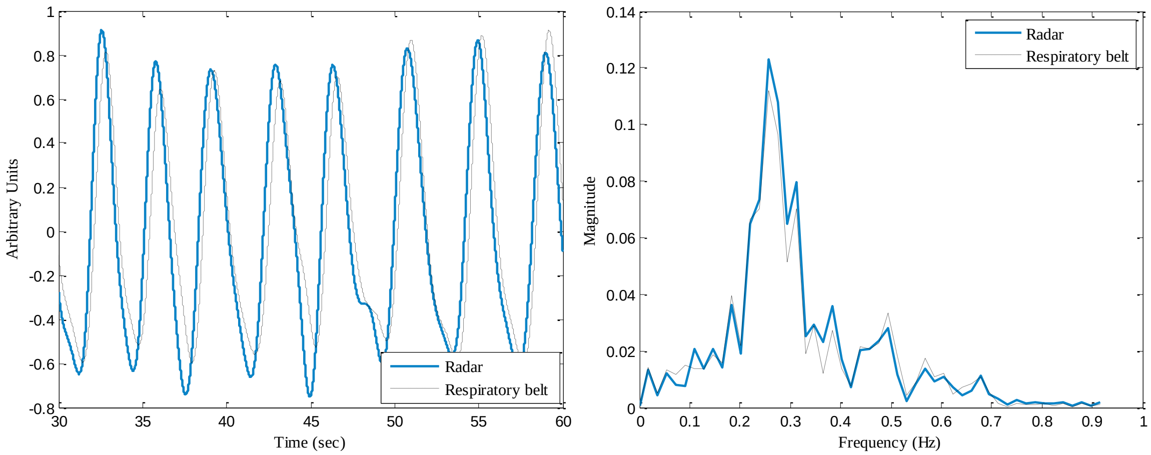


Figure 5.4. The respiration waveform extracted from the radar signal and the respiratory belt signal for a 30-second window (left). The dominant frequencies of the two signals (right) are selected and the difference between the assessed RR is calculated.

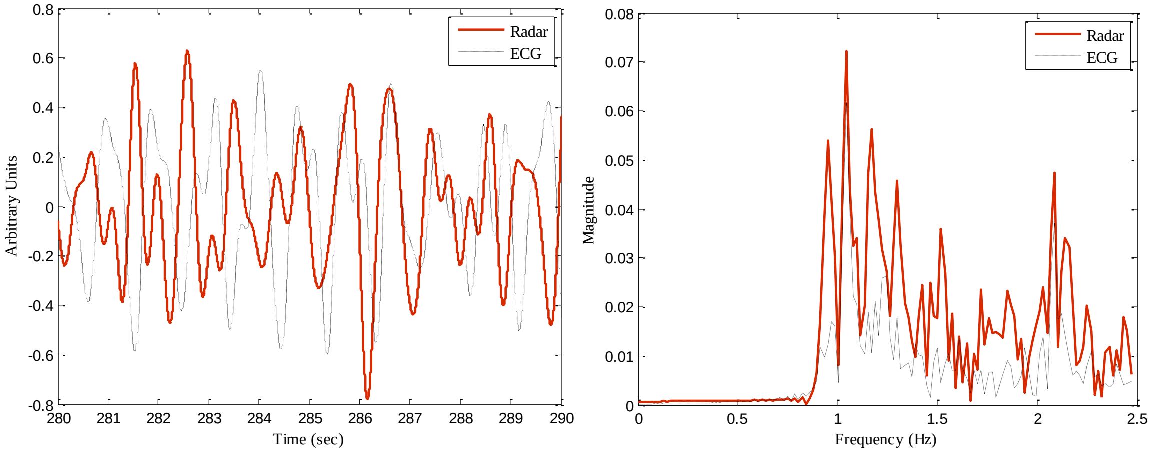


Figure 5.5. The cardiovascular pulse waveform extracted from the radar signal and the ECG signal for a 30-second window (left). The dominant frequencies of the two signals are used to calculate the difference between the assessed HR.

using the FFT of 30-second non-overlapping windows, and choosing the highest magnitude frequency component to represent the RR or HR (Figures 5.4-5.5).

5.1.4 Comparison to ground truth from contact sensors

The chest respiratory effort signal is filtered with the cutoff frequencies of the expected respiratory rate range, and the ECG signal is filtered with the cutoff frequencies for the expected heart rate range. Ground truth RR and HR values are obtained through the same process as the radar and MWIR signals using 30-second non-overlapping windows.

Frequencies are converted from units of Hz (s^{-1}) to min^{-1} by multiplying by 60 so that RR is reported in breaths per minute, and HR in beats per minute. For each of the 30 second windows, the difference between the ground truth and estimated values d_{RR} and d_{HR} are calculated as

$$d_{RR} = RR - RR', \text{ and} \tag{16}$$

$$d_{HR} = HR - HR', \tag{17}$$

where *RR* and *HR* are the ground truth values, and *RR'* and *HR'* are the estimated values from either the MWIR imagery or the radar signal. The accuracy of the estimated respiration rate is represented by the percentage of 30-second windows having a difference from the ground truth within either one or two breaths per minute. Similarly, the accuracy of the estimated heart rates is represented by the percentage of the windows having a difference from the ground truth within either two or six beats per minute. Accuracy estimates are calculated independently for each portion of the experiments: baseline, mental task and physical task.

5.2 Results

5.2.1 Vital sign estimation from thermal imagery

Using 30-second windows to analyze the estimated vital signs yields 150 windows for baseline, 168 for the mental task, and 387 for the physical task. The temperature signal from Segment 19 estimated the subject’s respiratory rate to within one breath per minute for 60.0% of the windows, and within two breaths per minute for 70.7% of the windows during baseline. During the mental and physical tasks, these percentages fell to 43.5% and 32.6%, respectively, for within one breath per minute, and 54.8% and 40.1%, respectively, for within two breaths per minute (Table 5.1). A distribution of all d_{RR} values displays more positive values than negative, indicating that the estimated RR is more often less than the ground truth RR (Figure 5.6).

The estimated HR from the temperature signal of Segment 29 was accurate to within two beats per minute for 12.7% and 8.93% of the windows, and within six beats per minute for 24.7% and 18.5% of the windows at baseline and mental task, respectively (Table 5.2). During the physical task, these percentages increased to 20.2% for within two breaths per minute, and 26.4% for within six breaths per minute. A distribution of all d_{HR} values shows that nearly 55% of the windows have difference greater than 18 beats per minute, meaning that the estimated HR values were much lower than the ground truth HR (Figure 5.6).

Table 5.1. Percentages of the estimated RR from the temperature signal having a difference *d* within one or two breaths per minute from the ground truth for the three experimental collections.

	Baseline	**Mental Task**	**Physical Task**
-1 < d_{RR} < 1 breath per minute	60.0%	43.5%	32.6%
-2 < d_{RR} < 2 breaths per minute	70.7%	54.8%	40.1%

Table 5.2. Percentages of the estimated HR from the temperature signal having a difference *d* within two or six beats per minute from the ground truth for the three experimental collections.

	Baseline	**Mental Task**	**Physical Task**
-2 < d_{HR} < 2 beats per minute	12.7%	8.9%	20.2%
-6 < d_{HR} < 6 beats per minute	24.7%	18.5%	26.4%

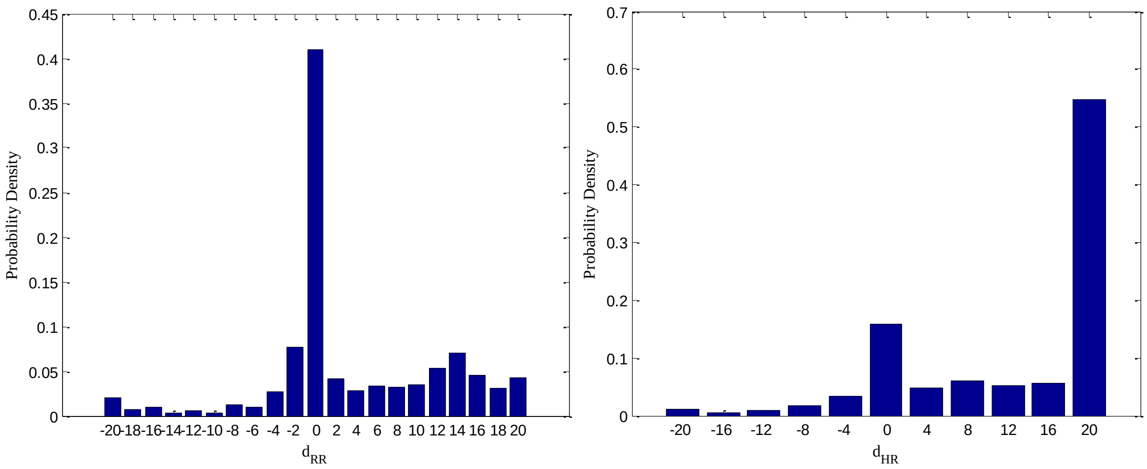


Figure 5.6. A distribution of the differences between ground truth RR and estimated RR (left) and between ground truth HR and estimated HR (right) by MWIR imagery for 30-second, non-overlapping windows across all tasks. The bin widths of each plot are set to the measured accuracy interval for each estimated parameter.

5.2.2 Radar vital signs

The radar signal estimated the subject’s respiration rate within one breath per minute for 67.1% of the windows and within two breaths per minute for 75.9% of the windows during baseline. These percentages decreased for the mental and physical tasks to 43.5% and 15.3% for within one breath per minute, and 51.2% and 21.9% for within two breaths per minute (Table 5.3). The distribution of all d_{RR} values display more positive values than negative, similar to the thermal imagery, so the estimated RR is more often lower than the ground truth (Figure 5.7). There are a total of 158 windows from baseline, 170 from the mental task, and 393 from the physical task.

The radar-estimated HR was accurate within two beats per minute for 34.2% of the windows, and within six beats per minute for 49.4% of the windows during baseline. During the mental and physical tasks the radar signal estimated the HR within two beats per minute for 22.9% and 32.8% of the windows, and within six beats per minute for 41.1% and 49.3% of the windows (Table 5.4). The distribution of all d_{HR} values shows that many of the windows had an estimated HR with a difference of more than 30 beats per minute from ground truth (Figure 5.7).

Table 5.3. Percentages of the estimated RR from the radar signal having a difference within one or two breaths per minute from the ground truth for the three experimental collections.

	Baseline	**Mental Task**	**Physical Task**
-1 < d_{RR} < 1 breath per minute	67.1%	43.5%	15.3%
-2 < d_{RR} < 2 breaths per minute	75.9%	51.2%	21.9%

Table 5.4. Percentages of the estimated HR from the radar signal having a difference within two or six beats per minute from the ground truth for the three experimental collections.

	Baseline	**Mental Task**	**Physical Task**
-2 < d_{HR} < 2 beats per minute	34.2%	22.9%	32.8%
-6 < d_{HR} < 6 beats per minute	49.4%	41.1%	49.3%

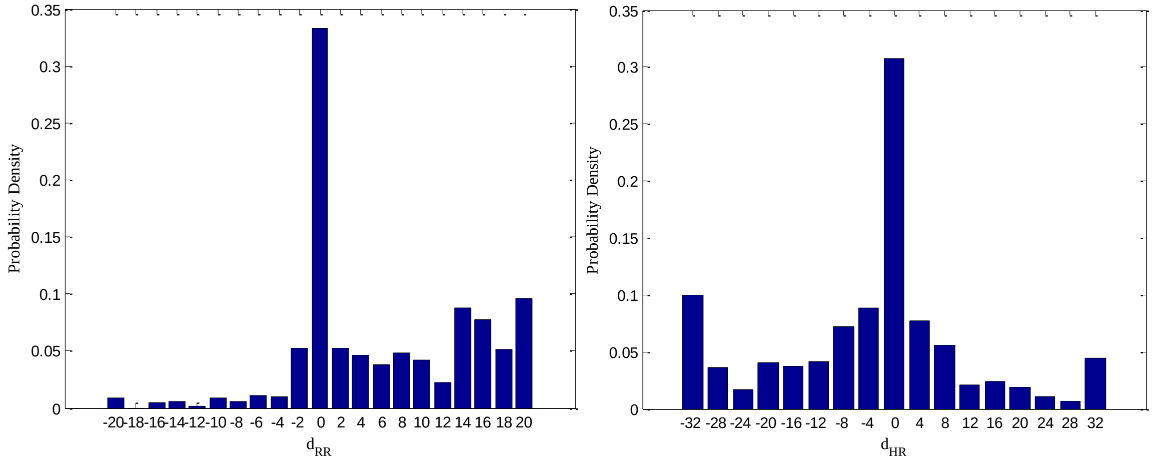


Figure 5.7. A distribution of the difference between ground truth RR and estimated RR (left) and ground truth HR and estimated HR (right) from radar for 30-second windows across all tasks. The bin widths are set to the measured accuracy interval.

5.3 Discussion

Detection of RR was similarly successful using thermal imagery and the radar signal for baseline and mental task data collections. However, during the physical task the RR detected by thermal imagery were more accurate than those from the radar system. This is not an unexpected result, due to the increased subject movement while pedaling the exercise bicycle during the physical task. Face tracking compensated for this movement during acquisition of the temperature signal, but the chest displacement signal from the radar was impacted by the additional motion that introduces noise in the respiratory pattern. Also, as subjects' RR increase during exercise, breathing tends to become more shallow, the chest displacement becomes smaller, and the SNR decreases.

The radar system was more successful than thermal imaging for detecting HR in all three data collections. Even after filtering, the cardiovascular pulse waveform of the temperature signal has low SNR since the neck segment covers a larger area of the skin than is necessary. A more focused ROI directly over the carotid artery would provide a better pulse waveform. Using thermal imagery, higher HR accuracy was achieved during the physical task than for baseline and the mental task. This is expected since subjects' HR increase and the heart's contractile force is greater during physical exercise, resulting in more blood being pumped through the carotid artery than when at rest, creating a higher SNR. The greater force of contraction also allows HR to be more successfully detected through the radar-acquired chest displacement signal during the physical task; the accuracy values were similar to those from baseline and greater than those from the mental task.

The differences between the ground truth and estimated RR and HR were often very large, indicating that the respiration or pulse waveform was not accurately detected or filtered, and that its frequency was not dominant in the signal. Instead, a noise frequency in the low frequency region of the filter's passband dominated the signal (Figure 5.10). The low cutoff frequency for the RR filter was 4.8 breaths per minute, so a RR around that value would often be selected when the respiration waveform was not detected in the signal. The HR filter had a low cutoff frequency of 54 beats per minute, which would very often be at least 20 beats per minute less than the true HR. Low frequency noise in the HR signal likely stems from subject motion that causes slight inaccuracies in segment tracking. Gibbs phenomenon may have also contributed low frequency oscillations that would affect the ability to correctly detect the RR or HR frequencies.

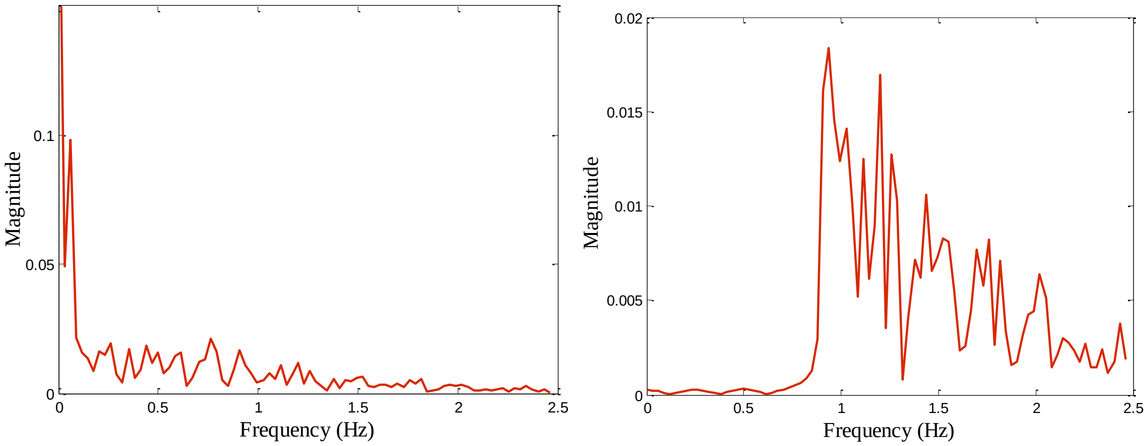


Figure 5.8. Unfiltered (left) and filtered (right) spectra of thermal segment 29 show that a peak at the lower frequency region of the passband, being the dominant frequency, would be selected as the estimated HR.

To compensate for such errors when an artifact or noise peak dominates the frequency spectrum, manual peak selection may be performed. This is accomplished by first identifying the frequency spectrum peaks for cases where the difference between ground truth (or expected RR or HR) and extracted frequencies are very large. In instances where a secondary peak appears in the normal expected range for the vital sign frequency, it may be selected to replace the more dominant noise peak (Figure 5.9). These manual overrides were performed for the extraction of HR and RR from thermal imagery during the baseline collections; 35% of HR windows and 17% of RR windows were updated. The accuracy of the estimated RR improved to within one breath per minute for 72.0% of the windows and to within two breaths per minute for 86.7% of the windows (Table 5.5), a more satisfactory result. The accuracy of the estimated HR improved to within two beats per minute for 26.7% of the windows and to within six beats per minute for 53.3% of the windows (Table 5.6).

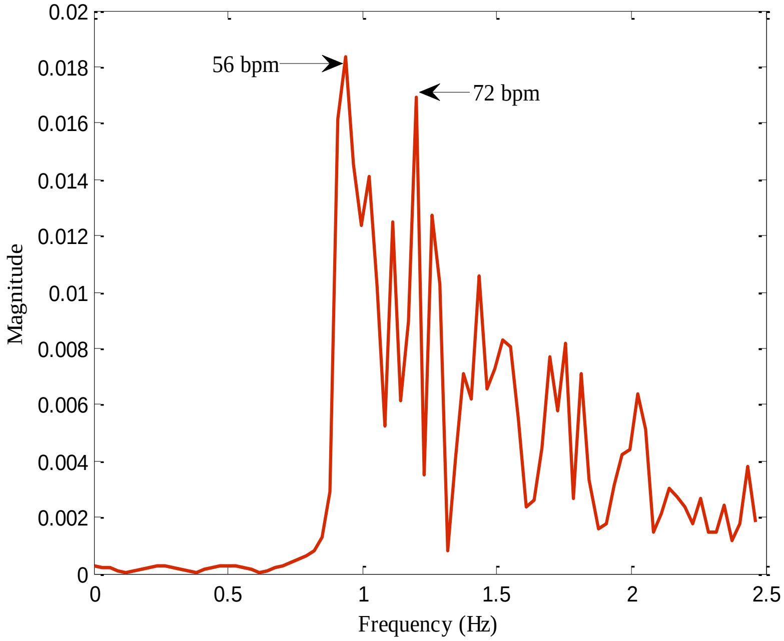


Figure 5.9. The low frequency peak (56 bpm) is contributed by noise in the low region of the passband. The 72 bpm peak is manually selected to provided a better estimate of HR.

Table 5.5. Percentages of the estimated RR from the temperature signal having a difference within 1 or 2 breaths per minute from the ground truth at baseline for original and adjusted results.

	Baseline	**Baseline (Adjusted)**
-1 < d_{RR} < 1 breath per minute	60.0%	72.0%
-2 < d_{RR} < 2 breaths per minute	70.7%	86.7%

Table 5.6. Percentages of the estimated HR from the temperature signal having a difference within 2 or 6 beats per minute from the ground truth at baseline for original and adjusted results.

	Baseline	**Baseline (Adjusted)**
-2 < d_{HR} < 2 beats per minute	12.7%	26.7%
-6 < d_{HR} < 6 beats per minute	24.7%	53.3%

Although acquisition conditions were more challenging and more stringent metrics were used to evaluate performance, the results of this experiment to detect vital signs through thermal imagery were not as successful as reported in previous works [15-18]. However, several aspects of our techniques make them more feasible for practical applications. For instance, no manual ROI selection or initialization was needed and the region was continuously tracked throughout the collection, allowing subjects complete freedom of motion within the camera's field of view. Also, the frontal view of the face was used to detect both RR and HR, even though this is a non-ideal pose for HR detection. This experiment also demonstrated the ability to unobtrusively detect vital signs during mental and physical tasks, which was not investigated in other works.

6. Operational Limitations and Confounding Factors

6.1 Procedure

Three separate data collections were conducted with 12 subjects each (6 male, 6 female) to evaluate the impacts of confounding factors on the stability of thermal signatures relevant to stress detection. The collections included investigations of day-to-day variability, impacts of distance and pose, and effects of activation of facial muscles. For these studies, only image data were collected (i.e., no physiologic or radar signals were gathered). Data from the 32 subjects who performed the mental and physical tasks were also used to test the impact of topical skin products on thermal signatures. The data processing procedures of image registration, facial feature tracking, and segmentation (described in Chapter 3) were followed by analyses unique to each of the different collections.

6.1.1 Day-to-day variability

These experiments were designed to explore both intra-day and longer-term variability under normal conditions. Image data were collected of each subject three times per day (morning, afternoon and evening time periods) for five consecutive days. Subjects were seated and remained still for approximately one minute while being imaged. From the 15 collections, the (temporal) mean is computed over ten seconds of the (spatial) mean pixel data of each of the 29 segments. To compare segment stability, the coefficients of variation (CV) from each of the 12 subjects are averaged to obtain the mean CV by segment; this is referred to as the *pooled* CV. Intra-day variances are calculated using the three mean pixel values of a segment from same-day collections. Therefore, each subject has five intra-day variances per segment, and these are averaged across the 12 subjects. To examine the impact of data collection time, the CV of each the morning, afternoon, and evening collections is calculated using the five mean pixel values of a segment, and averaged across the 12 subjects.

6.1.2 Distance and pose

At distances of 5, 10, 15, 20, and 25 feet, subjects were imaged while standing at angles of 0°, ±30°, ±60°, and ±90° to the camera, with 0° being normal (frontal pose) to the camera. The ability to analyze the images at a particular position depends on the success of the facial feature tracker in locating the subject’s face and overlaying the subject-specific mask. The mask is created for a subject using a neutral frame from the image set taken at 5 feet and 0°, and is then applied to the remaining image sets to initialize tracking. If the tracker is unsuccessful in overlaying the mask, that image set is not analyzed.

Each segment’s mean pixel value is averaged over ten seconds for successfully tracked image sets. The 29 segments are grouped into seven facial regions (Figure 6.1), and a mean for each region is computed. The variances in the mean pixel values at different distance and pose angles are computed by region for each subject, and the mean variance across the 12 subjects is found.

6.1.3 Activation of facial muscles

Image data were collected for 12 subjects after each of three trials of inflating a balloon. Following a one-minute baseline collection, subjects were given roughly two minutes to inflate a balloon using their breath and were imaged for one minute immediately after the balloon was inflated. A brief rest period followed. The balloon inflation task was repeated two more times with images collected after each trial. The average pixel values for the seven segment regions (Figure 6.1) were computed for the first ten seconds of image data taken in each of the three trials. Average baseline values, taken from the last ten seconds of the baseline collection, are subtracted from the average pixel values of each trial.

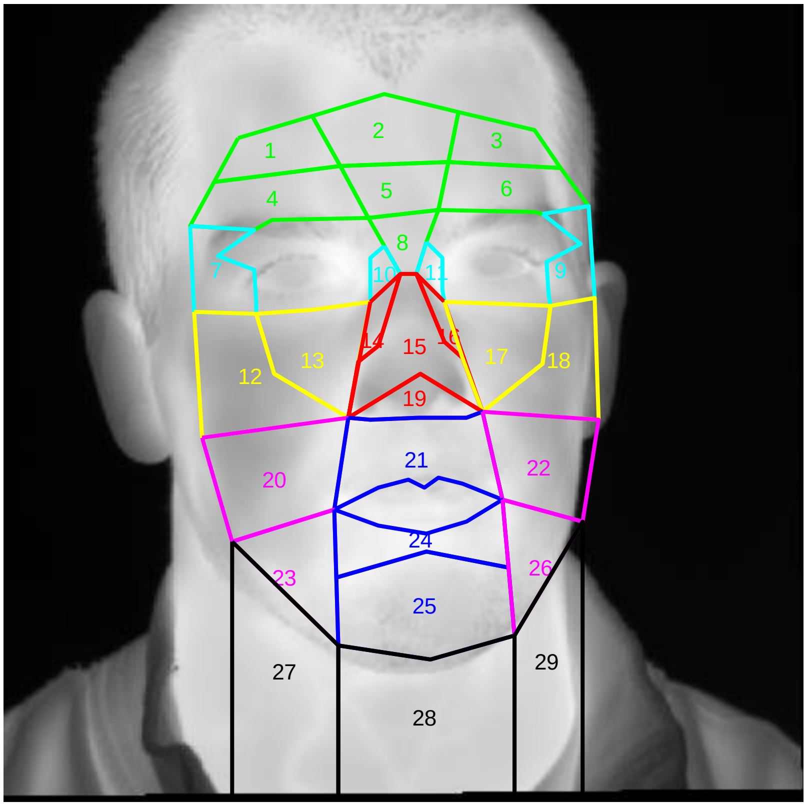


Figure 6.1. Grouping of segments into seven regions: forehead (green), perioptic (cyan), nose (red), upper cheeks (yellow), lower cheeks (pink), mouth (blue), and neck (black).

6.1.4 Impact of topical skin products

To examine the impact of topical skin products, 16 of the 32 participants of each sex (8 males and 8 females) were selected at random to apply one of four products to the right side of his or her face prior to the experiment. The products included an aftershave balm (Nivea Double Action Balm), SPF 30 sunblock (Neutrogena Ultra Sheer), and two cosmetic products: powder (CoverGirl Aquasmooth) and liquid foundation (Revlon PhotoReady). The subjects were then imaged according to the protocol in Chapter 4, including baseline, mental task, and physical task data collections. A 10-second average of the mean (spatial) pixel data in the cheeks and forehead regions is calculated to determine if the product causes a significant apparent temperature difference.

6.2 Results

6.2.1 Day-to-day variability

The segments having the highest pooled CVs across the 15 collections, relative to the rest of the segments, are (in order from highest to lower values) 19, 15, 14, and 16, which are all from the nose region (Figure 6.2). The ordered lowest pooled CVs (lowest to higher) are of segments 9, 10, 7, and 11 of the perioptic region. Although this region exhibits the lowest variability, the CVs are not substantially lower than the mean CV of the face.

Segments 19 and 15 have intra-day and pooled variances that are substantially higher than the remaining segments, and segments 16, 14, and 21 also have high intra-day variance (Figure 6.2). Four of these five segments are from the nose region, and one is from the region just below the nose and above the mouth. The pooled and intra-day CVs show good correlation, R = 0.989, with the intra-day variances respectively lower than the pooled CVs for all segments.

The variations across the collection time periods were found to be significantly different ($p = 0.045$), precluding the combination of a day’s three CV value for calculating inter-day variance. Instead, data from the morning, afternoon and evening were analyzed independently. The CVs across the morning, afternoon, and evening data collections are highly correlated (R = 0.781, 0.899, 0.937, respectively). Again here, the nose segments (15 and 19) consistently have the highest variation and the perioptic segments (7, 9, 10 and 11) the lowest variation (Figure 6.3). Notably, the relative magnitude of the variation for the nose segments is lower in the morning versus the afternoon and evening. Further, a majority of the segments exhibit lower variance in the afternoon collection versus the other two time periods.

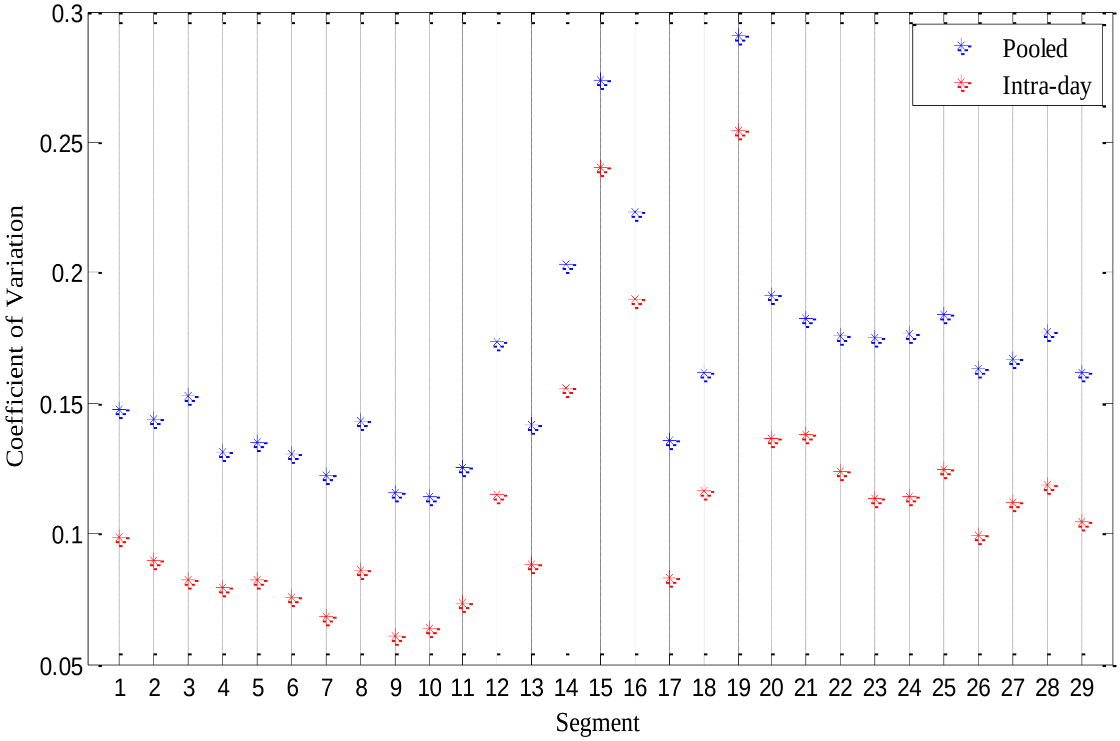


Figure 6.2. The pooled and intra-day CVs are the greatest for segments 19 and 15 of the nose. Segments 7, 9, 10, and 11, of the perioptic region are the lowest.

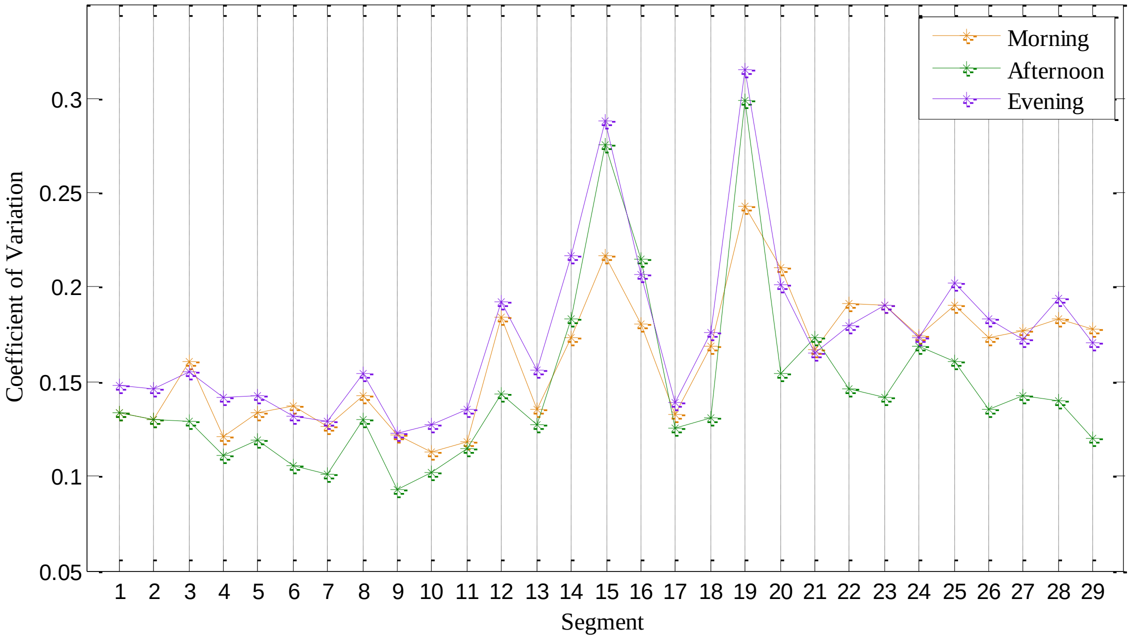


Figure 6.3. The aggregate CVs for the morning, afternoon, and evening collections are significantly different but are highly correlated. Again, segments of the nose region have the highest variation and segments of the perioptic have the lowest variation.

An alternate analysis was performed using three-factor ANOVA to determine whether there was a significant source of variation in the mean pixel values due to the time of collection, day, or subject. The returned p-values by segment for each of the factors indicate significance if $p < 1\text{-}\alpha$, where we chose $\alpha = 0.95$. The time of collection is found to be a significant factor for all segments except for Segments 1, 4, 5, 8, 10, 11, 13, and 17, representing the forehead, perioptic, and upper cheeks (Figure 6.4). The day of collection is found to be a significant factor for all segments except for Segments 15, 16, and 19, representing the nose region, as well as Segment 21, located just below the nose (Figure 6.5). A significant source of variation is found across subjects for every segment, which is expected due to person-to-person variability.

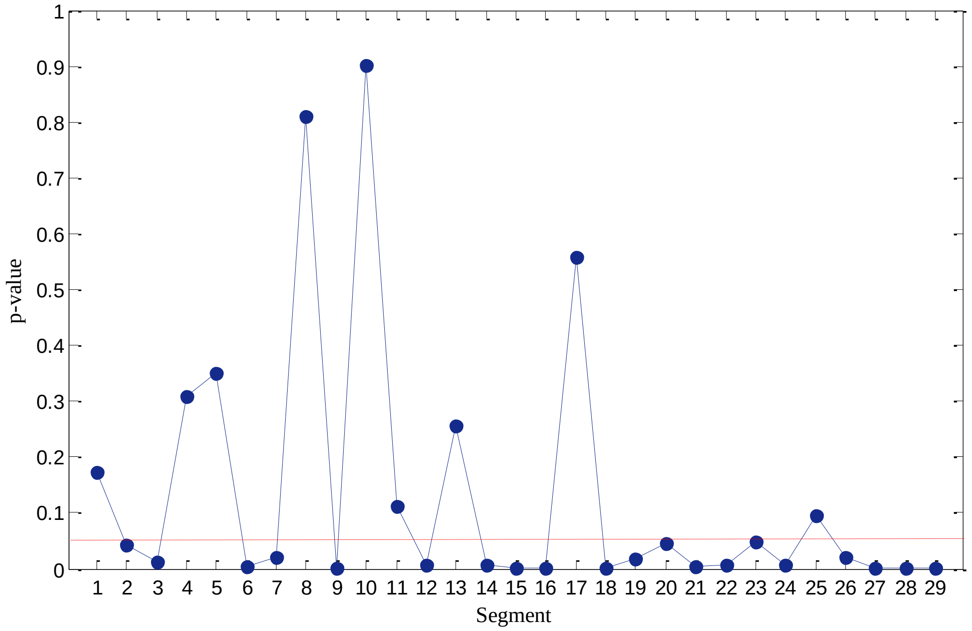


Figure 6.4. *p*-values by segment, returned from three-factor ANOVA, that indicate the significance of the *time* of collection as a source of variance. The dashed line is drawn at *p* = 0.05. For segments of the forehead, perioptic, and upper cheek region, collection time was not a sigificant factor.

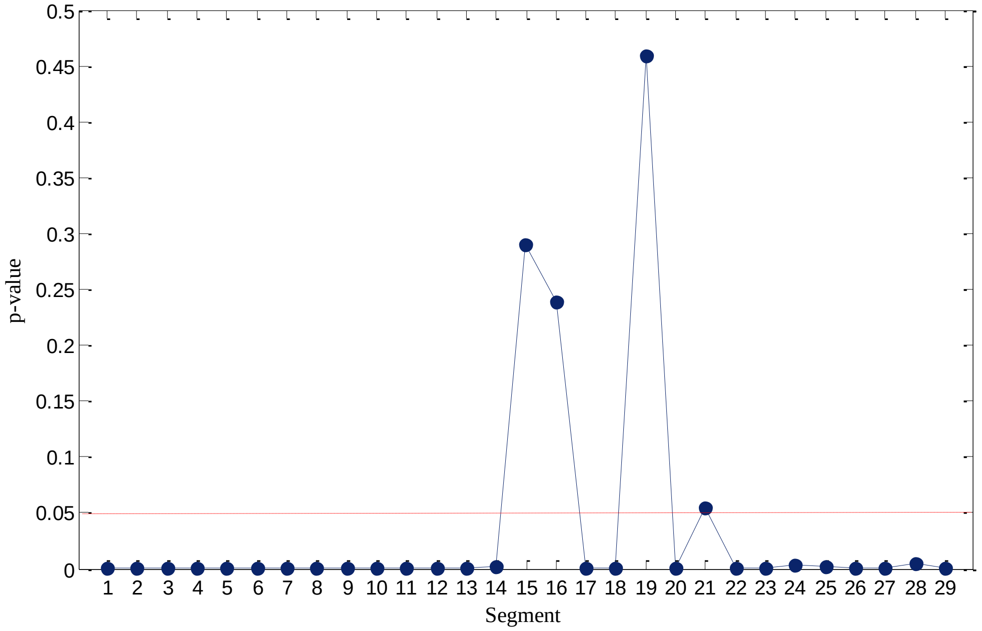


Figure 6.5. *p*-values by segment, returned from three-factor ANOVA, that indicate the significance of the *day* of collection as a source of variance. The dashed line is drawn at *p* = 0.05. For segments of the nose and the segment just below the nose, day of collection was not a significant factor.

6.2.2 Distance and pose

Facial feature tracking was not successful at distances over 15 feet for all subjects, and for angles greater than 30 degrees for all subjects. Images of all 12 subjects were successfully tracked for distances of 5 and 10 feet at the 0° pose, and images of 10 subjects were successfully tracked at 15 feet and 0° pose (Table 6.1). At 10 feet, images of all 12 subjects were successfully tracked for the ±30° poses.

Table 6.1. Number of subjects successfully tracked at each distance and pose angle from the imager, where 0° corresponds to the frontal pose.

Pose Angle [°] Distance [ft]	0	± 30	± 60	± 90
5	12	8 both, 4 either	0	0
10	12	12 both	0	0
15	10	4 both, 4 either	0	0
20	2	0	0	0
25	0	0	0	0

The largest mean CV of the pixel values at 5, 10 and 15 feet occurs in the nose region, 5.08%, followed by the forehead region at 4.20% (Table 6.2). Whereas the mean pixel value in the nose region at the three distances is fairly similar for most subjects, the spread is very large for a few subjects (Figure 6.6, top). Also, there is no apparent trend between the pixel value and subject-to-imager distance for any of the segments. The regions having the smallest mean CVs at distances of 5, 10 and 15 feet are the mouth, upper and lower cheeks.

The perioptic region has the largest mean CV for pose angles of -30°, 0°, and +30° at 10 feet, 4.55%, followed by the nose region at 3.55%. In comparison to the effect of increasing distance, the mean pixel value of the perioptic region is somewhat more

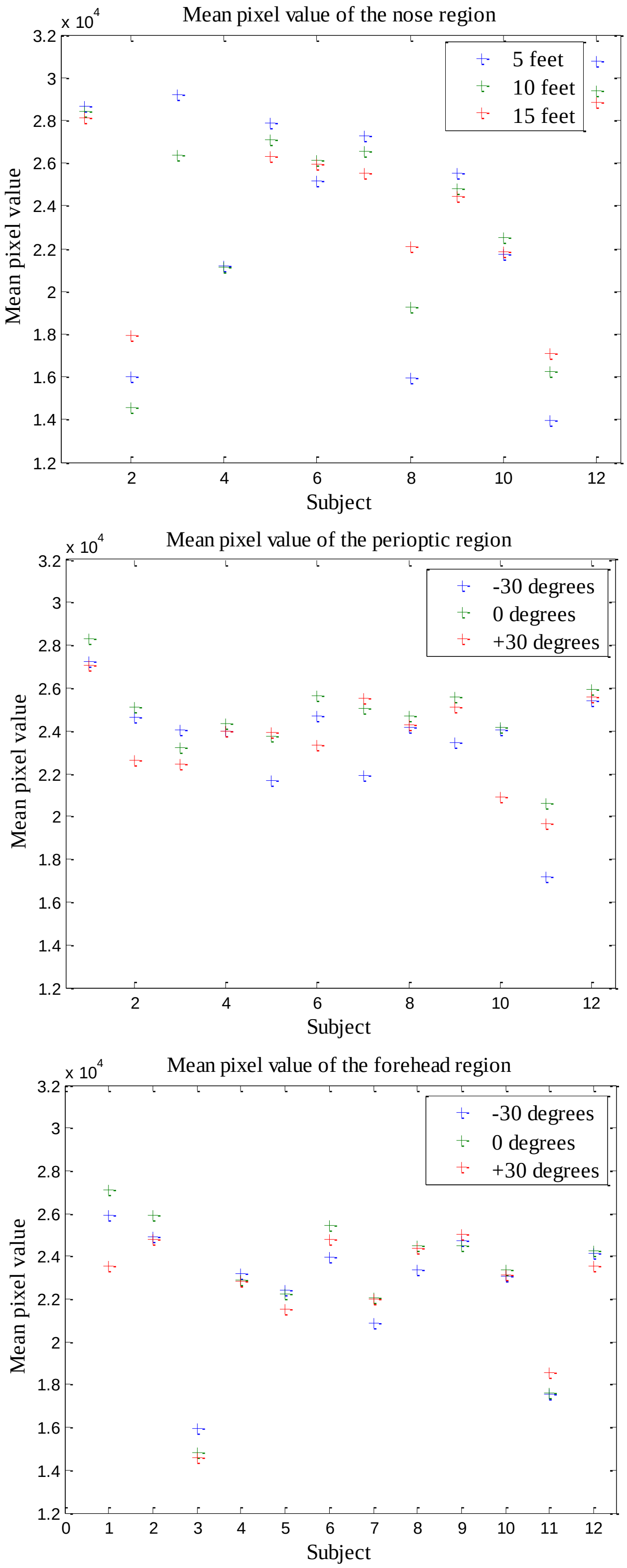


Figure 6.6. Mean pixel values of the nose region (top) having the highest CV across distances, and the forehead region (middle) and perioptic region (bottom) having the highest and lowest CVs across poses, respectively.

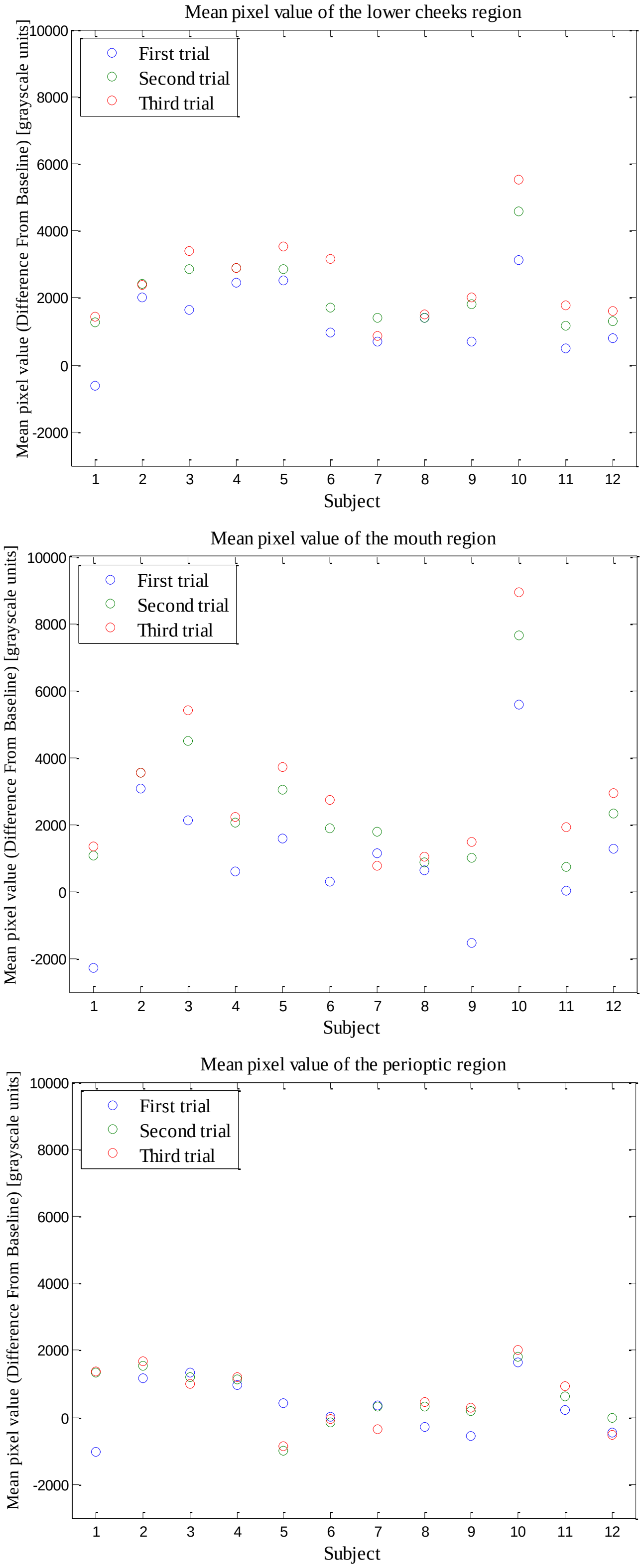
consistent over pose angle (Figure 6.6, middle); the forehead region shows even more consistent values over pose angle (Figure 6.6, bottom). The mouth, upper cheeks, and forehead regions have the smallest mean CVs under these pose angle (2.24 to 2.96%). As the imaging distance is constant, the range of mean CVs was small.

Table 6.2. Mean CV across 12 subjects of the pixel values in a given region, at distances of 5, 10, and 15 ft, and pose angles of -30°, 0°, and 30° at 10 feet.

Region	Mean CV for Distances	Mean CV for Angles
Forehead	0.0420	0.0268
Perioptic	0.0403	0.0455
Upper cheeks	0.0344	0.0296
Nose	0.0508	0.0355
Lower cheeks	0.0355	0.0302
Mouth	0.0342	0.0224

6.2.3 Activation of facial muscles

As expected, after the first balloon inflation trial, the lower cheeks region showed the greatest change in pixel value from baseline, with an average increase of 1,353 grayscale units, corresponding to a temperature increase of about 0.71 °C, across 12 subjects (Figure 6.7, top). The change in grayscale units are converted to the change in temperature using a linear fit to the calibration data. Although the actual calibration curve is nonlinear, the region lying within the observed skin temperature range is nearly linear. The mouth region showed the second greatest change, corresponding to about 0.55 °C (Figure 6.7, middle). Except for the perioptic region, where the average pixel intensity decreased slightly from Trial 2 to Trial 3, all regions showed temperature increases between subsequent trials (Figure 6.7, bottom). After the second and third trials, the


Mean pixel value of the lower cheeks region

Mean pixel value of the mouth region

Mean pixel value of the perioptic region

Figure 6.7. Mean pixel values (difference from baseline) of the lower cheek (top), mouth (middle), and perioptic (bottom) regions for 12 subjects, after each trial of blowing up a balloon. The lower cheeks and mouth showed the greatest temperature increase, while the perioptic region showed the least.

mouth region showed the greatest changes in temperature, 1.33 °C and 1.57 °C, respectively, and the lower cheeks showed the second greatest changes, 1.12 °C and 1.31 °C, respectively. The forehead region showed the smallest changes in pixel values for each of the three trials, and the perioptic region showed the second smallest changes (Table 6.3).

Table 6.3. Mean pixel value (difference from baseline) across 12 subjects, after each trial of blowing up a balloon. Segments are merged together into seven regions.

Segment region	Trial 1	Trial 2	Trial 3
Forehead	145	464	532
Perioptic	322	615	601
Upper cheeks	841	1318	1529
Nose	722	1281	1804
Lower cheeks	1353	2140	2508
Mouth	1043	2543	3007
Neck	640	1148	1282

6.2.4 Impact of topical skin products

Baseline, mental task, and physical task data, with four subjects wearing each of four products, were compared. Using a paired *t*-Test, $\alpha = 0.95$, with the null hypothesis being that there is no temperature difference between left and right sides of the face, only the liquid-based makeup caused an apparent difference in skin temperature during the mental task ($p = 0.0039$) and physical task ($p = 0.0373$), and only in the region of the cheeks, and the increase in skin temperature was less than 1 °C for each subject.

6.3 Discussion

The segments found to have the highest day-to-day variability are those within the nose region. Mainly consisting of cartilage tissue, and having no major blood vessels in the proximity, the nose does not exhibit good thermoregulation, and the overlying skin is greatly affected by activity and ambient conditions. The room temperature during the experimental collections varied within a few degrees which would contribute to nose segment temperature variability. More likely, of greater impact are the effects of a subject's activities prior to the arrival. As our data collection was conducted in the month of January, subjects may have spent time outdoors just before the experimental session. Collection during different phases of respiration, inspiration or expiration, also likely contributes to variability. Temporal averaging may have captured most of either one of the phases for a particular collection, which was not controlled.

The perioptic region was found to have the lowest day-to-day variability. In contrast to the nose, this region exhibits very stable temperatures under normal conditions due to having a minimal amount of tissue between the superficial vessels and overlaying skin. Therefore, this region remains close to core body temperature and may only experience variations due to responses of the autonomic nervous system. Within this region, Segment 9, containing the area to the outside of the left eye, was observed to have the lowest intra-day and pooled variances (Figure 6.8, left). Segment 19, the tip of the nose and nostril region, was observed to have the highest intra-day and pooled variances (Figure 6.8, right).

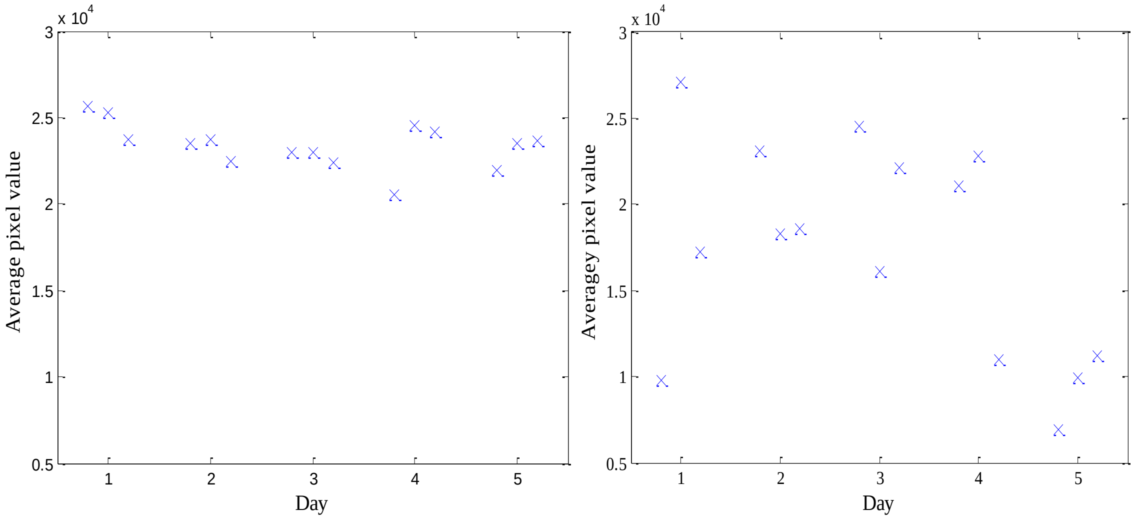


Figure 6.8. Segment 9 of the perioptic region has the lowest variability across all subjects, and Segment 19 of the nose has the highest variability. Here, the mean pixel values of Segment 9 (left) and Segment 19 (right) during 15 collections over five days are shown for a particular subject.

The facial feature tracker was unsuccessful in tracking subjects at distances greater than 15 feet and at angles greater than 30°. These findings are consistent with the limitations quoted by the developer [33]. The minimum allowable size of the face in the image is approximately 80 pixels wide, which is roughly the observed size at the distance of 20 feet. The developer also quotes that head rotation is tracked up to approximately 45°, in agreement with our findings that tracking was unsuccessful at 60° or greater. Our expectation is that tracking would be more successful for a continuous head rotation from the frontal pose to a larger degree, rather than discretized collections at each angle.

Our approach to segmenting the neck region was developed for images collected at a subject-to-imager distance of six feet, and was invalid at greater distances. Due to the unavailability of fiducial coordinates on the neck output by the tracker (at any distance), vertical lines drawn from four tracked points along the jaw and chin line to the bottom of the image were used to segment the neck region into three rectangular-shaped segments

At distances greater than six feet, a large portion of the subject's upper body is, therefore, included in these segments. The background subtraction step did not suppress these regions because the pixel values in the chest region are more similar to those of the face than to background pixel values.

The highest variance observed in the image sets at varying distances was found to be from the nose region. As previously mentioned, the variability of the nose is likely impacted by the ten-second collection time occurring during the inspiration or the expiration phase. As this factor cannot be controlled in real-world scenarios, our experiments did not aim to consistently collect data during one or the other phase. The cheeks and mouth regions were found to have the smallest variances in image sets at varying distances. These regions are composed of segments with large areas and good separation between the coordinates, allowing the tracker to consistently locate these regions even when their size is decreased. The perioptic region was found to have the highest variance across the image sets at varying pose angles. The small size and short distances between coordinates of these segments, which become even shorter as the head is turned, cause the tracker to be more inaccurate in locating these regions.

Overall, each of the regions showed very low variability with varying distances and pose angles. All CVs were less than approximately 5%. The expected effect of increased subject-to-imager distances was decreased pixel values due to a loss of energy by atmospheric absorption of photons along the transmission path [3]. This trend was not found in our results, but is likely to be more of an issue at distances greater than 100 ft.

The activation of facial muscles after the trials of balloon inflations caused pixel values in the mouth and lower cheeks regions to increase by the greatest amount. All other regions also displayed an increase in pixel value from baseline after each trial. Even though facial muscles may not be activated in each of the regions, the increase of blood flow to the active muscles causes temperatures in all regions of the face to increase. The forehead and perioptic segment regions displayed the smallest increase in pixel values, as expected from being the furthest distance away from the active muscles around the mouth.

7. Conclusions

7.1 Summary of Findings

This research has evaluated the performance of thermal imaging for the detection of physiological indicators of stress in humans. Results of three separate investigations were given that describe the ability to detect and discriminate stress through thermal imaging, the utility of stand-off modalities to monitor physiological vital signs that are related to stress responses, and the stability of thermal signatures in terms of the limitations and possible confounding factors in this application.

Feature-based classification was found to be successful for the two-class problems including paradigms of high versus low stress as well as mental versus physical stress. Four classifiers were tested (ANN, NB, LDA and SVM) with LDA providing the highest classification accuracies. LDA achieved 100% accuracy in classifying high mental stress, and nearly 99% in the classification of mental versus physical stress (the other classifiers were also successful). Sequential feature selection was found to be very useful in reducing the dimensionality of the feature space, which ultimately yielded faster computation times and better performance. This also allowed for the identification of salient features and specific facial regions related to expected physiological indicators of stress.

We demonstrated that thermal imagery is capable of detecting vital signs (RR and HR) at stand-off distances in a similar manner to that of the mmW radar system, although not at the performance levels reported in previous works. However, our method utilized a full face tracker that requires minimal manual input, and evaluated an ROI with a larger area than found necessary in previous methods. Further, in contrast to previous works, we conducted these experiments under more realistic operating conditions, collecting vital sign data during mental and physical tasks. Relative to baseline measurements, accuracy typically decreased during task performance, with the exception of HR detection during the physical task, which was more accurate than baseline measurement due to higher cardiac output that improved SNR.

In the investigation of the stability of human thermal signatures, the nose segments were found to have the greatest day-to-day variability and the perioptic segments were found to be the least variable. This is a desirable result as the perioptic region was found to provide useful features for the detection of stress. Because this region is less susceptible to environmental and inter-individual differences and sensitive to stress-related changes, keying on this region is advised. The one caveat is that this region is unavailable in subjects who are wearing glasses, as the thermal signatures will not be transmitted through glass or polycarbonate materials. The distance and pose at which the subject may be positioned from the imaging system was found to be limited by the facial feature tracker. Subjects' faces were not able to be located at distances greater than 15 feet and pose angles greater than 30°, consistent with the tracker's specifications. All regions had very low coefficients of variation (typically less than 5%) across collections at varying distances and pose angles. The smallest variances were observed in the cheek and forehead regions, making these areas suitable for evaluation under realistic surveillance conditions. As predicted, the mouth and lower cheeks showed the greatest temperature increase after balloon inflation. This effect propagated to all other facial regions, confirming that local muscle activation is a confounder in the interpretation of facial thermal signatures. Liquid makeup was the only skin product to confound thermal signatures. Although the emissivity of the product was not measured, its primary component, water, has an emissivity (0.99), which is slightly greater than that of human skin. This increased emissivity would contribute to the apparent increase in skin temperature.

7.2 Suggestions for Future Work

Rather than the Stroop test, the use of a more realistic scenario for the mental stress-inducing task might generate results that could be extended to surveillance domains. One such experiment is the guilty knowledge technique (GKT) which has been reported to produce valid physiological responses for a subject attempting to be deceptive [34]. This experiment involves a subject enacting one, both, or neither of two mock-crimes. Afterward, the subject is interrogated with questions related to one of the crimes. Such an experiment engages the subject, when guilty, into recalling things that were truly witnessed, providing a more realistic response to stress. Further improvements to non-contact vital signs detection would include the implementation of a peak detection algorithm to provide a weighted average of frequency peaks within an expected range, rather than selection of a single dominant frequency. The expected frequency range may be narrowed by using the estimated values of prior time windows in a continuous monitoring application. Also, a more focused ROI over the external carotid would increase SNR of the cardiovascular pulse signal. Increasingly challenging test conditions could be tested to assess system robustness. The use of moving subjects and the introduction of scene clutter and obscurations would present additional challenges to tracking and feature extraction.

Appendix

Table A.1. Calibration measurements for temperatures ranging from 19.9 to 38 °C. Radiance values are calculated with a source emissivity of 0.95, source area of 25 cm^2, and source distance of 170 cm. The pixel value is a spatial average of an ROI on the blackbody.

Temperature (°C)	**Radiance (W/(sr·cm^2))**	**Pixel value [14 bit]**
19.9	0.000145	8490.715
20.9	0.000150	8701.461
21.9	0.000155	8923.866
23	0.000161	9177.802
23.9	0.000167	9383.433
25	0.000173	9651.062
26	0.000179	9903.951
27	0.000186	10161.342
28	0.000192	10427.250
29	0.000199	10703.815
30	0.000206	10987.870
31	0.000213	11279.208
32	0.000221	11573.264
33	0.000228	11888.059
34	0.000236	12200.358
35	0.000244	12533.391
36	0.000252	12865.278
37	0.000261	13217.797
38	0.000270	13572.122

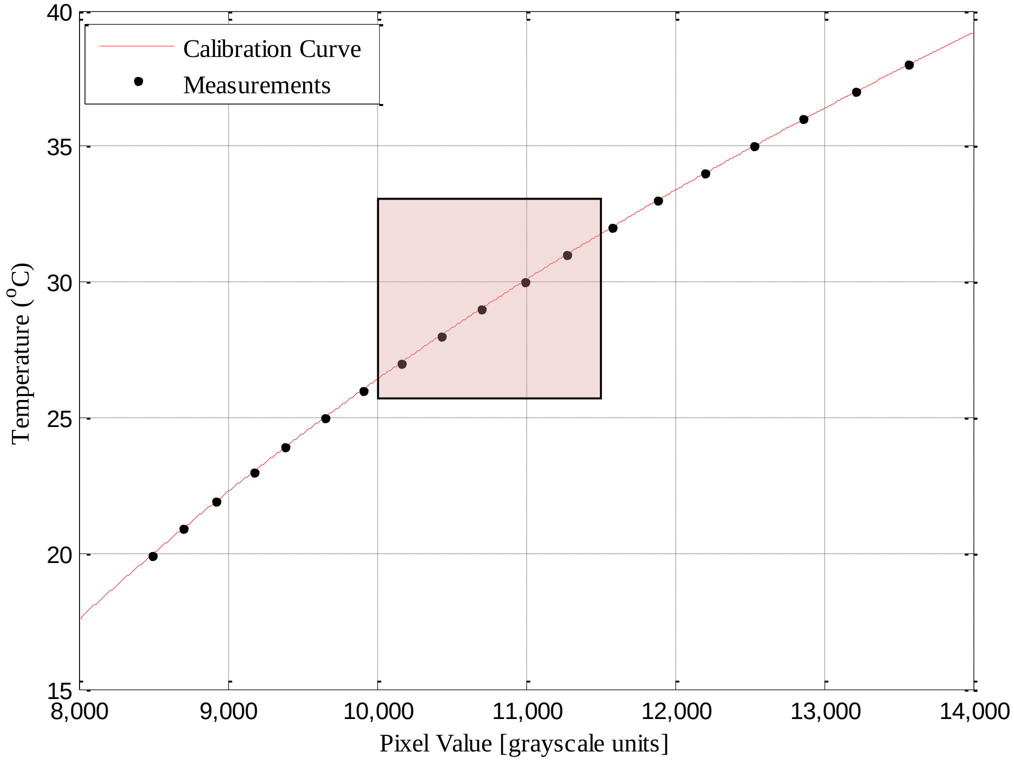


Figure A.1. Calibration curve, 3rd degree polynomial fit, to the temperature and pixel value measurements of the blackbody. The shaded square represents the normal range of pixel values measured, where a linear fit is approximated for conversion to temperatures.

Table A.2. For each physiological variable (HR, GSR and RR), the difference from baseline for minutes 1, 2, and 3 of the mental task. Normalized data are given as a Z-score by subtracting the mean of each column and dividing by the standard deviation. The composite Z-score is calculated by normalizing the sums of the nine individual Z-scores. Shaded cells indicate manual adjustments made to the extracted RR.

Subj. #	Raw Data									Normalized Data									Z Sum	Z Composite
	Δ HR [beats/m]			Δ GSR [μSiemens]			Δ RR [breaths/m]			Z HR			Z GSR			Z RR				
1	-3.49	-4.28	-1.87	0.93	0.86	0.88	18.13	13.18	13.73	-1.18	-1.25	-0.91	1.96	1.92	1.79	2.06	1.37	1.23	6.99	1.59
2	13.37	7.76	9.72	0.11	0.11	0.01	N/A	3.49	2.75	0.58	0.41	0.47	-0.75	-0.66	-0.92	-1.44	-0.67	-0.91	-3.89	-0.88
3	5.05	3.34	1.38	0.08	0.04	0.01	1.65	4.39	4.94	-0.29	-0.20	-0.52	-0.85	-0.90	-0.92	-1.12	-0.48	-0.48	-5.77	-1.31
4	2.60	0.35	4.27	1.16	1.11	1.19	10.44	9.89	10.44	-0.54	-0.62	-0.18	2.72	2.79	2.74	0.58	0.68	0.59	8.76	1.99
5	7.52	7.32	8.42	0.29	0.29	0.30	3.30	2.75	3.30	-0.03	0.34	0.32	-0.17	-0.04	-0.01	-0.80	-0.83	-0.80	-2.02	-0.46
6	9.68	3.84	9.99	0.61	0.58	0.61	10.98	5.49	11.53	0.19	-0.14	0.51	0.88	0.95	0.95	0.69	-0.25	0.80	4.58	1.04
7	2.22	0.27	0.24	0.32	0.21	0.24	3.29	6.04	8.24	-0.58	-0.63	-0.66	-0.05	-0.30	-0.19	-0.80	-0.13	0.16	-3.19	-0.72
8	7.96	5.97	4.99	0.20	0.21	0.24	3.85	-0.55	2.20	0.01	0.16	-0.09	-0.46	-0.32	-0.18	-0.69	-1.52	-1.02	-4.12	-0.94
9	-5.02	-4.69	-1.77	0.25	0.24	0.31	6.04	6.59	2.75	-1.34	-1.31	-0.90	-0.29	-0.21	0.02	-0.27	-0.02	-0.91	-5.22	-1.19
10	0.87	-0.68	0.59	1.13	1.08	1.24	5.49	3.85	6.04	-0.72	-0.76	-0.62	2.61	2.71	2.90	-0.37	-0.60	-0.27	4.88	1.11
11	23.02	9.34	7.40	0.22	0.10	0.10	6.59	6.04	5.49	1.58	0.62	0.20	-0.39	-0.69	-0.63	-0.16	-0.13	-0.38	0.02	0.01
12	-1.76	2.26	5.35	0.15	0.14	0.16	-0.55	1.10	0.55	-1.00	-0.35	-0.05	-0.62	-0.54	-0.43	-1.54	-1.18	-1.34	-7.06	-1.60
13	-6.61	-9.08	-6.55	0.05	0.07	0.05	18.68	19.78	20.32	-1.50	-1.92	-1.47	-0.95	-0.80	-0.79	2.17	2.76	2.52	0.02	0.00
14	2.33	3.33	4.22	0.56	0.36	0.07	0.55	2.19	-2.75	-0.57	-0.21	-0.18	0.75	0.20	-0.72	-1.33	-0.94	-1.98	-4.99	-1.13
15	2.85	6.82	1.04	0.31	0.30	0.28	6.59	10.44	10.44	-0.52	0.28	-0.57	-0.10	-0.01	-0.08	-0.16	0.79	0.59	0.22	0.05
16	28.23	24.73	17.34	0.32	0.30	0.33	9.33	6.04	9.34	2.12	2.74	1.38	-0.06	0.00	0.08	0.37	-0.13	0.38	6.88	1.56
17	0.85	-1.51	-2.19	0.50	0.46	0.42	7.14	7.69	7.14	-0.73	-0.87	-0.95	0.53	0.54	0.37	-0.06	0.21	-0.05	-1.00	-0.23
18	24.71	11.56	11.72	-0.03	-0.05	-0.06	9.89	9.89	10.44	1.76	0.93	0.71	-1.23	-1.19	-1.11	0.47	0.68	0.59	1.60	0.36
19	24.40	10.11	11.86	0.33	0.33	0.32	14.83	16.48	16.48	1.72	0.73	0.73	-0.03	0.10	0.06	1.43	2.07	1.77	8.57	1.95
20	22.71	20.63	36.11	0.14	0.13	0.16	2.20	2.20	5.49	1.55	2.18	3.63	-0.67	-0.59	-0.46	-1.01	-0.94	-0.38	3.31	0.75
21	5.85	3.92	1.50	0.32	0.26	0.28	14.83	13.18	15.38	-0.21	-0.12	-0.51	-0.06	-0.15	-0.07	1.43	1.37	1.56	3.23	0.73
22	7.88	6.38	7.06	0.33	0.30	0.25	4.39	3.85	3.85	0.01	0.21	0.16	-0.04	0.01	-0.15	-0.59	-0.60	-0.70	-1.69	-0.39
23	-4.35	-4.04	-6.41	0.46	0.41	0.47	9.34	9.89	9.34	-1.27	-1.22	-1.45	0.41	0.37	0.52	0.37	0.68	0.38	-1.22	-0.28
24	10.55	10.54	8.77	0.12	0.09	0.08	4.94	4.94	8.24	0.28	0.79	0.36	-0.72	-0.70	-0.68	-0.48	-0.37	0.16	-1.36	-0.31
25	5.40	3.42	3.62	0.11	0.09	0.09	13.73	3.85	2.20	-0.25	-0.19	-0.26	-0.77	-0.73	-0.66	1.22	-0.60	-1.02	-3.27	-0.74
26	15.38	10.70	9.19	0.12	0.07	0.09	7.14	1.65	4.39	0.78	0.81	0.41	-0.74	-0.79	-0.65	-0.06	-1.06	-0.59	-1.88	-0.43
27	8.26	0.41	1.22	0.25	0.17	0.16	8.79	6.59	7.69	0.04	-0.61	-0.54	-0.30	-0.43	-0.44	0.26	-0.02	0.05	-1.98	-0.45
28	8.81	6.30	14.17	0.16	0.15	0.19	6.59	6.04	7.69	0.10	0.20	1.00	-0.59	-0.53	-0.35	-0.16	-0.13	0.05	-0.40	-0.09

Table A.3. Features selected by SFS for the classification of mental stress under our three different paradigms. MEAN, MAX, and TTM are the raw features of a segment's mean pixel value, maximum pixel value, or mean of the top 10% hottest pixels, respectively. S1-S3 are the slopes of the feature data for minutes 1, 2, or 3, and M1-M11 are the means of the 30-second windows with a 15-second slide.

Segment	High v. Remaining	High v. Low	High v. Low v. Neutral
1	MEAN M5		
2			
3	TTM S1		MAX S1 MAX S2
4			
5			
6		MEAN M2	
7		TTM S3	
8			
9			
10			
11			
12			
13			
14			
15		MAX S3	
16			
17			
18	TTM M1	MEAN M2	
19			
20			
21			
22		MEAN M7	
23	TTM S1		TTM S1
24			MEAN M7 MEAN M9
25			
26	MAX S1		MEAN M11 MAX S1
27			
28	MAX M8	MAX S3	MAX M3
29	MAX M3		

Table A.4. Features selected by SFS for the classification of mental versus each set of physical data, representing three different time spans of the physical task. MEAN, MAX, and TTM are the raw features of a segment's mean pixel value, maximum pixel value, or mean of the top 10% hottest pixels, respectively. S1-S3 are the slopes of the feature data for minutes 1, 2, or 3, and M1-M11 are the means of the 30-second windows with a 15-second slide.

Segment	Beginning of exercise (Minutes 1-2-3)	End of exercise (Minutes 3-4-5)	After exercise (Minutes 6-7-8)
1	MEAN S1		
2			
3	MEAN M6		
4			
5			
6			
7	MAX S1		
8			TTM S2
9			TTM M1
10	TTM M4	TTM S1	
11	MEAN M9 TTM M1	MEAN M7	MAX S2 MAX M5 TTM M3
12			
13			
14			
15	TTM S1	MEAN S1	MAX S2
16			
17		MEAN M1	
18			
19			TTM M7
20	MEAN S2	MAX M3	
21			MEAN S3
22			
23			
24	MEAN S1	MEAN S1 MEAN M10	
25			
26			MAX M7
27	MEAN S1	MAX M6	
28	MAX M9	MAX M2	MAX M4 TTM S3
29			

References

[1] R.R. Seeley, T.D. Stephens and P. Tate, *Anatomy & Physiology,* New York: McGraw-Hill, pp. 9, 580-581, 2008.

[2] B.F. Jones and P. Plassmann, "Digital infrared thermal imaging of human skin," *IEEE Engineering in Medicine and Biology Magazine,* 21(6), pp. 41-48, 2002.

[3] H. Kaplan, *Practical Applications of Infrared Thermal Sensing and Imaging Equipment*, 3rd ed., Bellingham: SPIE Press, pp. 17-19, 24-25, 2007.

[4] F.P. Incropera, D.P. DeWitt, T.L. Bergman and A.S. Levine, *Fundamentals of Heat and Mass Transfer*, Hoboken: John Wiley & Sons, pp. 737-739, 744, 2007.

[5] J. H. Hong, J. Ramos and A. K. Dey, "Understanding physiological responses to stressors during physical activity," *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pp. 270-279, 2012.

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

[7] A. Barreto, J. Zhai and M. Adjouadi, "Non-intrusive physiological monitoring for automated stress detection in human-computer interaction," Lecture Notes in Computer Science, *Human-Computer Interaction,* pp. 29-38, 2007.

[8] Y. Shi, M.H. Nguyen, P. Blitz, B. French, S. Fisk, F. De la Torre, A. Smailagic, D.P. Siewiorek, M. al' Absi, E. Ertin, T. Kamarck and S. Kumar, "Personalized stress detection from physiological measurements," *International Symposium on Quality of Life Technology*, pp. 28-29, 2010.

[9] I. Pavlidis, J. Levine and P. Baukol, "Thermal imaging for anxiety detection," *Proceedings of the IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, pp. 104-109, 2000.

[10] C. Puri, L. Olson, I. Pavlidis, J. Levine and J. Starren, "StressCam: non-contact measurement of users' emotional states through thermal imaging," *CHI Extended Abstracts on Human Factors in Computing Systems*, pp. 1725-1728, 2005.

[11] G. Seematter, M. Dirlewanger, V. Rey, P. Schneitter and L. Tappy, "Metabolic effects of mental stress during over- and underfeeding in healthy women," *Obesity*, 10(1), pp. 49-55, 2002.

[12] K. Hong, P. Yuen, T. Chen, A. Tsitiridis, F. Kam, J. Jackman, D. James, M. Richardson, W. Oxford, J. Piper, F. Thomas and S. Lightman, "Detection and classification of stress using thermal imaging technique," *Proceedings of the SPIE Europe Security+Defense Conference*, 748601, pp. 1-9, 2009.

[13] D.T. Petkie, E. Bryan, C. Benton and B. D. Rigling, "Millimeter-wave radar systems for biometric applications," *Proceedings of the SPIE Europe Security+Defense Conference*, 748502, pp. 1-7, 2009.

[14] D.T. Petkie, C. Benton and E. Bryan, "Millimeter wave radar for remote measurement of vital signs," *IEEE Radar Conference*, pp. 1-3, 2009.

[15] J. Fei, Z. Zhu and I. Pavlidis, "Imaging breathing rate in the CO_2 absorption band," *27th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 700-705, 2005.

[16] J. Fei and I. Pavlidis, "Thermistor at a distance: unobtrusive measurement of breathing," *IEEE Transactions on Biomedical Engineering*, 57(4), pp. 988-998, 2010.

[17] M. Garbey, N. Sun, A. Merla and I. Pavlidis, "Contact-free measurement of cardiac pulse based on the analysis of thermal imagery," *IEEE Transactions on Biomedical Engineering*, 54(8), pp. 1418-1426, 2007.

[18] S.Y. Chekmenev, A.A. Farag, W.M. Miller, E.A. Essock and A. Bhatnagar, "Multiresolution approach for noncontact measurements of arterial pulse using thermal imaging," *Augmented Vision Perception in Infrared*, London: Springer-Verlag, pp. 87-112, 2009.

[19] ThermoVision SC6700 User's Manual, FLIR Systems Inc., Wilsonville, OR, 2009.

[20] Basler A202k User's Manual, Basler AG, Ahrensburg, Germany, [Online] 2007, http://www.baslerweb.com/2/9/4/9/A202k_Users_Manual.pdf (Accessed: 10 January 2013).

[21] H.V. Karvir, "Design and validation of a sensor integration and feature fusion test-bed for image-based pattern recognition applications," Ph.D. dissertation, Wright State University, Dayton, OH, pp. 72-75, 2010.

[22] P. Viola and W.M. Wells III, "Alignment by maximization of mutual information," *Proceedings of the IEEE Conference on Computer Vision*, pp. 16-23, 1995.

[23] Ranging Sensor Heads, Ducommun Technologies Inc., Carson, CA, [Online], http://www.ducommun.com/pdf/SRR-Series.pdf (Accessed: 22 March 2013).

[24] SRR-35121010-D1 Data Sheet, Ducommun Technologies Inc., Carson, CA, [Online], http://www.ducommun.com/pdf/smd/SRR-35121010-D1.pdf (Accessed: 22 March 2013).

[25] J.H.M. Tulen, P. Moleman, H.G. Van Steenis and F. Boomsma, "Characterization of stress reactions to the Stroop Color Word Test," *Pharmacology, Biochemistry & Behavior*, 32(1), pp. 9-15, 1989.

[26] R.A. Robergs and R. Landwehr, "The surprising history of the "HRmax = 220 – age" equation," Journal of Exercise Physiology Online, 5(2), 2002.

[27] R.F. Thompson and W.A. Spencer, "Habituation: a model phenomenon for the study of neuronal substrates of behavior," *Psychological Review*, 73(1), pp. 16-43, 1966.

[28] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, Burlington: Academic Press, 2009.

[29] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," *IEEE International Joint Conference on Neural Networks*, 3(1), pp. 17-21, 1990.

[30] M. Aly, "Survey on multiclass classification methods," Technical Report, California Institute of Technology, pp. 1-9, 2005.

[31] S.H. Sik, C. Lee and M. Lee, "Ideal filtering approach on DCT domain for biomedical signals: index blocked DCT filtering method," *Journal of Medical Systems,* 34(4), pp. 741-753, 2010.

[32] D.C. Dugdale, (2011, Feb 20), "Vital Signs," *Medline Plus Medical Encyclopedia* [Online] Available: http://www.nlm.nih.gov/medlineplus/ency/article/002341.htm (Accessed: 25 March 2013).

[33] FaceTrack, Visage Technologies Face Tracking & Animation, [Online] 2012, http://www.visagetechnologies.com/products/visagesdk/facetrack (Accessed: 10 January 2013).

[34] D.T. Lykken, "The validity of the guilty knowledge technique: the effects of faking," *Journal of Applied Psychology*, 44(4), pp. 258-262, 1960.