https://doi.org/10.2352/ISSN.2470-1173.2017.13.IPAS-207

Thermal Facial Signatures for State Assessment during Deception

Nilesh U. Powar, Tamera R. Schneider, Julie A. Skipper, Douglas T. Petkie, Vijayan K. Asari, Rebecca R. Riffle, Matthew S. Sherwood and Carl B. Cross

Abstract

This study explored emotional and physiological states in response to stressful situations involving deception, and their relation to facial imaging. Male participants were evaluated (i.e., appraised) impending stressors. Stressor appraisals ranged from threat (i.e., appraisals that there are a lack of sufficient resources to meet the demands of a stressor) to challenge (i.e., appraisals that there are adequate coping resources to manage stressor demands). We used appraisals to discriminate two appraisal groups: threat versus challenge. The capacity to detect state changes of the human face was examined using several standoff sensing technologies, including 2D visible (VIS) and mid-wave infrared (MWIR) cameras. Psychophysiological determination of the human state was achieved through self-assessment and physiological measures which formed the ground truth for the classifier systems. Two deception studies, a false opinion study and a false behavior study, served as high-stakes stressors to understand facial changes with respect to human stress states. The methodology extracted MWIR statistical features from facial regions in response to changes in stress state. Using the statistical features from MWIR sensor and implementing Eigen analysis allowed classification of the threatened and challenged participants for the false behavior study with an accuracy of 85%.

Introduction

In ancient China, lie detection included putting dry rice in the mouth of the accused. Upon spitting it out, those with rice sticking to their tongue were considered guilty. Stress causes a dry mouth, preventing enough saliva to help spit out all of the rice. Presumably, those guilty would suffer more stress than innocents. Assessment of human emotional states is a complex multidisciplinary field with theoretical roots in psycho-physiology. To empirically investigate emotional states, scientists trained in psychology (e.g., clinical, social, neuroscience and cognitive) utilize different assessments including self-reports, physiological monitoring and observation, to name a few. To study and categorize emotions, for example, self-reports assess emotional experience, physiological assessments assess blood flow and sympathetic nervous system or amygdala activity, and behaviors indicate approaching or avoiding a stimulus. Our research aims to analyze the face during a high-stakes deception situation, using past-validated psycho-physiological metrics as well as advanced imaging and pattern recognition techniques to predict the human emotional state via non-contact methods.

The use of non-contact sensing for detecting intentions to deceive, if valid, can be of great value. People often rely on nonverbal cues of deception (eye contact patterns, nervous laughter, and hurried speech) and are rarely accurate, with success rates hovering around 50 percent [3]. Some deceivers will have concerns that they will be found out, leading to high levels of anxiety [20]. Complicating the accurate detection of deception is that some deceivers will be unconcerned with discovery and have low emotional reactions, whereas those innocently accused may have high levels of anxiety. Consequently, we assume that stress reactivity varies under deception conditions, because the act of interrogation can be a stressor itself.

The stress process begins with a persons evaluation or appraisal of an event, and people appraise situations differently, causing their stress responses to differ [18], [19]. Appraisals range on a continuum from challenge to threat. Challenge appraisals result when people evaluate an impending situation as personally relevant and demanding, but one with which they may be able to cope and even gain mastery. Threat appraisals result when people evaluate the demands of the situation as outweighing their ability to cope; they may become overwhelmed. Research shows that challenged participants experience more positive and less negative affect than threatened participants [18], [19]. These two groups have also been distinguished by their physiological pattern differences. During a stressor, challenged participants have increased cardiac output into a more accepting vasculature, whereas threatened participants have somewhat more blood pumped into a more resistant, constricted vasculature [18],[19],[12]. High-stakes situations are stressors that engage stress responses in participants. Research indicates that threat and challenge appraisals initiate the stress process and are robust predictors of various stress responses. Stressor appraisals are appropriate for discerning the human state and emotional activity. We examined stress responses such as appraisals, affect, and physiology and investigated the validity of remote sensing of facial activity to do the same.

The human state and emotional assessment can be defined as a participants stress response to specific stimuli that includes behavioral, physiological and visual components. Behavioral measurements in stress studies are subject to modification, based on what an individual is able and motivated to display, whereas physiological measurements are more implicit, where individuals are often not aware of and furthermore cannot control their physiological responses to any great degree. Visual facial observations (facial expression, head gesture, eye movements, gaze, etc.) can provide quality indicators of stress [7]. The domain of human state assessment and emotion recognition from a psychological perspective has been expanding over several years. Ekman and Friezen [5] have shown that facial expressions may be linked to specific (i.e., basic) emotions, and that these basic emotions are relatively universal across human cultures. Stemming from his work with Tomkins, Ekman has developed a systematic method of

categorizing human facial expressions known as the Facial Action Coding System (FACS) [6]. J. Bailenson [1]et al. performed real time classifications of emotions (sadness or amusement) using facial feature tracking and physiological responses and recorded a classification rate of nearly 95%. One of the unique things is that their paper analyzes physiological features and uses the physiological features along with facial features to recognize emotions. In some cases, it has been shown that the physiological levels in detection of sadness outperform facial expression data analysis. The paper uses trained psychological coders labels for participants emotions. The ground truth reliability is heavily dependent on the inter-coder reliability to correctly labeled emotions. Most of the visible cameras data help perform facial feature point tracking that could be used to analyze emotions. This may not be enough to understand complicated emotions such as challenge/threat responses, so there is a need to explore other sensors like thermal sensors. Our research focuses on thermal imagery to detect behavioral indicators of emotional changes during a high-stakes situation which, to our knowledge, is the first analysis in this area.

Recent technological advances in imaging, computing and pattern recognition have made it possible to effectively analyze facial visual modalities. Human emotions are thought to trigger specific facial activity as external signals (although facial activity is clearly not for the sole purpose of emotional expression). We aimed to capture these external signals using non-contact sensors. Several approaches have been used to classify human affective states using facial expressions. One of the most common non-contact sensors is the electro-optics (EO) sensor that operates in the visible light (VIS) waveband, which can capture 2D static images as well as 2D dynamic video sequences. Yacoob et al. proposed a mid-level symbolic representation for spatial and temporal data using linguistic and psychological considerations [23]. The system achieved a recognition rate of 86% for smile, 94% for surprise, 92% for anger, 85% for fear, 80% for sadness, and 92% for disgust. Essa and Pentland have developed an advanced computer vision system to probabilistically characterize facial motion and muscle activation, thus developing a new and more accurate representation of human facial expressions termed FACS+ [8]. Their expression recognition accuracy is close to 98% on a database of 52 sequences. Tian et al. recognized the changes in the action units (AUs) of the FACS system using static and dynamic 2D images [21] and making use of geometric-based feature (using a set of points that represent a facial component) and appearance-based features (using texture information).

The use of thermal imaging is increasing in the surveillance, security, military and health industries. Another important advantage of thermal imaging is its lack of sensitivity to varying illumination conditions, unlike traditional VIS imagers. O'Kane et al. have demonstrated noticeable changes in the thermal signature of the human face during breathing, muscle tension, aerobic exercise, and during the playing of aggressive video games [16], suggesting that thermal imaging can be used to gain insight on the human state. These results suggest that thermal imaging is a promising technology that could be used to gain insight on the perception of human state assessment and also to understand underlying internal states. I. Pavlidis et al. [17] have captured high definition thermal images of the face that were useful for deceit detection. Exploring thermal images and detecting deceit has accuracy comparable to the polygraph examination. To attach electrodes and to perform a security screening at the airport using a polygraph mechanism for each individual person is almost impossible because of the amount of time needed. Using thermal imaging of a face gives a specific thermal signature for different emotions. In the paper [17], an experiment is conducted with twenty participants and they were asked to stab a mannequin, rob it for $20, and then prove that they are innocent. The thermal imaging was successful in correctly classifying 6 out of 8 participants who were guilty. Using thermal imagery as a stand-off sensor in turn helps to measure and analyze psychological responses without contact sensors.

Another study [14] measures the startling effect using thermal imaging. Facial thermal signatures changes have been seen near the periorbital and cheek regions for subjects after fright eliciting experiments. The study in [14] shows that thermal signatures of the face help us to determine the psychological state of a person. However, the above mentioned studies did not provide any pattern recognition analysis of thermal signatures to classify the emotions. Next, we discuss studies which provide a detailed pattern analysis of the respective thermal signatures.

Using facial temperature difference images and binary patterns from specific facial regions as inputs into an artificial neural network (ANN) for classification, Yoshitomi et al. could distinguish happy, surprised and sad expressions with a recognition accuracy of 90% [24]. Liu and Wang analyzed facial temperature sequences from samples of the USTC-NVIE (natural visible and infrared facial expression) database and computed statistical and temperature difference histogram features. Hidden Markov Models (HMM) were then used to discriminate happiness, disgust and fear with recognition rates of approximately 68%, 57%, and 52%, respectively [15]. This research suggests that thermal cameras could be an effective non-contact modality for sensing temperature changes of the face. Jarlier et al. [10] show that the thermal changes of the face are caused by the changes in the facial muscle contractions. The FACS coders are trained to produce different action unit combination at various intensities. These changes in action unit combination eventually cause the thermal patterns which can be classified using a PCA decomposition of the thermal signal. One of the things to be noted is that all the coders are forced to certain emotions; which makes it difficult to detect and characterize spontaneous expressions.

The primary objective of the current research effort was to develop an emotional state assessment system using multimodal non-contact sensors. We employed 2-D (VIS) and midwave infrared (MWIR) imagers with a goal of developing a state classifier based on specific features from the contact sensor data and various the psycho-physiological indices as ground truth. Ultimately, we aim to create a sensor model that establishes associations between human facial signatures captured and a separately validated psycho-physiological state. To validate our approach, two human subject studies were conducted: a false opinion study and a false behavior study.

Method

Participants

The false opinion study included 44 male participants, with 3 participants eliminated due to missing data for the self-report (n = 1), physiological (n = 1) , or sensor data (n = 1), yielding

a dataset from 41 participants. The false behavior study included 50 male participants, with 2 participants eliminated due to missing data, yielding a dataset of 48 participants. Given the resources to collect only a small sample size, we included males only to reduce physiological variability (autonomic and hormonal variability) in stress responses to the deception task.

Measures

Surveys

Self-assessment surveys were administered pre- and post-task and included questions that address demographics, state anxiety, positive and negative affect, and stressor appraisals. To measure self-reported appraisals, the Stressor Appraisal Scale (SAS) [18] was administered in both studies. With the SAS, multiple questions assess perceptions of the personal relevance (i.e., stressor demands) of the experimental task and the persons perceptions of his ability to manage those stressor demands (coping resources). Personal relevance is assessed via primary appraisal questions, and includes: How demanding do you expect the upcoming task to be? and How important is it for you to do well on this task? Coping resource questions include: How well will you be able to perform this task? The primary and secondary items resulted in two reliable subscales (Cronbachs alphas exceeded .70). To arrive at the appraisal ratio, an average of the personal relevance items is divided by the coping resource perceptions (demands/resources). An individuals ratio value indicates where he lies on the threat-challenge continuum, with higher values indicating greater threat and lower values representing greater challenge. To measure self-reported emotions we use the Positive and Negative Affect Scale (PANAS) which asked participants to rate how they are feeling at the present moment. There are 10 positive affect indicators, including inspired, strong, and attentive, and 10 negative affect indicators, including afraid, distressed, and nervous. The subscales for positive and negative affect were both reliable (Cronbachs alphas exceeded .70).

Physiology

Cardiovascular signals, including electrocardiography (EKG), were obtained with an ambulatory impedance cardiograph (Mindware, Inc.). Baseline impedance (Z0) and the rate of change in impedance on a given heartbeat (dZ/dt) are used to derive measures of cardiac performance. These signals along with the EKG are used to estimate stroke volume, cardiac output (CO; the amount of blood pumped out of the heart over time), and contractile force, to determine cardiac reactivity. CO is combined with mean arterial blood pressure (MAP) (CO * 80 / MAP) to estimate vascular resistance, and to determine vascular reactivity. For both studies blood pressure assessment was made on the non-dominant arm. For the first study, we used an Oscar Ambulatory BP Monitor, obtaining two assessments at baseline, one pre-task and two during the task. For Study 2, we used a continuous noninvasive arterial blood pressure monitor (CNAP). Reactivity values were obtained by subtracting the last minute of baseline (resting rate) from the first minute of each task. The key physiological variables of interest, based on our conceptualization of stress and prior robust findings in this literature, were CO (derived from EKG and stroke volume estimates) and vascular resistance (derived from CO and MAP estimates). To enhance power for analyses, the reactivity values for CO and vascular resistance were transformed into a single physiological variable where higher values denote greater challenge physiology.

Imaging

The imaging platform consists of scientific-grade, high-resolution imagers including a VIS camera (Basler A202k, Basler, Inc., Exton, PA) and a MWIR imager (SC6700, FLIR Systems, Boston, MA). The cameras are placed at a distance of approximately two meters from the participant during data capture. Image data are acquired at 30 Hz, with the VIS and MWIR cameras operating synchronously. The MWIR camera is calibrated with a blackbody source and the ambient temperature and humidity are automatically sampled and logged throughout the experimental period.

Procedure

For both studies, the participants were seated in a comfortable chair and seven physiological sensors were attached to their torso to assess heart rate and blood flow (Figure 1). Baseline physiological measures were recorded for five minutes, whereas MWIR and VIS image data were collected during the last 10 seconds of baseline.

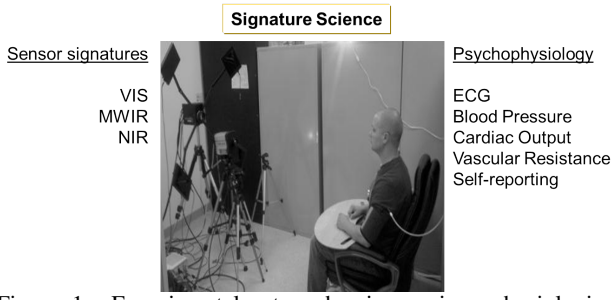

Figure 1: Experimental setup showing various physiological probes attached to the participant and three cameras that are positioned approximately two meters from the participant.

Study 1

For Study 1, after the physiological baseline, participants were administered a baseline survey that assessed state emotion, demographics, and their opinions about abortion and gay marriage. The researcher determined the participants strongest held opinion, and participants were randomized to speak either towards or against their own opinion with approximately (n=22 participants) in each category. Participants were told that if deemed by the researchers as telling the truth, they would receive a financial bonus, but if deemed as lying they would relinquish half of their original participant pay (high stakes scenario). After these instructions, appraisals and emotions were assessed. Participants were given one minute to prepare the speech and two minutes to deliver their speech. After the preparation time, participants were reminded to speak either to or against their opinion and the speech task commenced. When the task was complete, we again measured appraisals and emotion, and their beliefs about how effective they were in persuading the researchers. All participants were told that the researchers were mixed in discerning their truthfulness, so each received the participant pay plus bonus. Participants were asked to not share study details with others, and were told that other participants might not be judged similarly and

could lose their participation money. For the analysis of this paper, we examined the data from 20 participants who were in the false opinion condition (i.e., spoke against their opinion).

Study 2

Study 2 was similar in many ways to Study 1, in that the measures for physiology, psychological self-report and the multimodal imaging remained the same. However, the nature of the deception differed. In describing the task, the experimenter gave the participant an envelope that either did or did not contain money. The participant was to take the money and hide it, or pretend (rattle the envelope) to take the money from an empty envelope (the presence of money evoked the truth and deception condition, and was randomized). Researchers were blind to condition. Again, the task was to deliver a speech and be deemed as telling the truth that they did not take money. As before, if deemed truthful by the researchers, participants would receive a bonus, otherwise half of the participant remuneration was to be forfeited. After these instructions, self-reported appraisals and emotions were assessed. As in Study 1, this included a one-minute preparation and two-minute speech task. After the preparation, participants were reminded to convince the judges that they did not take the money and then the speech task began. After the task, self-report appraisals and emotions were measured. Participants were delivered the same debriefing and asked not to share the study details with others as in Study 1, and all were given their remuneration plus bonus. The analysis for this study included 23 participants who spoke about not taking the money when they actually did take it.

Image acquisition

VIS and MWIR image data are acquired synchronously during the last 10 seconds of the baseline period and during the entirety of each task. The VIS and MWIR images were co-registered using test pattern images (that are collected pre-baseline) as inputs to our mutual information based registration algorithm [11]. The VIS and MWIR images are co-registered using test pattern images as inputs to a mutual information based registration algorithm that we tailored for thermal imaging based on coarse and fine region registration. Since the object (i.e. human face) depth is not large, an assumption of affine geometry is considered and thus, requires parameters for translation, rotation and scale differences. Fifteen fiducial points are selected for each of the test pattern images. Note that one-to-one correspondence is not needed during the selection process as the algorithm will examine all combinations of available points to find the set of analogous pairs. Six transform coefficients are returned, and then each MWIR is transformed to match the analogous VIS image using bilinear interpolation.

Extraction of thermal features

In summary, facial feature tracking was performed on the VIS images using a real-time face tracker (Visage T-Tracker, Visage Technologies AB, Linkping, Sweden). Following manual initialization of a 3D mesh mask that is customized to each participants facial geometry, the software tracks the face in each frame at video rates. The tracking procedure is explained in [4] where 49 fiducial points are selected out of the total 84 points returned by the face tracker. In addition to the set of fiducial points returned by the software, we derive points that enable facial segmentation into 29 non-overlapping segments (Figure 2). The segment vertices are then transferred to the co-registered MWIR images. Approximately 90% of our participant videos were tracked successfully. Predictably, tracking is lost when head pitch or yaw is too great or when participants obscure part of their face with their hands. Tracking generally resumes within a few frames once proper position is resumed or the obscuration is removed. From each segment of each frame of the MWIR im-agery, we extracted a number of features, including several statistical features (mean, minimum, maximum, and standard deviation of pixel intensities) as well as the mean of the top 10% thermally hottest pixels in a segment. The mean of top 10% thermally hot pixels in a segment is motivated from the method described in the research article [22] where the mean temperature of the 10% hottest pixels from within the periorbital region of interest is used to classify deceit. Using this feature, a classification rate of 87.2% is achieved for 39 subjects, which is almost on par with success rate achieved by highly trained psycho-physiological experts. The facial image region is divided into local areas and features are extracted from each region independently and then these features are concatenated to form a global descriptor of the face [9]. There are several ways that a facial image region can be divided into rectangular regions 4 x 4, 5 x 5, 7 x 7, etc. We have experimented with several grids and experimentally found that 4 x 4 grid performs better with a histogram-based feature. An implementation of a histogram of thermal feature is implemented, wherein each thermal image is segmented into a regular 4 x 4 grid, the histogram of pixel intensities in each grid segment is calculated and the cumulative distribution function for these data is formed. The segmentation of the face resulted in 16 separate regions with each region containing a specific part of the face. A histogram of intensity values are computed independently within each of the 16 regions. The resulting 16 histograms are combined yielding a (165 bins=80) dimensional feature vector.

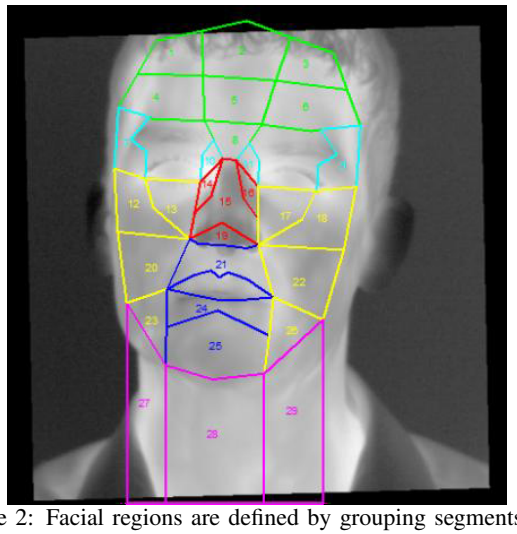

Figure 2: Facial regions are defined by grouping segments into forehead (green), eye/periorbital (cyan), nose (red), cheek (yellow), mouth (blue), and neck (magenta) regions.

Among the features computed, the most promising feature was the mean of the top 10% hottest pixels in a segment. We have used this particular feature in most of our analysis, as it can be tracked over the whole stressor time cycle from baseline through

Table 1: Summary data for participants in the lie condition for study 1 and study 2

Study 1 (n=20)			Study 2 (n=23)		
	Mean	SD		Mean	SD
Appraisal Ratio	1.10	0.32	Appraisal Ratio	1.06	0.55
Positive Effect	2.75	0.61	Positive Effect	3.25	0.78
Negative Effect	1.62	0.59	Negative Effect	1.66	0.80
Physiology	-0.11	0.80	Physiology	0.25	0.84

Table 2: Summary data for most challenged person and most threatened person for for study 1 and study 2

Most challenged person			Most Threatened person		
	Study 1	Study 2		Study 1	Study 2
Appraisal Ratio	0.75	0.24	Appraisal Ratio	1.82	2.30
Positive Effect	2.50	4.40	Positive Effect	2.30	3.50
Negative Effect	1.00	1.00	Negative Effect	2.90	4.40
Physiology	-1.02	1.58	Physiology	0.54	0.26

the first 30 seconds of the task for all 29 segments of the face. 130 seconds of data (10 seconds of baseline and 120 seconds of task data) yield 3,600 frames for analysis. Analyzing just a single feature from each facial segment results in over 100k features per subject. Data reduction was achieved by temporal sampling using various schema such as sliding windows (details will be published elsewhere) to yield both sub-sampled and time-averaged features. Features from defined time epochs, such as slopes and differences from baseline values, were included in the final feature pool of 261 features per participant.

Classification

We employed Eigen analysis on the spatio-temporal data set to facilitate dimensionality reduction and improve classification results. We combined spatial segments to form six distinct regions as shown in Figure 2, and then focused on the forehead and nose regions to reduce the feature set to 300 time series features (30 distinct frames * 10 segments) and 90 time-epoch-based features (9 epochs * 10 segments) per participant. In this particular study, statistical and histogram-based thermal features from the false behavior dataset were input into a basic k-Nearest Neighbor classifier wherein the appraisal ratio served as ground truth (GT).

Results

False opinion (Study 1) vs. false behavior (Study 2)

There appeared to be more variability in self-reports of appraisal and emotional state for Study 2 (Table 1). In Study 2, we also found that the most challenged person appeared to have higher positive affect, lower negative affect, and challenge-like physiology, compared to the most challenged person in Study 1 (Table 2). A similar examination of the most threatened persons in Study 1 versus Study 2 revealed that Study 2s most threatened person appeared to have higher threat appraisals, lower positive affect, and higher negative affect, than the most threatened person in Study 1 (Table 2). However, the physiology appeared to be similar for the most threatened person in Studies 1 and 2. Overall, the psychophysiological metrics appear to be more varied in Study 2, and in the right direction for discriminating between challenged and threatened participants. Physiology may have been less reliable for Study 1 given the non-continuous measurement of blood pressure. The physiological index for each study was not related to self-reported stressor appraisals or affective states, as it has been in past research [18],[19], [13],[2]. Consequently, subsequent analyses focused on those variables that were related to one another as expected from the theoretical and empirical literature.

Given the above examination of the data for each study, we utilized the Study 2 appraisal ratio as our ground truth for analysis. The research and findings have shown that an appraisal ratio distinguishes important outcomes for people along the threat-challenge continuum [18],[19], [13],[2]. Our preliminary facial pattern analysis for both studies indicated that Study 2 uncovered interesting and discriminatory results.

Data partitioning using appraisal ratio

In Study 2, the appraisal ratio served as ground truth for the classification task and led to the identification of three classes of participants: threatened, challenged and neutral (Figure 3).

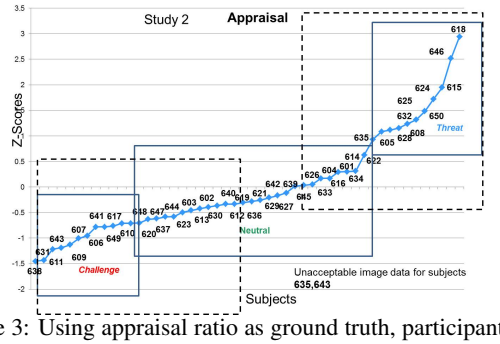

Figure 3: Using appraisal ratio as ground truth, participants were divided into three categories: threatened (z-score $>$ 0.5, n=10), challenged (z-score $<$ -0.5, n=10) and neutral (-0.5 $<$ z-score $<$ 0.5, n=28).

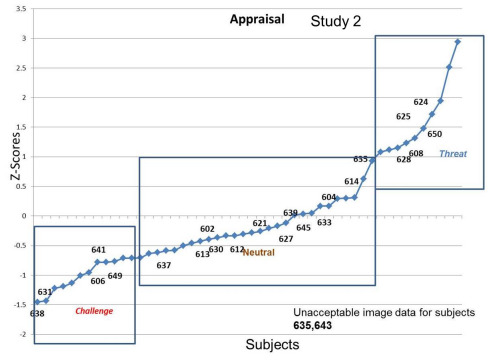

Figure 4: Using appraisal ratio as ground truth, participants who were being deceitful were divided into three categories: threatened (z-score $>$ 0.5, n=5), challenged (z-score $<$ -0.5, n=5) and neutral (-0.5 $<$ z-score $<$ 0.5, n=13).

This analysis included all Study 2 participants with usable data and not just the subset in the deception condition. The computed z-scores ranged from -2.0 to 3.0, where higher Z-scores indicate threat and lower Z-scores indicate challenge conditions.

Partitioning the Z-scores at thresholds of -0.5 and 0.5 results in 10 subjects in the challenged class (Z-score $<$ - 0.5), 10 subjects in the threatened class (Z-score $>$ 0.5) and 28 subjects in the neutral class (i.e., less challenged, less threatened).

To further analyze only those subjects who are being deceitful, we re-partition the data in a similar way as above, but consider only participants who are in the deception condition (i.e., those who took the money and convinced the judges that they did not). Of the 48 Study 2 participants, 23 meet this criterion (Table 1); these subjects are identified in Figure 4. Five participants fall into the threatened category, five participants fall into the challenged category and the remaining participants fall into the neutral category.

State classification using thermal features

Analyses using the whole face yielded 870 time-series features and 261 epoch-based features per participant. Following these initial analyses, we noted that features from the forehead and nose regions were most successful in class discrimination, and we subsequently restricted our focus to Segments 1-6, 14-16 and 19. This reduced the feature set to 300 time-series features and 90 epoch-based features per participant. Eigen analysis revealed that the first principal component (PC1) captures the lowest frequency information, whereas the second and third principal components (PC2 and PC3) capture high frequency information. Further investigation showed that the facial changes in response to the stressors are reliably captured by PC2 and PC3. Plots of PC2 and PC3 for the time-series and epoch-based features (Figure 5 top and bottom, respectively) show that the separation between challenged (blue) and threatened (red) participants is similar. In fact, the relative positions of participants within the distributions remain fairly consistent, indicating that the time-series and epoch-based analyses support the same conclusions. Most importantly, this analysis validates our strategy to reduce 300 time-series features and 90 epoch-based features to two features (each). Using

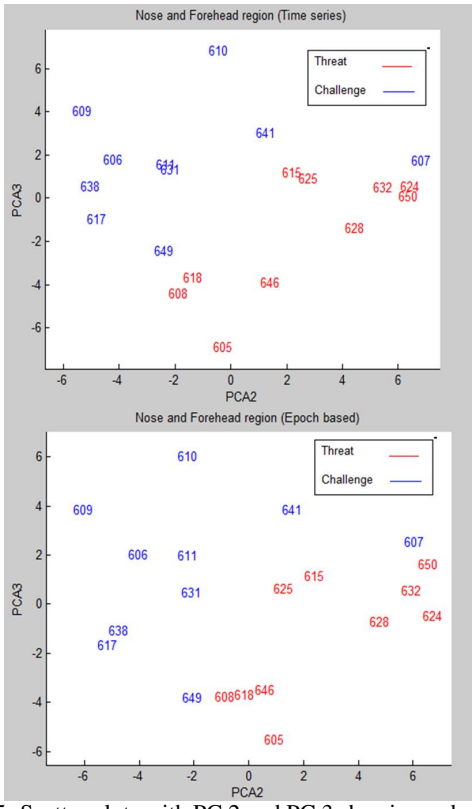

Figure 5: Scatter plots with PC 2 and PC 3 showing reduced time series features in the top plot and reduced epoch features in the bottom plot.

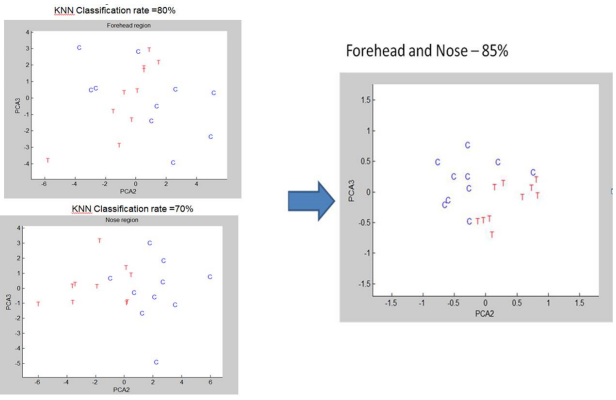

Figure 6: Scatter plots of PC3 versus PC2 for the forehead (upper left), the nose (upper right), and the fusion of the forehead and nose regions (bottom), using data from the threatened (T) and challenged (C) participants in Study 2 Classification accuracies using k-NN for the forehead, the nose and the combined forehead and nose were 70%, 80%, and 85%, respectively.

the k-NN classifier with k=3, we found that features from certain regions of the face are more effective than others for our classification problem (Table 3). Specifically, the forehead and nose regions were most useful for the discrimination of threatened and challenged participants, a result that is consistent with other research findings [17]. Combining the data from these segments yielded a classification accuracy of 85% (Figure 6). This is in contrast to an accuracy of 73% observed when inputting features from the entire face.

Table 3: Region based classification accuracies calculated using k-NN classifier

Region	Segments	Classification accuracy
Forehead	1, 2, 3, 4, 5, 6	80%
Nose	14, 15, 16, 19	70%
Eye	7, 8, 9, 10, 11	50%
Cheek	12, 13, 20, 23	55%
Mouth	21, 24, 25	35%
Neck	27, 28, 29	30%
Face (Region fusion)	[1....29]	73%
Forehead + Nose (region selection)	1, 2, 3, 4, 5, 6, 14, 15, 16, 19	85%

Pattern analysis for test data set with 30 participants for Study 2

The majority of our pattern analysis discussed involved using the extreme threatened and challenge participants. In this analysis, some new participants are included that belong to the category of lesser threatened and lesser challenged class (neutral) as shown in the Figure 7. The dotted box in the challenge category includes the original extreme challenge participants and also five new participants that belonged to the lesser challenged class. Similarly, the dotted box in the threat category includes the original extreme threat participants and also five new participants that belonged to the lesser threatened class.

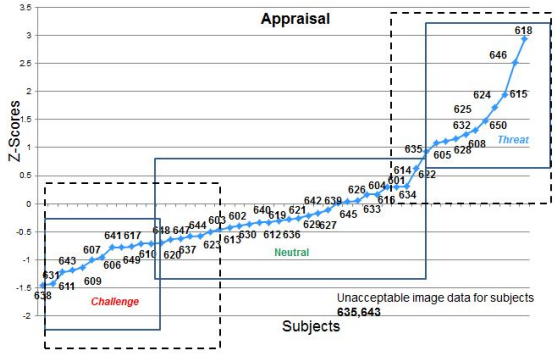

Figure 7: Data partitioning indicating additional (five lesser threat + five lesser challenge) test participants shown in the dotted rectangular box.

Accuracy All PCs	Accuracy PC 1, 2, 3	Accuracy PC 2, 3
53.33%	56.67%	70%

	T	C
T	0.67	0.53
C	0.27	0.4
N	0.07	0.07

	T	C
T	0.6	0.33
C	0.27	0.53
N	0.13	0.14

	T	C
T	0.8	0.2
C	0	0.6
N	0.2	0.2

Figure 8: Accuracy estimation along with confusion matrices for the fusion of nose and forehead for lesser threat and lesser challenge participants

A total of 30 participants are used for the analysis. In the threat class, 10 participants belonged to extreme threat class and remaining five belonged to lesser threat class. In the challenge class, 10 participants belonged to extreme challenge class and remaining five belonged to lesser challenge class. The fused forehead and nose region provides better accuracy as shown in Table 4. Henceforth, fused forehead and nose regions epoch data are used for most of the analysis. The statistical feature (mean of top 10% hot pixels) is estimated for all the 8 segments of the forehead and nose and for the first two epochs (9 seconds) of the particular emotion. The feature vector would be 8 sub-regions x 2 epochs x 1 feature=16 values row feature vector for each participant. The cross validation involves a previously trained system with 20 extreme threat and challenge participants and testing involves using all the 30 participants.

Finally, accuracy is calculated for three different cases using k-NN classification as shown in Figure 8. In the first case accuracy is estimated using all the principal components (PCs), while in the second case first 3 PCs are used for estimating accuracy. Finally, the third case involves using the second and third PC to estimate the accuracy. For each case a confusion matrix is calculated where the T indicates threat value, C indicates challenge value and N indicates that the predicted value does not belong to T or C and therefore, it belongs to neutral category or less threatened and less challenged category. The accuracy calculated for each case indicates that PC 2 and PC 3 provide a better separation for the lesser threatened and challenged data indicating that the thermal changes on the face are mostly higher order frequency changes. However, the accuracy dropped from 85% to 70% indicating that the lesser threat and challenge participants probably do not exhibit a characteristic threat or challenge response.

Pattern analysis for test data set with 40 participants for Study 2

The majority of our pattern analysis discussed involved using the extreme threatened and challenge participants. In this analysis, additional new participants are included that belong to the category of lesser threatened and lesser challenged class (neutral) as shown in the Figure 9. The dotted box in the challenge category includes the original extreme challenge participants and also ten new participants that belonged to the lesser challenged class. Similarly, the dotted box in the threat category includes the original extreme threat participants and also ten new participants that belonged to the lesser threatened class.

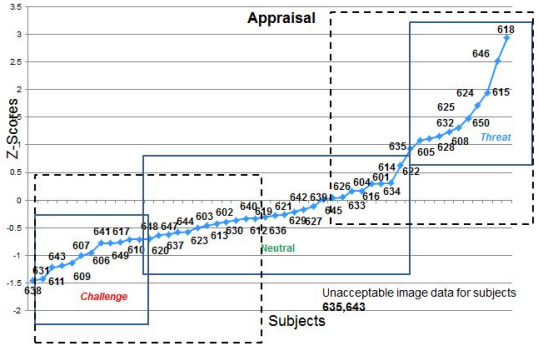

Figure 9: Data partitioning indicating additional (ten lesser threat + ten lesser challenge) test participants shown in the dotted rectangular box.

A total of 40 participants are used for the analysis. In the threat class, 10 participants belonged to extreme threat class and remaining ten belonged to lesser threat class. In the challenge class, ten participants belonged to extreme challenge class and remaining ten belonged to lesser challenge class. From the Table 4, it is evident that the fused forehead and nose region provides better accuracy. Therefore, fused forehead and nose regions are used for most of the analysis. The statistical feature (mean of top 10% hot pixels) is estimated for all the 8 segments of the forehead and nose and for the first two epochs (9 seconds) of the particular emotion. The feature vector would be 8 sub-regions x 2 epochs x

1 feature=16 values row feature vector for each participant.

Accuracy All PCs	Accuracy PC 1, 2, 3	Accuracy PC 2, 3
50 %	52.50%	52.50%

	T	C
T	0.5	0.35
C	0.4	0.5
N	0.1	0.15

	T	C
T	0.55	0.35
C	0.25	0.5
N	0.2	0.15

	T	C
T	0.55	0.4
C	0.3	0.5
N	0.15	0.1

Figure 10: Accuracy estimation along with confusion matrices for the fusion of nose and forehead for lesser threat and lesser challenge participants.

The cross validation involves a previously trained system with 20 extreme threat and challenge participants and testing involves using all the 40 participants. Finally, accuracy is calculated for three different cases using k-NN classification as shown in Figure 10. The accuracy calculated for each case indicates that PC 1, PC 2 and PC 3 together provide similar classification as PC 2 and PC 3. However, the accuracy dropped from 85% to 52.50% indicating that the lesser threat and challenge participants probably do not exhibit a characteristic threat or challenge response.

Deception classification using thermal features

We have reported results from Study 2 for the challenged versus threatened classes (Figure 3), which are comprised of both deceptive and non-deceptive participants. A similar analysis was performed using only the deceptive Study 2 participants, who also populate the challenged-threatened continuum (Figure 4).

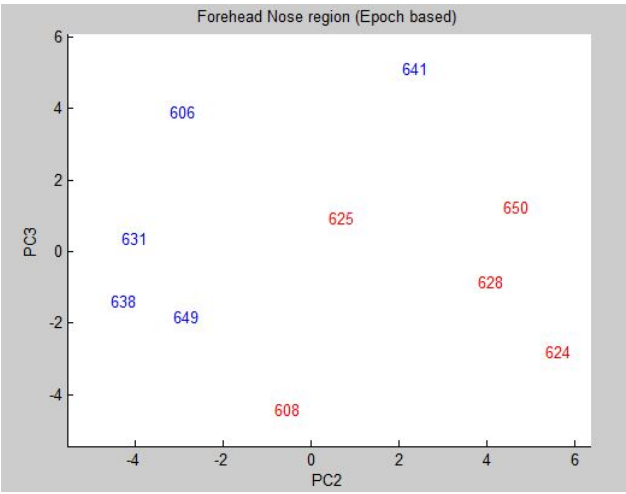

Figure 11: Scatter plots of the PC2 and PC3 features for Study 2 participants who were in the deceptive category. A clear separation exists between the Challenged (blue) versus Threatened (red) subjects in this (albeit small) dataset.

Eigen analysis of the deceptive participants showed good separation between the challenged versus threatened classes (Figure 11), leading to a k-NN classification accuracy of 90%. The k-NN analysis was performed using a leave one out cross validation strategy. However, given the small sample size (n=23), these results may not be generalizable. However, our classification findings suggest that both the deceptive group and the mixed group (deceptive and non-deceptive) participants can be readily classified into challenged and threatened classes using thermal features extracted from the face.

Discussion

The goal of this study was to examine whether thermal imaging is useful for detecting differences between specific human states, which we identify as threatened and challenged states. We applied Eigen analysis and a simple k-NN classifier to a thermal feature dataset to successfully discriminate emotional states. Our approach was tested with a minimum of ten threatened and challenged participants whose spontaneous changes in thermal signatures were captured and validated with psychological self-reporting ground truth information. From the above sections, the accuracy rate drops considerably from 85% to 70% and further drops to 52.5% as the lesser threatened and lesser challenged participants are included in the testing dataset as shown in Table 4. The pattern analysis agrees with the psychological ground truth where the participants close to zero z-score are the typical neutral participants.

Table 4: Accuracy estimation for all experiments calculated using k-NN clasifier using leave one out cross validation technique

Threat and challenge participants	Accuracy all PCs	Accuracy PC 1, 2, 3	Accuracy PC 2, 3
20 participants	60%	60%	85%
30 participants	53.33%	56.67%	70%
40 participants	50%	52.50%	52.50%

The results suggest that our method could be used reliably for state discrimination. Further, the methodology allows insight into other questions about the duration and pattern of the thermal response to evoked emotion. For example, we noted that after the first few time epochs, additional data did not improve classification accuracy. We also observed spatial patterning that suggests that the forehead (frontalis) and nose (nasalis) regions provide discriminatory information for classification of threatened versus challenged individuals. Temperatures in the nasalis region increased for threatened individuals and decreased in challenged individuals. Temperature changes were also observed in other regions of the face, but these features did not as effectively discriminate between the two states. Lastly, by comparing classification accuracies of individual and combined spatial regions, we found that pooling the data from the forehead (80%) and nose (70%) regions led to improved performance (85%). In summary, thermal imaging appears to capture meaningful temperature changes in the face that are useful for classifying distinct human emotional states.

Conclusion

We used thermal imaging to discriminate threatened from challenged participants during a deception (high stakes) stressor. Participants were not uniform in their evaluations of the stressor; rather, appraisals of engaging in deception ranged from threat to challenge. Among various multi-modal sensor outputs, data captured from the MWIR camera provided statistical features that when transformed, appeared to appropriately distinguish threatened and challenged responses in individuals. Using data from specific facial regions and over defined temporal ranges, we were able to classify with an average recognition rate of 85% or greater.

Acknowledgments

The authors wish to thank Meena El-Shaer, Gary Kash, James Trame, Jeremy Lewis, Natalie Gilbert, Adrian Johnson and Melissa Peterson for their efforts in data collection and image truthing. The authors also wish to thank Dr. Erik Blasch for his valuable suggestions throughout the research. This work was supported by a grant from the Office of Naval Research (Grant N00014-10-1-0295).

References

[1] Jeremy N Bailenson, Emmanuel D Pontikakis, Iris B Mauss, James J Gross, Maria E Jabon, Cendri AC Hutcherson, Clifford Nass, and Oliver John. Real-time classification of evoked emotions using facial feature tracking and physiological responses. *International journal of human-computer studies*, 66(5):303–317, 2008.

[2] J Blascovich and WB Mendes. Challenge and threat appraisals: The role of affective cues. inj. forgas (ed.), feeling and thinking: The role ofaffect in social cognition (pp. 59–82), 2000.

[3] David B Buller and Judee K Burgoon. Interpersonal deception theory. *Communication theory*, 6(3):203–242, 1996.

[4] Carl B Cross, Julie A Skipper, and Douglas Petkie. Thermal imaging to detect physiological indicators of stress in humans. In *SPIE Defense, Security, and Sensing*, pages 87050I–87050I. International Society for Optics and Photonics, 2013.

[5] Paul Ekman and Wallace V Friesen. Constants across cultures in the face and emotion. *Journal of personality and social psychology*, 17(2):124, 1971.

[6] Paul Ekman, Wallace V Friesen, and Silvan S Tomkins. Facial affect scoring technique: A first validity study. *Semiotica*, 3(1):37–58, 1971.

[7] Paul Ekman and Erika L Rosenberg. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[8] Irfan A Essa and Alex P Pentland. Facial expression recognition using a dynamic model and motion energy. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 360–367. IEEE, 1995.

[9] Ms SS Ghatge and VV Dixit. Face recognition under varying illumination with local binary pattern. *International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering*, 2(2), 2013.

[10] Sophie Jarlier, Didier Grandjean, Sylvain Delplanque, Karim N'diaye, Isabelle Cayeux, Maria Inés Velazco, David Sander, Patrik Vuilleumier, and Klaus R Scherer. Thermal analysis of facial muscles contractions. *IEEE Transactions on Affective Computing*, 2(1):2–9, 2011.

[11] Hrishikesh Karvir. *Design and validation of a sensor integration and feature fusion test-bed for image-based pattern recognition applications*. PhD thesis, Wright State University, 2010.

[12] Robert M Kelsey, Jim Blascovich, Christopher L Leitten, Tamera R Schneider, Joe Tomaka, and Stefan Wiens. Cardiovascular reactivity and adaptation to recurrent psychological stress: The moderating effects of evaluative observation. *Psychophysiology*, 37(6):748–756, 2000.

[13] Robert M Kelsey, Jim Blascovich, Joe Tomaka, Christopher L Leitten, Tamera R Schneider, and Stefan Wiens. Cardiovascular reactivity and adaptation to recurrent psychological stress: Effects of prior task exposure. *Psychophysiology*, 36(6):818–831, 1999.

[14] James A Levine, Ioannis Pavlidis, and Murray Cooper. The face of fear. *The Lancet*, 357(9270):1757, 2001.

[15] Zhilei Liu and Shangfei Wang. Emotion recognition using hidden markov models from facial temperature sequence. In *Affective computing and intelligent interaction*, pages 240–247. Springer, 2011.

[16] Barbara L O ' Kane, Philip Sandick, Todd Shaw, and Mike Cook. Dynamics of human thermal signatures. In *Proceedings of the Inframation Conference. Las Vegas*, 2004.

[17] Ioannis Pavlidis, Norman L Eberhardt, and James A Levine. Human behaviour: Seeing through the face of deception. *Nature*, 415(6867):35, 2002.

[18] Tamera R Schneider. The role of neuroticism on psychological and physiological stress responses. *Journal of Experimental Social Psychology*, 40(6):795–804, 2004.

[19] Tamera R Schneider. Evaluations of stressful transactions: what's in an appraisal? *Stress and Health*, 24(2):151–158, 2008.

[20] Kamila E Sip, David Carmel, Jennifer L Marchant, Jian Li, Predrag Petrovic, Andreas Roepstorff, William B McGregor, and Christopher D Frith. When pinocchios nose does not grow: belief regarding lie-detectability modulates production of deception. 2012.

[21] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[22] Panagiotis Tsiamyrtzis, Jonathan Dowdall, Dvijesh Shastri, Ioannis T Pavlidis, MG Frank, and P Ekman. Imaging facial physiology for the detection of deceit. *International Journal of Computer Vision*, 71(2):197–214, 2007.

[23] Yaser Yacoob and Larry Davis. Computing spatio-temporal representations of human faces. In *IEEE Computer Society Conference On Computer Vision And Pattern Recognition*, pages 70–70, 1993.

[24] Y Yoshitomi, N Miyawaki, S Tomita, and S Kimura. Facial expression recognition using thermal image processing and neural network. In *Robot and Human Communication, 1997. RO-MAN'97. Proceedings., 6th IEEE International Workshop on*, pages 380–385. IEEE, 1997.

Author Biography

Nilesh U. Powar received his Bachelor of Engineering in Electronics from University of Bombay, India in 1999, and a M.S. in Computer Engineering from Wright State University, Dayton, OH in 2002 and a PhD degree in engineering from University of Dayton, Dayton, OH in 2013. He is currently a senior researcher at University of Dayton Research Institute (UDRI) and had joined UDRI in 2003. His research interests include image processing, statistical pattern recognition and system integration.

Tamera R. Schneider is currently Social Psychology Program Director, in the Social, Behavioral, and Economic Sciences Directorate, at the U.S. National Science Foundation. She earned a Ph.D. in Social/Health Psychology at the Stony Brook University in 1997, was a postdoctoral fellow and associate research scientist at Yale University (1997-2000), and joined Wright State University where she is currently Professor of Psychology. Dr. Schneider has expertise in the area of emotions and psychophysiological stress resilience, and persuasion to promote health behaviors and collaborative behaviors toward women in STEM.

Julie A. Skipper received her B.S. in Biomedical Engineering in 1991 and her Ph.D. in Biomedical Sciences (Medical Physics and Engineering track) in 2003, both from Wright State University. She is currently an associate research professor at Wright State University, teaching courses in radiation physics, engineering biophysics and medical image processing. Her research focuses on imaging systems and image analysis for medical and military applications, including multimodal sensing and information fusion.

Douglas T. Petkie received the BS degree in Physics from Carnegie Mellon University, Pittsburgh, PA, in 1990 and a PhD degree in physics from Ohio State University, Columbus, OH, in 1996. He is currently the Department Head and Professor of Physics at Worcester Polytechnic Institute in Worcester, MA. His research interests include the development of millimeter-wave and terahertz technologies for spectroscopy, imaging and radar applications. His current research projects include the development of continuous wave radar systems to measure biometric signatures, a compact and highly sensitive MEMS-based photo acoustic spectrometer for gas phase spectroscopy, and NDE techniques for health monitoring of composite materials.

Vijayan Asari is a Professor in Electrical and Computer Engineering and Ohio Research Scholars Endowed Chair in Wide Area Surveillance at the University of Dayton, Dayton, Ohio, USA. He is the director of the Center of Excellence for Computer Vision and Wide Area Surveillance Research (UD Vision Lab) at UD. Dr. Asari received his Bachelors degree in electronics and communication engineering from the University of Kerala, India in 1978, M Tech and PhD degrees in Electrical Engineering from the Indian Institute of Technology, Madras in 1984 and 1994 respectively.

Rebecca Riffle received a B.S./Health Science and B.A./Psychology from Cleveland State University (2008) and a M.S./Human Factors Psychology from Wright State University (2011). She conducted research on psychophysiological stress responses and human performance while at WSU and with NAMRU-D, respectively. She is currently a Clinical Research Coordinator (medical device studies) at MetroHealth Medical Center, Cleveland, OH, and the Technical Writer/Communications Specialist for the APT Center of Research and the CVAMREF, both at the Cleveland VA.

Matthew S. Sherwood received the BS and MS degrees in biomedical engineering from Wright State University in 2011 and 2013. He is currently pursuing a Doctorate in Engineering at Wright State University where he is examining neural correlates associated with neural augmentation via real-time functional magnetic resonance imaging. Mr. Sherwood also holds a position as a research engineer at Wright State University, Dayton, Ohio.

Carl B. Cross received his B.S. degree in biomedical engineering in 2011 and M.S. degree in biomedical engineering in 2013 from Wright State University, Dayton, OH. Carl was a research assistant at Wright State University where his research interests included stand-off physiological sensing, stress detection, thermal imaging, and computer vision. He currently works as a Biomedical Engineer at the Loma Linda VA Medical Center in Loma Linda, CA.