Physiological Measurement for Emotion Recognition in Virtual Reality

Lee B. Hinkle
Department of Computer Science
Texas State University
San Marcos, Texas, 78666
leebhinkle@txstate.edu

Kamrad Khoshhal Roudposhti
Department of Computer Engineering
Lahijan Branch, Islamic Azad University
Lahijan, Iran
kamrad@liau.ac.ir

Vangelis Metsis
Department of Computer Science
Texas State University
San Marcos, Texas, 78666
vmetsis@txstate.edu

***Abstract*—In this work, various non-invasive sensors are used to collect physiological data during subject interaction with virtual reality environments. The collected data are used to recognize the subjects' emotional response to stimuli. The shortcomings and challenges faced during the data collection and labeling process are discussed, and solutions are proposed. A machine learning approach is adopted for emotion classification. Our experiments show that feature extraction is a crucial step in the classification process. A collection of general purpose features that can be extracted from a variety of physiological biosignals is proposed. Our experimental results show that the proposed feature set achieves better emotion classification accuracy compared to traditional domain-specific features used in previous studies.**

***Index Terms*—Physiological measurement, emotion recognition, virtual reality, feature extraction, feature selection.**

I. Introduction

With the continued development of automated systems that interact closely with humans in a more natural manner, the ability of such systems to determine the emotional state of nearby subjects and adapt accordingly is increasingly important. Applications include the broad area of affective computing as well as more specific areas, such as evaluating the effectiveness of virtual reality based treatment for social phobias. The primary goal of this research is to determine emotional state of human test subjects through the non-invasive collection and analysis of physiological data during subject interaction virtual reality or other immersive audio/visual environments. This paper covers two elements of the research: the key findings related to the experimental design, and the results of the data analysis and emotion recognition.

Early work regarding emotional states includes that by Ekman and Friesen on universal facial behaviors. This study used six emotional states: happiness, anger, sadness, disgust, surprise, and fear [1]. The interactions of the sympathetic and parasympathetic nervous system cause physiological changes that are measurable [2], [3]. Work in this area includes emotion classification based on physiological bio-signals, e.g. [4]. Image processing and machine learning techniques have also been used to classify emotion based on facial images and voice recordings [5]. More recent studies suggest that emotion should not be classified into discrete states, and propose a circumplex model of affect, more widely known as the arousal-valence model [6]. We adopt that model in this study.

A key pattern that emerges is most studies using physiological biosignals for emotion recognition is that a specific set of biosignals types are used, and a limited set of domain-based features are extracted, based on the specific type of application at hand, to maximize accuracy. We hypothesize that with the blooming of sensors and machine learning, and the generation of ever larger amounts of data, a more general approach for data acquisition and feature extraction may be needed. Our experimental results support our hypothesis.

One of the key challenges in this area is the acquisition of accurate measurements. In medical settings, such as sleep studies or brain scans, the subject is typically still and has been connected to the sensors by a team of trained personnel. For the purposes of therapy evaluation and ultimately real-world data collection the subject will be moving, perhaps extensively. In addition the ease of connecting and unconnecting the sensors will need to be simplified. There are a number of aspects that need to be considered. Some are inherent in the biology of the subjects; these include the fact that the signals themselves are very small and the body generates a variety of different electrical potentials in response to brain activity, muscle activity, and involuntary reactions to stimuli [7].

On the sensor side, there is a constant struggle between the need for very good electrical conductivity to the skin and the non-invasive least intrusive goals for a system which would ideally be used continuously. For example, highly accurate measurement of muscle activity typically entails the use of a needle electrode in the muscle itself [8]. For the best non-invasive signal on the skin surface, skin preparation such as an alcohol scrub to remove dead cells and a strong adhesive electrode with pre-applied electrolyte gel is typically used. If the location has hair it may be shaved beforehand to facilitate a better connection and lessen the difficulty in removal of the electrode [9].

Given these hassles, immediate widespread adoption is unlikely. However, it's easy to foresee that electrodes will be incorporated into clothing, jewelry, and other worn accessories. Currently the signals available through these types

This research was conducted while K. K. Roudposhti was a postdoctoral research associate at the Intelligent Multimodal Computing and Sensing (IMICS) Laboratory at Texas State University.

Corresponding author: Vangelis Metsis (vmetsis@txstate.edu).

of sensors are far weaker and noisier than securely attached disposable electrodes [10]. However, technological advances will probably help overcome these problems. Furthermore, the current hardware using attached electrodes can still be utilized in special settings, such as therapy sessions.

In this work we attempt to answer the following research questions:

- Evaluate the extent to which emotional responses caused by interaction with virtual reality environments can be measured through physiological biosignals.
- Establish a methodology and discuss the challenges of biosignal data acquisition using current hardware capabilities.
- Develop a general-purpose feature extraction strategy, which is independent of the types of signals acquired, and evaluate the discriminatory power of such features compared to traditional domain-based features.

In the following sections, we first explain our data collection methodology in section II. Next, in section III, we discuss our data preprocessing steps. Subsequently, we introduce our feature extraction and feature selection approaches and present the experiment results for emotion classification in section IV. Finally, we discuss our experimental findings in section V, and conclude this paper in section VI.

II. Data Collection Methodology

The experimental methodology includes two distinct elements that are detailed in the following sections. The first is the collection methodology of the physiological signals including challenges induced by the fact that the subjects were moving during the data collection. The second is the stimuli based on the virtual reality sessions along with the annotation of the subjects' emotional state to provide label information. Five male subjects participated in this study, with ages ranging for 20 to 50 years old.

A. Physiological Data Collection

Data collection was limited to non-invasive techniques, which include disposable adhesive electrodes on the skin, wearable type sensors, such as respiration straps, finger electrodes, etc. Audiovisual data were also collected to facilitate off-line data annotation. The physiological signals were measured using two BioRadio Wireless Physiology Monitors and associated peripherals from Great Lakes Neurotechnology [11]. The BioRadio provides 4 differential inputs that can be configured for a variety of electrical biosignals and an additional expansion pod that can be used for temperature sensing and pulse oximetry. As described in the BioRadio User's Guide: "The BioRadio is worn by the person and is designed for acquiring physiological signals from sensors attached on the body. Physiological signals are amplified, sampled, and digitized, which can be wirelessly transmitted to a computer Bluetooth receiver and/or recorded to onboard memory for post-analysis." More details can be found in the user manual which is available online [11].

TABLE I
Physiological Signals Measured.

Physiological Signal	***Location***
EEG f4	High right forehead
Electrooculography (EOG) Horizontal	Outside of eyes
Electrooculography (EOG) Vertical	Above and below right eye
Electromyography (EMG) Zygomaticus "smile" muscle	Right cheek
Electrodermal Activity (EDA)	Right index & pointer finger
Electrocardiogram (ECG)	Left and right wrists
Chest Respiration (RIP)	Chest strap
Abdomen Respiration (RIP)	Stomach strap
Peripheral Temperature	Right pinkie finger
Heart Rate via PulseOx	Right ring finger
Blood Volume (PPG) via PulseOx	Right ring finger
Blood Oxygen (SpO2) via PulseOx	Right ring finger
Head Acceleration and Rotation	Rear of head
Body Acceleration and Rotation	Right hip

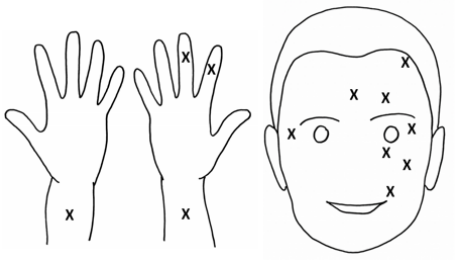

Fig. 1. Electrode placement.

Respiration was measured using inductive interface cables and respiratory inductance plethysmography (RIP) belts made by SleepSense (S.L.P. Inc). The RIP belts and interface module provide a voltage input to the BioRadio which varies based on the measured chest and abdomen volume. This technique has been shown to reliably measure respiration [12]. The finger based pulse oximeter used was Model 3012LP by Nonin Technologies. Pulse Oximetry utilizes two types of LEDs to measure the absorption of light within the finger or earlobe to determine pulse, SpO2, and peripheral blood volume [13]. A total of 24 signals were measured, including 6 head and 6 body acceleration sensors (linear and angular acceleration in 3 axis). The signals are listed in Table I.

In the following paragraphs we elaborate on the details of applying the sensors to the test subjects in order to, first, support the validity of our approach, and second, to facilitate possible future replication of similar experiments by other researchers.

1) Skin Electrode Placement: Each subject was provided with an instruction sheet guiding them in the skin preparation and electrode application. They were asked to clean the skin where the electrode will be attached with an alcohol swab to remove any oils, lotions, makeup, etc. as well as dead skin. The electrodes used in this study were MVAP-II Electrodes containing a Silver/Silver Chloride Sensing Element with Hydro Gel and manufactured by MVAP Medical Supplies. Placement locations of the adhesive electrodes are shown in Fig. 1.

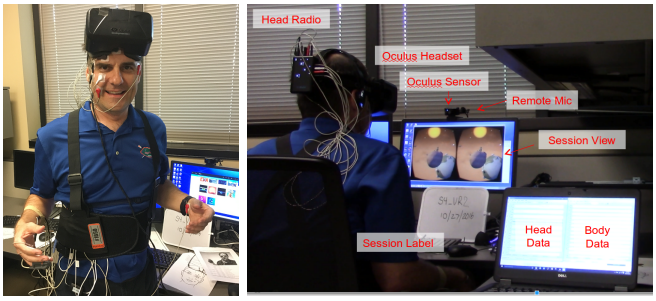

Fig. 2. Left: a picture of a test subject right before starting the VR session. Right: a picture of the test subject during a VR session.

2) Equipment Mounting and Cabling: A seemingly simple but significant issue was mounting the BioRadio uniformly and securely. The BioRadio is equipped with a removable belt clip. However, since the accelerometer and gyroscope are internal to the device, a uniform mounting independent of clothing was desired. The characteristics of the device attached to an elastic waist band might vary significantly versus one clipped to a tighter belt.

For the head BioRadio, the first attempts were to simply clip the device onto the straps of the Oculus DK2. The Oculus itself is fairly immobile on the head. In this setup, however, the BioRadio moved significantly. A big improvement in mounting was made by using a plastic mounting bracket modified from an inexpensive LED headlamp which was secured to the straps at the cross point directly in the back of the head.

In order to eliminate the effect of clothing on the motion capture of the "body" radio, a back supporting belt was utilized. This belt attaches securely around the waist and the Velcro closures were used to hold the body radio tightly against the right side of the waist. Additionally, it was possible to also use the Velcro flaps to secure the cables and sensor pods associated with the RIP belts.

Cable management remained somewhat cumbersome, and this was a problem for several reasons. First the subjects' movement is somewhat limited. This did not prove to be an issue for these activities that consisted only of sitting and standing in a limited area, but it will be a greater problem as movement is increased. Directly related to this is the fact that the cables can become snagged and disconnect during the activities. This required special attention especially with respect to the ergonomic armrests and adjustment knobs on the chair. Finally cable movement can induce noise into the signals. Best practices include taping or affixing the cables tightly to the body. This however, was not practical in a non-medical setting. Future setups would benefit from a fixed harness where the cables are joined into less cumbersome bundles.

Fig. 2 shows a two pictures of a test subject wearing the necessary gear for data collection and interaction with VR.

3) Video Collection and Event Markers: All data collection experiments were recorded with a high definition video recorder and external microphone. External speakers in parallel with the subject's headphone audio were used to record the virtual reality audio track. This video and audio record proved very valuable when labeling the sessions. In addition, the marker functionality of the BioRadio was used to mark key segments such as the start of a VR application in the data set. Care was taken to be very deliberate when hitting the marker key so that is was both visible and audible on the video recording.

4) Informed Consent: As this research involved human subjects, the research team obtained Institutional Review Board approval before performing any data collection or experiments.

All subjects signed an informed consent form prior to the virtual reality sessions. Immediately before each VR session, subjects were briefed on the content and approximate duration of the session. Subjects were periodically reminded that they were free to stop at any time without repercussion or harm to the research. Subject were also told to be aware of the possibility of nausea and read/accepted the in session Oculus health warning. Only one of five subjects reported nausea and none requested to stop the VR sessions.

B. Virtual Reality Sessions and Stimuli

The virtual reality sessions were chosen from applications available for the Oculus Rift [14] that were also compatible with the Oculus DK2 headset with the following general criteria:

- Readily available content - free on Oculus web store.
- A range of relaxing to exciting scenarios, while avoiding disturbing or mature content.
- Both passive (movies, demos) and active (games) subject involvement.

The generation of intense physiological responses spanning a range of emotions is a challenge. Ethical aspects and subject health must be taken into account when planning experiments. Some of the least upsetting stimuli include colors, music, and self-directed thought experiments e.g. "Imagine a time when you were very happy". With such an experimental setup the degree to which the emotion is experienced, and consequently the level of physiological response, is limited [15]. In order to elicit stronger emotional responses different and likely more upsetting or deceitful techniques have been used. Given the immersive potential of the virtual reality environment extreme content was not included during this experimentation. Some elements included, such as views from high vantage points, could be considered startling for people fear of heights, however the individuals who participated in this experiment did not report any known phobias which could elicit a strong emotional response. The virtual reality sessions used are shown in Table II.

The targeted total time wearing the VR headset was slightly more than *60 minutes*. Early experiments showed that after approximately 1 hour subjects began to tire of the VR environment. The last session was a game which could be terminated at any time based on the subject's fatigue and desire to continue. The total time, including the time required to set up the data collection hardware, describe and launch

TABLE II
VIRTUAL REALITY SESSIONS AND THE CORRESPONDING EXPECTED EMOTIONAL RESPONSES FROM THE TEST SUBJECTS.

VR Session	Expected Emotional Response
Intro to Virtual Reality Demo	First-time VR exposure excitement.
The Rose and I Movie	Relaxation, mild sadness.
Two Discovery Action Videos	Excitement, stress.
InCell Game	Arousal due to fast-paced nature.
Lost Movie	Stress and surprise.
Oculus Dream Deck	Relaxing and exciting videos.
Luckys Tale Game	Arousal, concentration.

the application plus the recording of subjects' responses after each session added up to a total of about 2 and a half hours per subject. This included skin preparation and application of the adhesive electrodes, connection of the electrodes to the BioRadio, adjusting and donning the RIP belts, and mounting of both BioRadios, as well as removal of the equipment after the conclusion of the VR sessions.

Initial experimentation showed that getting consistent and reliable feedback from the subject is challenging. It is quite difficult to describe many of the experiences in consistent emotional terms. Indeed, the question "How did that make you feel?" is an opened ended one, often used in therapy. Even members of the research team who were familiar with the classification of emotion and the arousal-valence model of affect [6], struggled to enunciate what types of emotion a specific video or game induced.

The final methodology employed was a simplified version of the arousal-valence model. Instead of removing the headset to fill out a form, the subjects were asked verbally the following questions after each session:

- Did you find this [movie, game, demo] exciting, relaxing, or neutral?
- Did you find this [movie, game, demo] pleasant, unpleasant, or neutral?

While there was still some hesitation on the subjects' part, especially during longer sessions that had multiple scenes, this simplified oral response method worked much better than the prior methods.

III. DATA PRE-PROCESSING

A. Data Retrieval, Merge, Cleaning, and Labeling

The collected signals were streamed to a laptop running Windows 10 and two instances of the BioCapture program, one for each BioRadio device. For consistency, the recording was started on the head device first and then on the body device. The typical offset involved with switching instances and setting up the second recording was approximately 17 seconds. The keyboard was configured for a marker in the head radio instance of BioCapture. All marker data was captured in the head instance.

After the session completion, the files were exported from the BioCapture program into standard comma delimited text (.csv) format. Given the 250 Hz sampling rate, each minute of collected data generated 15,000 rows in the table. The total number of rows depends on the session length and ranges from approximately 70,000 to 140,000 rows for sessions 1 - 6. Due to the fact that it was left to the subject as to how long to continue Session 7 has a broader range and can exceed 250,000 rows if the subject completes the entire game. The head configuration has 13 columns and the body configuration has 16 columns, depending on the number of data channels collected by each device.

After conversion to .csv format, each of the 14 data files (7 sessions x 2 for Head and Body) was imported in to a MATLAB table. MATLAB Version R2016a was used for this analysis. The separate head and body recordings were joined prior to labeling. Unfortunately, there is no common signal nor ability to add a marker in each file. There are known techniques of synchronizing the files based on cross correlation however given that each of the signals is discrete and the times involved are based on human reaction the files were synchronized using the time-stamp label of each data row available via the BioRadio.

Specifically, an offset was calculated by subtracting the delay from the start of the head recording to the start of the body recording and this was used to align the rows prior to performing a table join. In addition, the start and stop times were also used to discard the setup and takedown segments of each session.

A separate table containing column vectors with the subject and session metadata was also created. The table and its column vectors were joined to the initial table. The BioRadio sampling was very consistent during the data collection and there were no missing data points, however in two cases the data had to be adjusted. The first case is the PulseOximeter data where the sensor generates an out-of-range value when it is unable to get a proper reading which only occurred briefly during periods of subject movement. In this case the out-of-range data was replaced with the last known good value. The second case affected only one subject and was due to an electrode cable becoming disconnected during the data collection. In this case the erroneous data was replaced by an average of the prior readings.

B. Emotional Response and Labeling

The subjects' responses to the two questions regarding level of excitement and pleasantness were graphed on a 9 square grid, a discretized version of the arousal-valence model shown in Fig. 3. The summed up responses for each VR segment for all subjects are shown in Fig. 4. The valence scale was measured as "Unpleasant", "Neutral", or "Pleasant" (from left to right) and the arousal scale was measured as "Relaxing", "Neutral", or "Exiting" (from bottom to top).

The results are very asymmetrical with the majority of the sessions rated as "Exciting" and "Pleasant". This is likely due to the conservative selection of stimuli and the limited number of subjects. For example, only one subject expressed any trepidation regarding heights, and therefore the "Pendulum Swing" and "City Scene which", which both involved a view from very high perspective with a large potential drop off

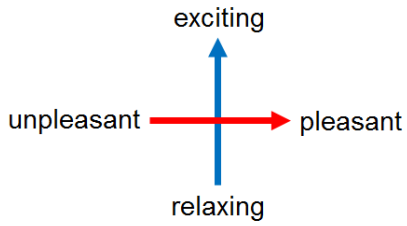

Fig. 3. Subject response graph indicating "pleasantness" and "excitement" as described by the arousal-valence model of affect [6].

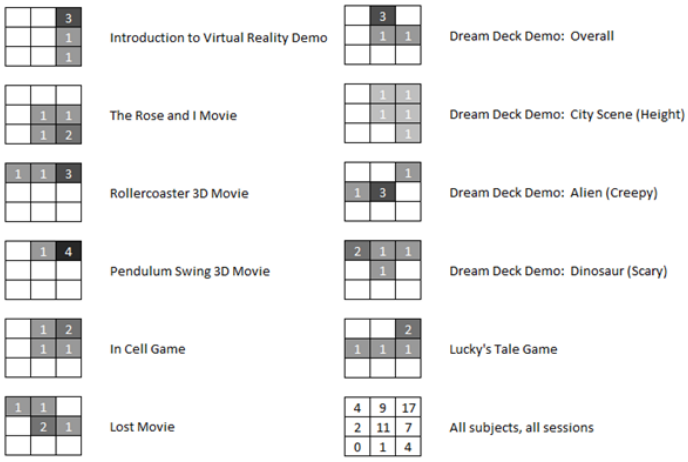

Fig. 4. Subject response results. Each cell shows how many subjects reported feeling the corresponding level of "pleasantness" and "excitement" for each VR session.

was not rated as unpleasant. Given the lack of unpleasant valence results and the limited negative arousal results the classification was limited to the arousal only during the Rose and I, Pendulum, and Roller-coaster Movie sessions.

Classification labeling within each segment was much more manual and required the time information gleaned from the video. A point was made to include a start mark (by pressing the 'S' button on the laptop) in the head file that was audibly and visually visible in the video. Since this data was also clearly present in the data file it served as the alignment index between the video data and the signal data. For several of the sessions the activities were further broken down into segments as previously described in the subject self-reporting section. The video was viewed and time manually input into a spreadsheet to convert to elapsed time in the data file.

IV. Feature Extraction and Classification

Three methodologies were employed for feature extraction. First we tried a simple mean/standard deviation analysis. Second we performed an analysis using domain specific knowledge and extracting features known to be relevant to the specific physiological signals. Finally, a more comprehensive feature set was generated, extracting 90 different general-purpose signal descriptors (features) from each signal channel.

For classification, a leave-one-out cross validation approach was followed, where in each round the data collected from one subject was used as the test set, and the data collected from the remaining subjects was used for training the classifier. This approach ensures that no data samples from the same subject end up in both the training and testing set, which would artificially increase accuracy due to over-fitting on the idiosyncrasies of the absolute signal values collected from each subject.

In our classification experiments, we segmented the signals into 10-second non-overlapping windows, and each segment (window) was labeled and was treated as an independent data sample. Thus, each session yielded multiple data samples, depending on its duration.

Finally, due to the small number of responses in the "relaxing" and "neutral" arousal categories, during classification, the latter two categories were merged into one, thus, leading to a simplified *binary classification problem of "high-arousal" or "moderate/low arousal"*.

Naive Bayes (NB), K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) classifiers were tested, with SVM yielding the highest accuracy in most cases.

In the following subsections we present the arousal classification accuracy results for each approach.

A. Mean and Standard Deviation Features

For an initial analysis, and to establish a baseline, simple mean and standard deviation features were extracted from each signal channel, with a 10 second (2500 sample) sliding non-overlapping window. The resulting MATLAB table contained 48 features, 2 for each of the 24 time based input signals. For the simple mean and standard deviation feature set only a few feature selection experiments were run and did not yield significant improvement to the classification accuracy. For the mean and standard deviation feature set utilizing all 48 features (without feature selection) the leave-one-out accuracy was **74%** using a Support Vector Machine classifier.

B. Domain-Based Feature Extraction

Feature extraction algorithms were developed based on existing knowledge regarding the behavior of the physiological signals with respect to emotional response, similar to ones found in previous studies, e.g. [4]. The 15 features that were computed are described briefly below.

- *meanHR* - the mean of the heart rate as reported by the pulse oximeter was calculated for each segment and was normalized by subtracting the mean of the subject's heartrate for the entire dataset (the base heart rate).
- *magPPV* - the magnitude of the peripheral blood volume as reported by the pulse oximeter was calculated for each segment by subtracting the minimum value from the maximum value.
- *slopeGSR* - the slope of the electrodermal activity (EDA) or skin resistance was calculated by subtracting the value of the last sample in the segment from the value of the first sample in the segment.

- *meanGSR* - the mean value of the skin resistance was calculated as the average of all samples within each segment.
- *slopePT* - the slope of the peripheral skin temperature (SKT) was calculated for each segment by subtracting the minimum value from the maximum value.
- *mECGHR* - the mean heartrate based on the ECG signal was calculated by counting the number of peaks which were greater than 0.5 seconds apart during each segment.
- *HRV* - heart rate variability is a better predictor of emotion than raw heartrate [3]. The variability of the heart rate was computed by taking the maximum distance between adjacent peaks minus the minimum distance between adjacent peaks divided by the average distance between peaks for a given segment.
- *minHRV* - is calculated as HRV except the min peak distance only is used.
- *maxHRV* - is calculated as HRV except the max peak distance only is used.
- *respA, respC* - a similar peak counting method as mECGHR is applied with a 2 second peak to peak minimum for the abdomen and chest RIP signals.
- *respVA, respVC* - the minimum peak to peak distance divided by the mean peak to peak distance is computed for each segment from the abdomen and chest RIP signals.
- *meanfR* - the mean value of the f4 EEG signal is computed. This signal is previously normalized by subject
- *magEMG* - the magnitude of the EMG signal is calculated by subtracting the minimum value in the segment from the maximum value in the segment.

Multiple combinations of features were run during the development of the domain based features as well as some automated testing. The accuracy was found to be highest with the following set of features: *meanHR, magPPV, slopeGSR, mECGHR, HRV*. When each feature was evaluated individually the top 3 performing features were *meanHR*, *magPPV*, and *HRV*. The highest overall leave-one-out accuracy of **80%** was achieved using a Support Vector Machine and the following five features: *meanHR, magPPV, slopeGSR, mECGHR, HRV*.

C. General Purpose Feature Extraction

For the last round of analysis, we utilized a multitude of general purpose signal descriptors, subsets of which have been previously used in various signal processing applications, but to the best of our knowledge, have not been previously combined and tested in physiological biosignal analysis applications. From each signal, the 90 features shown in Table III were extracted resulting in 2160-dimensional feature vectors. The signal descriptors used as features here are only listed by name. The reader may refer to the literature for details about each descriptor type.

Using all 2160 features the leave-one-out accuracy was **57%**. The low classification accuracy is justified by the fact that many of the features included could be just introducing unnecessary noise, and also by the fact that the vector dimensionality is high compared to the number of available data samples.

TABLE III
LIST OF GENERAL PURPOSE FEATURES EXTRACTED.

#	**Signal Descriptor**
1	Average
2	Standard deviation
3-8	PSD[1] - peaks frequency
9-14	PSD - peaks amplitude
15	Energy
16	Zero crossing rate
17	Energy entropy
18	Spectral centroid
19	Spectral spread
20	Spectral entropy
21	Spectral Rolloff point [16]
22-26	MODWT[2] - Energy of Wavelet [17]
27-31	MODWT - Percentage of Energy of Wavelet
32-36	MODWT - Standard deviation of Wavelet
37-41	MODWT - Mean of Wavelet
42	Tsallis entropy [18]
43	Renyi entropy
44	Shannon entropy
45-54	RSP[3] of sub-bands
55	RSP - Slow wave bands-spectral bands Delta
56	RSP - Slow wave bands-spectral bands Theta
57	RSP - Slow wave bands-spectral bands Alpha
58-72	Harmonic parameters [19]
73	Hjorth parameters – Activity
74	Hjorth parameters – Mobility
75	Hjorth parameters – Complexity
76	Skewness
77	Kurtosis
78-87	Autoregressive parameters
88-90	Percentile 25, 50, 75 amplitude

Apparently, a feature selection step is necessary in this situation to reduce the dimensionality and eliminate noisy features. To that end, we experimented with a variety of feature selection methods including two embedded feature selection methods, RFS [20] and HSSL [21], as well as MSVM-RFE [22], mRMR [23] and ReliefF [24]. Table IV summarizes the best accuracy that was achieved by each feature selection method. The embedded feature selection methods in combination with SVM were clear winners in this case, with the method found in [20] (RFS) achieving slightly better accuracy, with a smaller number of features compared to the method found in [21] (HSSL). The best accuracy achieved in this experiment was **89.19%**. For the winning feature selection method, in Table V we present in more detail the accuracy achieved for different amounts of selected features.

It should be noted that some feature selection methods, like SVM-RFE and mRMR return the actual subset of best features selected, whereas other methods, like the two embedded methods used here and Relief-F, only return a ranking of all the features from best to worst. For the latter case, we performed a two-step incremental search to determine the best number of

[1]PSD: Power Spectral Density (multiple frequency ranges (sub-bands)).
[2]MODWT: Max overlap discrete wavelet transform (mult. sub-bands).
[3]RSP: Relative Spectral Power (multiple sub-bands).

TABLE IV
CLASSIFICATION ACCURACY RESULTS USING THE GENERAL PURPOSE FEATURES EXTRACTED AND A COMBINATION OF FEATURE SELECTION AND CLASSIFICATION ALGORITHMS.

Feature Selection Algorithm	Classification Algorithm	Best % Accuracy	Optimal # of Features
RFS [20]	NB	76.12	20
	SVM	**89.19**	42
	KNN	77.11	20
HSSL [21]	NB	82.65	30
	SVM	88.2	60
	KNN	76.76	20
MSVM-RFE [22]	NB	77.42	287
	SVM	77.72	211
	KNN	73.18	6
mRMR [23]	NB	59.55	2
	SVM	63.42	7
	KNN	58.18	2
Relief-F [24]	NB	77.78	10
	SVM	80.34	46
	KNN	73.14	12

TABLE V
CLASSIFICATION ACCURACY RESULTS FOR A VARYING NUMBER SELECTED FEATURES USING THE RFS [20] FEATURE SELECTION METHOD.

# of features	NB	SVM	KNN
10	66.27	69.6	70.51
20	76.12	73.8	77.11
30	75.79	86.56	68.28
40	54.37	88.86	73.19
42	54.05	**89.19**	72.53
50	52.73	88.86	71.23
60	53.4	82.96	63.71

features to keep. In the first step, we added features (stating from the best) in batches of 10, and re-trained and tested the classifier with each round. In the second step, we performed a more refined search, around the multiple of 10 that gave the best result in step 1, by adding one new feature at a time.

V. DISCUSSION AND FUTURE CONSIDERATIONS

Experimental design is critical to the success of this research. Testing the stimuli and subject self-reporting responses prior to the more complex experiments with the full sensor setup would be beneficial. Unfortunately, this will likely require a much larger subject population as some level of desensitization will occur, therefore, it would be best to perform the data collection on subjects who have not previously participated in the virtual reality simulations. In order to better cover the relaxing portion of the arousal-valence space, calming or meditative segments should be added. More challenging will be the identification of emotionally unpleasant stimulus. Given the immersive nature of the virtual reality environment care must be taken not to cause undue stress with explicit or unsettling content. A thorough IRB review of the stimulus, subject selection, and procedures would be warranted with this type of content.

Some physiological signals varied significantly between subjects, with no universal telltale markers found. If possible, selecting subjects based on presence or absence of sensitivity to the intended stimuli would be beneficial. For example, a dataset with 50% of the subjects expressing a fear of heights and 50% having no fear of heights engaged in a virtual reality simulation that involves height and drop simulations would be very interesting.

In order to increase the number of subjects several elements will need to be improved. Less cumbersome equipment with better cable management and/or wireless sensors along with sensors built into headset or other wearable type garments instead of adhesive type electrodes would make the simulations much more pleasant and also decrease the required setup time. Event marking should be as automated as possible. Video recording with remote audio capabilities was invaluable for this preliminary research but reviewing each video to determine event label start and stop points required a large amount of time. Additionally, this manual technique increases the risk of error and timing inaccuracies.

Feature extraction seems to play a significant role in the outcome, however, there are no standard feature types universally acceptable for physiological biosignal analysis. Domain knowledge will likely need to be combined with general purpose features for each type of signal to further improve accuracy. For example, the rise and recovery of a GSR event is well documented and fairly easy to spot visually on a graph however specific feature extraction will be required to form a marker for this type of event in the feature table. In addition, motion is likely not relevant for the direct prediction of emotional response, however it could prove to be valuable in detecting and eliminating noise and motion related artifacts from the signals that are relevant.

Furthermore, a qualitative examination of the types of features that are consistently ranked low by feature selection algorithms may indicate that they are inappropriate for the task at hand. If such signals tend to come from certain physiological channels, it may be possible to eliminate those channels altogether, without significant loss in accuracy, thus reducing the obtrusiveness of the data collection process.

With the continued progress in sensing technology and through the application of machine learning on large datasets, the classification of human emotion will help guide therapy, training, and the development of improved experiences with automated systems that include affective computing capabilities.

VI. CONCLUSION

In this work we presented our findings and observations towards emotion recognition from physiological biosignals collected during user interaction with virtual reality environments. As emotion is a subjective concept, the task of automating emotion recognition is a particularly challenging one due the inherent difficulty of generating reliable ground truth data. Furthermore, the current sensor technology for physiological signal acquisition has certain limitations which extend to both the inconvenience for the user and the quality of the data collected.

Nevertheless, significant effort has been put by researchers towards the goal of emotion recognition from physiological data with varying degrees success. One of the limiting factors is the absence of a universally acceptable methodology for feature extraction from a variety of physiological biosignals. Our work shows that analysis of biosignals does not require domain specific knowledge and that a generic set of features can be extracted from a variety of biosignals. This general purpose features in combination with strong feature algorithms can exceed the performance achieved by developing and using domain specific features.

ACKNOWLEDGMENT

L.B.H. thanks The Graduate College, Texas State University for the Research Support Fellowship that supported the purchase of additional equipment used during this research.

REFERENCES

[1] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[2] C. Collet, E. Vernet-Maury, G. Delhomme, and A. Dittmar, "Autonomic nervous system response patterns specificity to basic emotions," *Journal of the autonomic nervous system*, vol. 62, no. 1, pp. 45–57, 1997.

[3] R. D. Lane, K. McRae, E. M. Reiman, K. Chen, G. L. Ahern, and J. F. Thayer, "Neural correlates of heart rate variability during emotion," *Neuroimage*, vol. 44, no. 1, pp. 213–222, 2009.

[4] E.-H. Jang, B.-J. Park, S.-H. Kim, M.-A. Chung, M.-S. Park, and J.-H. Sohn, "Emotion classification based on bio-signals emotion recognition using machine learning algorithms," in *Information Science, Electronics and Electrical Engineering (ISEEE), 2014 International Conference on*, vol. 3. IEEE, 2014, pp. 1373–1376.

[5] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*. ACM, 2004, pp. 205–211.

[6] J. Posner, J. A. Russell, and B. S. Peterson, "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology," *Development and psychopathology*, vol. 17, no. 03, pp. 715–734, 2005.

[7] R. Ortner, E. Grünbacher, and C. Guger, "State of the art in sensors, signals and signal processing," *Guger Technologies OG, Graz, Austria*, 2013.

[8] M. Z. Jamal, *Signal acquisition using surface EMG and circuit design considerations for robotic prosthesis*. INTECH Open Access Publisher, 2012.

[9] S. Day, "Important factors in surface EMG measurement," *Bortec Biomedical Ltd publishers*, pp. 1–17, 2002.

[10] E. P. Scilingo, A. Gemignani, R. Paradiso, N. Taccini, B. Ghelarducci, and D. De Rossi, "Performance evaluation of sensing fabrics for monitoring physiological and biomechanical variables," *IEEE Transactions on information technology in biomedicine*, vol. 9, no. 3, pp. 345–352, 2005.

[11] Great Lake Neurotechnologies., "BioRadio User Guide, Rev E [Online]." https://www.glneurotech.com/bioradio/wp-content/uploads/2015/02/392-0050-Rev-E-BioRadio-User-Guide.pdf, accessed: 2017-03-31.

[12] P.-Y. Carry, P. Baconnier, A. Eberhard, P. Cotte, and G. Benchetrit, "Evaluation of respiratory inductive plethysmography: accuracy for analysis of respiratory waveforms," *Chest*, vol. 111, no. 4, pp. 910–915, 1997.

[13] J. E. Sinex, "Pulse oximetry: principles and limitations," *The American journal of emergency medicine*, vol. 17, no. 1, pp. 59–66, 1999.

[14] "Oculus Rift Applications [Online]." https://www.oculus.com/experiences/rift/, accessed: 2017-03-31.

[15] A. Gerrards-Hesse, K. Spies, and F. W. Hesse, "Experimental inductions of emotional states and their effectiveness: A review," *British journal of psychology*, vol. 85, no. 1, pp. 55–78, 1994.

[16] M. Kos, Z. Kacic, and D. Vlaj, "Acoustic classification and segmentation using modified spectral roll-off and variance-based features," *Digital Signal Processing*, vol. 23, no. 2, pp. 659 – 674, 2013.

[17] S. Khalighi, T. Sousa, G. Pires, and U. Nunes, "Automatic sleep staging: A computer assisted approach for optimal combination of features and polysomnographic channels," *Expert Systems with Applications*, vol. 40, no. 17, pp. 7046 – 7059, 2013.

[18] C. Tsallis, R. Mendes, and A. Plastino, "The role of constraints within generalized nonextensive statistics," *Physica A: Statistical Mechanics and its Applications*, vol. 261, no. 34, pp. 534–554, 1998.

[19] W.-C. Tang, S.-W. Lu, C.-M. Tsai, C.-Y. Kao, and H.-H. Lee, "Harmonic parameters with HHT and wavelet transform for automatic sleep stages scoring," *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, vol. 1, no. 9, pp. 414–417, 2007.

[20] F. Nie, H. Huang, X. Cai, and C. Ding, *Efficient and robust feature selection via joint $L_{2,1}$-norms minimization*, 12 2010.

[21] V. Metsis, F. Makedon, D. Shen, and H. Huang, "DNA copy number selection using robust structured sparsity-inducing norms," *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 11, no. 1, pp. 138–181, Jan. 2014. [Online]. Available: http://dx.doi.org/10.1109/TCBB.2013.141

[22] X. Zhou and D. P. Tuck, "MSVM-RFE: extensions of SVM-RFE for multiclass gene selection on DNA microarray data," *Bioinformatics*, vol. 23, no. 9, pp. 1106–1114, 2007.

[23] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[24] I. Kononenko, E. Šimec, and M. Robnik-Šikonja, "Overcoming the myopia of inductive learning algorithms with RELIEFF," *Applied Intelligence*, vol. 7, no. 1, pp. 39–55, 1997.