Int. J. Human-Computer Studies 110 (2018) 75–85

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Building predictive models of emotion with functional near-infrared spectroscopy

Danushka Bandara[a,*], Senem Velipasalar[a], Sarah Bratt[b], Leanne Hirshfield[c]

[a] Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244, USA
[b] Information School, Syracuse University, Syracuse, NY 13244, USA
[c] S.I. Newhouse School of Communications, Syracuse University, Syracuse, NY 13244, USA

ARTICLE INFO

Keywords:
fNIRS
Affective computing
Brain signal processing
Emotion classification
Valence classification
Arousal classification

ABSTRACT

We demonstrate the capability of discriminating between affective states on the valence and arousal dimensions using functional near-infrared spectroscopy (fNIRS), a practical non-invasive device that benefits from its ability to localize activation in functional brain regions with spatial resolution superior to the Electroencephalograph (EEG). The high spatial resolution of fNIRS enables us to identify the neural correlates of emotion with spatial precision comparable to fMRI, but without requiring the use of the constricting and impractical fMRI scanner. We make these predictions across subjects, creating the capacity to generalize the model to new participants. We designed the experiment and evaluated our results in the context of a prior experiment—based on the same basic protocol and stimulus materials—which used EEG to measure participants' valence and arousal. The F1-scores achieved by our classifiers suggest that fNIRS is particularly useful at distinguishing between high and low levels of valence (F1-score of 0.739), which has proven to be difficult to measure with physiological sensors.

1. Introduction

Accurately assessing human emotion has long been a goal of researchers and practitioners in human-computer interaction (HCI), as emotion is essential for understanding users' experiences with new technologies and for designing affect-based adaptive systems (Russell, 1980; Koelstra et al., 2012; Izzetoglu et al., 2004). Emotion is a complex phenomenon often difficult to recognize for humans, never mind machines (Plutchik, 2001). While emotions are frequently measured with self-report surveys, many HCI researchers recognize the shortcomings associated with self-report methods, such as the tendency to inaccurately assess personal emotions. Furthermore, these self-report techniques are administered after a task completion which interrupts the user experience and fails to capture real-time information about the user's changing emotional states during the task. For this reason, researchers have attempted to measure and predict changing emotional states using a variety of objective physiological sensors such as functional magnetic resonance imaging (fMRI), Electroencephalography (EEG), Galvanic Skin Response (GSR), and Heart Rate Variability (HRV), as detailed in the next section.

While much progress has been made in objectively measuring and predicting user emotions, further interdisciplinary research is needed to develop robust models for accurately predicting real-time changes in emotional state. As biologists and neuroscientists continue to analyze the physiology of emotion, biotechnology experts are developing new non-invasive sensors that are practical, robust to noise, and highly accurate (Parasuraman and Rizzo, 2008; Colibazzi et al., 2010). Meanwhile, computer scientists continue developing machine learning and data mining models capable of making real-time predictions from this wide array of multi-modal physiological sensor data (Berka et al., 2007; Grimes et al., 2008; Lee and Tan, 2006; Koelstra et al., 2012).

The focus of our research involves the use of functional near-infrared spectroscopy (fNIRS), a relatively new, non-invasive brain measurement technique that is resilient to noise, portable, and allows for naturalistic participant movement (as compared to fMRI). Further, fNIRS has higher spatial resolution than EEG and enables the localization of specific brain regions of activation while taking measurements under normal working conditions (Girouard et al., 2009; Hirshfield et al., 2007; Hirshfield et al., 2009; Hirshfield et al., 2011). Our goal is to leverage the high spatial resolution of fNIRS to develop machine learning classifiers capable of predicting valence and arousal in participants with a high degree of accuracy. In the experiment described in this paper, participants' brain function was measured with fNIRS while they viewed a variety of clips extracted from music videos. These videos have been shown to elicit various levels of valence and arousal. After each video clip they filled out the Self-Assessment Manikin (SAM) (Bradley and Lang, 1994) to indicate their valence and arousal. The self-report values from the

* Corresponding author.
E-mail address: dsbandar@syr.edu (D. Bandara).

https://doi.org/10.1016/j.ijhcs.2017.10.001
Received 6 October 2016; Received in revised form 25 August 2017; Accepted 4 October 2017
Available online 5 October 2017
1071-5819/

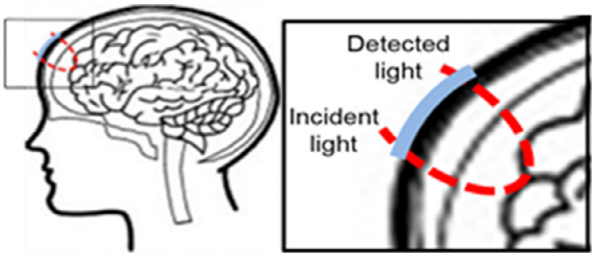


Fig. 1. Near-infrared light is emitted from a diode into the cortex, and detectors measure the light reflected out of the cortex.

SAM were used as labels during subsequent supervised machine learning classification.

This research makes two primary contributions in the realm of HCI: First, we demonstrate the capability of classifying and distinguishing between affective states on the valence and arousal dimensions using fNIRS, a practical non-invasive device. Our fNIRS results show that specific functional brain regions are recruited during changes in valence and arousal and these regions are consistent with those identified by fMRI research on emotion. Second, we develop models to classify and predict emotional states across subjects, creating the capacity to generalize the model to new participants rather than training each model per individual. The F1-scores achieved by our classifiers suggest that fNIRS is particularly useful at distinguishing between levels of valence, which has proven to be difficult to measure with physiological sensors.

The remainder of this paper proceeds as follows. We first provide related background information and a review of the relevant literature. Next, we describe our experimental set-up and protocol. We then report the analysis procedures and results and discuss findings in the context of our research goals. Last, we describe study limitations and possible avenues for future work stemming from this research.

2. Background and literature review

This section describes the fNIRS device and how it compares to other popular brain measurement techniques. We then describe conceptualizations of emotion, the measurement of emotion using subjective and objective brain measurement techniques, and describe challenges faced in conducting research using machine learning on cognitive data for emotional state predictions.

2.1. Review of brain measurement techniques

The measurement of brain activity has significant potential for evaluating the physiological correlates of emotion. Sensor technologies such as functional magnetic resonance imaging (fMRI) and positron emission tomography (PET) provide valuable insight into the functions and structures of the brain. However, they constrain subject movement and in the case of fMRI, require that subjects remain completely still. Further, they can expose subjects to hazardous materials (PET) or to loud noises (fMRI) (Izzetoglu et al., 2004) and are not ideal for assessing the neural activity of participants under normal working conditions. The use of the electroencephalograph (EEG) has attracted researchers interested in non-invasively measuring users' brain activity (Lee and Tan, 2006; Grimes et al., 2008). The EEG has gained popularity for research use because of its cost-effectiveness, ease of use, and granular temporal resolution. In the 1990s fNIRS was introduced, a tool which can augment and overcome some of the limitations of EEG and other brain-imaging devices (Chance et al., 1993). The fNIRS device pulses near-infrared light in the wavelength range (690–900 nm) into the brain (Fig. 1).

The primary absorbers of near-infrared light are deoxygenated hemoglobin (Hb) and oxygenated hemoglobin (HbO) in tissues. During hemodynamic and metabolic processes, these light values change in association with neural activity in the brain (Chance et al., 1993). These metabolic changes can then be detected through the measurement of the diffusively reflected light pulsed into the brain cortex (Chance et al., 1993; Izzetoglu et al., 2004; Son et al., 2005). The use of fNIRS includes measuring a range of cognitive states while computer operators engage in tasks during normal working conditions (Hirshfield et al., 2014, 2015; Ehring et al., 2010; Heilman et al., 2010; Colibazzi et al., 2010). In the next section, we describe the construct of emotion and provide a review of the literature on the subjective and objective measurement of emotion.

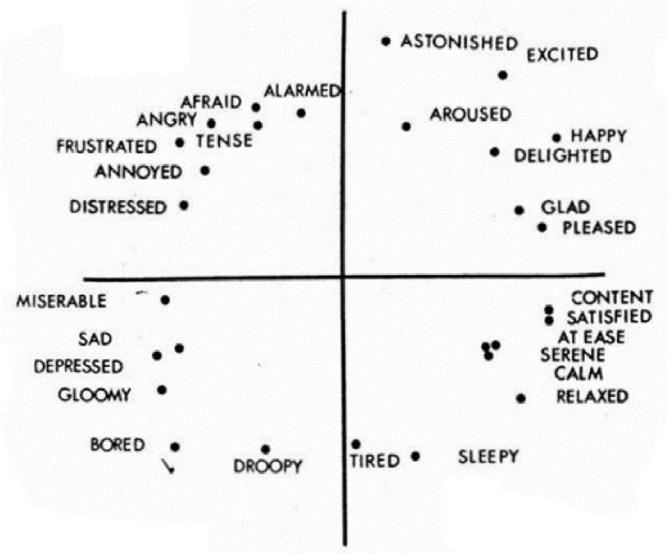


Fig. 2. The Circumplex model of affect. Horizontal axis shows degree of valence (pleasure/displeasure) and vertical axis shows degree of arousal (Russell, 1980).

2.2. The construct of emotion

Research on emotion has increased significantly over the past two decades with many fields contributing; including psychology (Ehring et al., 2010; Heilman et al., 2010; Colibazzi et al., 2010; Plutchik, 1980), neuroscience (Davidson et al., 2000; Lane and Nadel, 1999), medicine (Pressman and Cohen, 2005; Salovey et al., 2000), sociology (Massey, 2002; Scherer, 2005), and computer science (Izzetoglu et al., 2004; Hirshfield et al., 2007).

Among this recent surge, most researchers agree that emotions are affective states that exist over a relatively short period, with durations ranging from milliseconds to minutes, and are related to an event (Ortony et al., 1990; Picard and Picard, 1997). A frequently used metric for quantifying emotions is by mapping them to points in a two-dimensional space of affective valence and arousal. Valence represents overall pleasantness of an emotional experience and arousal represents the intensity level of an emotion, ranging from calm to excited (Russell, 1980; Bradley and Lang, 1994; Ball and Breese, 1999). These two dimensions enable us to differentiate between four basic categories of emotions. Some models of emotion identify nine categories of emotion by including a neutral section. In principle, an infinite number of other categories can be defined (Bosma and Andre, 2004), but Russell's (1980) circumplex model of affect is a widely used and well-vetted (Kring et al., 2003; Feldman and Russell, 1998; Russell, 1991) model in contemporary HCI research. Fig. 2 depicts the arousal and valence dimensions and their relation to emotional state (Russell, 1980).

Other conceptual models of emotion include Ekman and Friesen's (1971) model based on discrete sets of universal emotions and Plutchik's (2001). The literature on modeling emotion is reviewed further by Posner et al. (2005). They provide support for the two-dimensional model of affect with examples from empirical studies.

Self-report surveys, such as the Self-Assessment Manikin (SAM) or the Positive and Negative Affect Scale (PANAS) survey instruments (Watson et al., 1988) are most commonly used to locate a person's perceived emotion within the circumplex model (Russell, 1980). In our experiment, we use the SAM, a pictorial assessment technique for evaluating the pleasure, arousal, and dominance associated with subjects' affective reactions to stimuli.

2.3. Objective measures of emotion

To overcome the subjective limitations of self-report surveys, a range of objective sensors have been used to measure emotion. Charles Darwin formally documented variations in facial expressions associated with specific discrete emotions (Darwin, 1965). Schwartz and his colleagues built on the pioneering work of Darwin and recorded facial EMG of subjects engaging in pleasant and unpleasant mental imagery (Brown and Schwartz, 1980; Schwartz et al., 1976; Ekman et al., 1983). Facial Electromyography (EMG), the recording of electrical signals associated with facial muscle activity, is used extensively as a measure of emotional state (Fridlund et al., 1984; Dimberg, 1988; Warren et al., 2006). Early research found that conscious experiences of emotion evoke specific physiological activity (Cannon, 1927). Another view is that distinct experiences of emotions are produced by a continuous interaction of both mind and body (Schachter and Singer, 1962). More recent research has shown that psychological states evoke skin conductance changes when a user is presented with emotionally charged pictures (Lang et al., 1993), computer games (Bailey et al., 2009) and emotional films (Codispoti et al., 2008).

Extensive literature has examined the strong association of emotion with cognition and brain activity. Cacioppo et al. (1999) suggested that emotion helps construct cognition and cognition helps construct emotion. Further, methods for identifying neural networks associated with different semantic emotional states in the brain were developed (Kober et al., 2008) as well as using Russell's 2-dimensional valence/arousal model (Colibazzi et al., 2010). With its capability to localize specific brain regions of activation, fMRI studies have measured the neural correlates of emotion in the brain. For example, Viinikainen et al. (2010) showed participants affective images from the International Affective Picture System (IAPS; Lang et al., 1999) database while in the fMRI scanner. Results showed that both arousal and valence manifested different types of responses to negative and positive stimuli in the brain and suggesting that there are different valence and arousal representations in the brain for negative and positive (unpleasant and pleasant) stimuli. They also note that the medial prefrontal cortex (mPFC), a large region spanning the front portion of the human brain, is important in the processing of emotions. In another fMRI study, Colibazzi et al. (2010) found that changes in arousal directly affected the supplementary motor cortex, and several deep brain regions that support the limbic system. Changes in valence also effected the supplementary motor cortex, as well as the dorsolateral prefrontal cortex (DLPFC), inferior parietal cortex, and the frontopolar cortex. The researchers suggest that the more unpleasant the emotion the higher the activity in the DLPFC and frontopolar cortex. Also, the researchers note that the supplementary motor cortex may constitute an interface between limbic and motor-executive systems, whereby the brain transforms affective experiences into complex motor plans. For example, a feeling of excitement that drives a desire to dance activates the motor cortex even if the movement is not actually executed.

The emotion-related activation in the pre-motor cortex was complemented by recent research by Warren et al. (2006), who made concurrent fMRI and EMG recordings of participants while they listened to auditory sounds designed to elicit different levels of valence and arousal. Their results showed that positive auditory-induced emotions engage the pre-motor cortex, by causing the brain to automatically prepare for responsive facial gestures to the affective stimuli. In other words, sounds that induce positive emotions engage the pre-motor cortex, as the brain is preparing to create a facial gesture, such as producing a smile.

Several fMRI studies of music and emotional states also found the pre-motor cortex to be directly related to the emotional experience of music. For example, they suggested sad pieces of music contrasted with happy pieces by producing differing activations in the DLPFC, frontopolar cortex, and superior temporal gyrus. These regions have also been associated with emotional experiences, introspection, and self-referential evaluation (Brattico et al., 2011). Broca's area – the central region for language processing – has also been linked to the emotional experience of lyric-based music (Fadiga et al., 2009).

Recent work in EEG and fNIRS have also focused on measuring emotion in the brain. For example, León-Carrión et al. (2007) used fNIRS to measure the DLPFC region of participants' brains during their experiences of different emotional states. They found that increases in participants' subjective arousal correlates with activation in DLPFC. Rodrigo et al. (2016) conducted an experiment that compared the subjects' emotional response to neutral and fearful faces using fNIRS. They found that some regions of the PFC (right medial) showed increased activity when viewing fearful faces. In another study, Balconi et al. made concurrent measurements of EEG, Heartrate, and fNIRS data while participants viewed IAPS pictures and filled out the Self-Assessment Manikin for self-assessed valence and arousal ratings. Both the fNIRS and EEG results showed an increase in activation on the right side of the frontopolar region relating to negative emotions.

2.4. Challenges in using machine learning on cognitive data

Several of the studies described above use single trial classification on cognitive data to predict emotional state. It is worth noting that using machine learning (ML) techniques on cognitive data (whether it be from EEG, fNIRS, fMRI, or some other measurement technique) is non-trivial in terms of the difficulty of data preparation, cleaning noise artifacts, feature generation, and algorithm selection and parameter adjustment. Although ML has the potential to help researchers maximize the use of neurophysiological sensors in a variety of domains, there are significant research challenges to using ML on cognitive data. For example, the high dimensionality of sensor data coupled with small sample sizes produces datasets that can be susceptible to model overfitting. Smaller subject populations provide less data for ML algorithms to train on, making the development of across subject model development and generalizability particularly challenging (Picard and Picard, 1997).

Because the brain is a highly-individualized structure, most ML on brain data trains and tests classifiers at the individual level (Grimes et al., 2008; Berka et al., 2007). These classifiers have been found to improve dramatically as training time and model development improve. However, lengthy training sessions can be laborious. While well suited to research in areas such as Brain Computer Interfacing (where a user spends days, and even months training his or her medical system), long training sessions may not be ideal in the HCI domain for users. One way to increase training data is to merge datasets from multiple subjects, enabling the classifier to train and test models based on data from many people and test the models on new individuals (Churchill et al., 2014).

Across subject ML on cognitive data has been explored by a handful of researchers in the HCI domain, but classification accuracy tends to be lower than that achieved by training each model per individual. One reason for this finding is that each person's brain has slight differences, and the placement of sensors on each person's brain may differ spatially. Thus, one 'channel' of EEG or fNIRS data on one participant, may be quite different than the same 'channel' of data on another participant, making it difficult to generate and compare features in a meaningful way for inter-subject comparisons.

Furthermore, the state of the art fNIRS-based emotion research has its own set of issues, as highlighted by Bendall et al. (2016). They highlight the challenge of separating emotional activity from the other cog-

nitive processes in the prefrontal cortex. Mentioning the importance of good experimental design when it comes to the study of emotion using fNIRS. They also identify the lack of sufficient experimental conditions, where some studies choose to use only positive and negative emotional conditions (Hoshi et al., 2011) where others include positive, negative as well as the neutral condition (Glotzbach et al., 2011), which makes it difficult to compare results between experiments.

In addition, fNIRS signals can be affected by peripheral responses such as facial muscle movements and changes in cardiovascular activity (Doi et al., 2013). Another issue pointed out by Doi et al. is the possibility of the subjective emotional response and the neural responses lasting longer than the length of the stimuli. The length of stimuli needs to be decided with this consideration. They also mention how the selection of appropriate indicators of cortical activation (oxygenated blood flow vs deoxygenated blood flow) can affect the analysis of emotion. Also, the individual biological differences need to be taken into consideration when analyzing data between subjects. This effect has so far been difficult to investigate due to the small size of the fNIRS datasets available (15–60 participants).

3. Experiment

Our experiment goal was to induce a variety of emotional states in participants while measuring the hemodynamics of their brain with fNIRS. We aimed to demonstrate the use of the resulting fNIRS brain data to identify emotional state. Specifically, we aimed to develop across subject classifiers to accurately predict emotional state. For the sake of consistency, throughout the rest of the paper we will consider emotion to be an affective mental state as perceived by the person, as elaborated upon in Section 2, thus quantifiable with self-report surveys. We use the SAM measure of emotion as our 'ground truth' measure of emotion.

Twenty healthy, college age participants from a university in the Northeast took part in the experiment (13 male, 7 female). Upon arrival to the lab, participants provided informed consent and completed a pre-questionnaire to obtain their demographic data. They were then provided with instructions explaining the experiment and how to fill out the post task surveys.

3.1. Selection of stimulus material

The widely-used databases for emotion elicitation are International Affective Picture System (IAPS) and International Digitized Sound System (IADS) (Bradley, 1999). In this study, we chose music video clips from the Dataset for Emotions Analysis using Physiological signals (DEAP) dataset (Koelstra et al., 2012) because prior studies have found that visual-audio stimulus gives a better result than using either visual stimulus or audio stimulus (Baumgartner et al., 2006).

A subset of music videos from the DEAP dataset were selected as stimuli to elicit participants' emotions. The DEAP dataset experimenters preselected 120 videos using the emotion related tags from last.fm, and using a manual selection method. With this, their goal was to make sure to choose videos that fit in the four quadrants of the circumplex model (Fig. 2). Then they used a web-based subjective assessment experiment with 14 volunteers to further rate the music videos on valence and arousal scales. The resulting processed scores (mean/standard deviation) for valence and arousal were used as coordinates to place the music videos on the circumplex model. Then the final 40 videos were chosen that constituted regions in the circumplex model representing five experimental conditions of High Valence Low Arousal (HVLA), High Valence High Arousal (HVHA), Low Valence High Arousal (LVHA), Low Valence Low Arousal (LVLA) and Neutral Valence Neutral Arousal (N). We selected fifteen of these videos– three videos to represent each of the above five conditions. Videos were intentionally selected to maximize the expected emotional reaction of participants; that is, the HVLA, HVHA, LVHA, LVLA videos were handpicked from the results reported by Koelstra et al. that were as far away from circumplex model's "neutral" center as possible. The purpose of this selection was to ensure that each participant's brain state was maximally representative of the quadrants in the valence/arousal space.

3.2. Equipment setup

The experiment was performed in a controlled laboratory environment. The fNIRS signals from participants' brains were recorded using a Hitachi ETG-4000 fNIRS device with a sampling rate of 10 Hz. The device provides 52 channels of brain activity data from the frontal region of the participant's brain. (Hitachi Medical Systems, 2016)

Each participant was seated on the experiment chair and the chair was adjusted to his or her comfort level. The fNIRS probe (Fig. 3) was a 3×11 probe with 17 light sources and 16 detectors, resulting in 52 locations measured on the head. The distance between all light source and detector on the ETG-4000 is 3 cm, resulting in a measurement depth into the average adult brain of 2–3 cm (Cui et al., 2011). Once the fNIRS probe was in place, a 3d digitizer was used to record the locations of each fNIRS channel on that subject's head.

3.3. Protocol

After starting the recording of physiological data, the participant viewed a series of music videos. Each video was 60 seconds long. After the video ended, participants filled out the SAM survey for self-report assessment of valence (Likert ratings of 1–5) and arousal (Likert ratings of 1–5).

Since a hemodynamic response triggered by an event typically shows an increase in signal lasting 10–12 s to rise to peak and return to baseline (Crosson et al., 2010), the rest period between the videos was chosen to be 15 seconds. After this 15 s rest to allow neural activity to return to baseline, participants began watching the next video. Fig. 4 shows a screen shot of one of the music videos and a REST screen.

The protocol followed a block design format. The music videos were separated into three blocks, each containing videos from the DEAP dataset with five unique emotion labels, and included the conditions of Low Valence/Low Arousal (LVLA), High Valence/High Arousal (HVHA), Low Valence High Arousal (LVHA), High Valence/Low Arousal (HVLA), and Neutral Valence Neutral Arousal. A Note that the 'neutral' condition was included with music videos that were found by Koelstra et al. to be neutral on both the valence and arousal dimensions. The order within blocks was selected to ensure that within each block of videos the stimuli were presented in a random manner to the participant (so that the participants would not be able to easily guess which type of video would be played next, and to avoid the possible confounding effects of having the same-emotion-inducing video in a row, which would result in an intensified emotion effect), while still ensuring that each block in the experiment contained one video from each of the five conditions above.

4. Data analysis and results

4.1. Survey data analysis and results

Responses to the valence and arousal items from the Self-Assessment Manikin (SAM) are made on two 5-point scales. Before beginning analyses, we looked for agreement between the SAM survey data reported by our participants and the expected results, based on the label of each video within the DEAP dataset. It is well known in the emotion literature that individual and cultural differences effect one's emotional response to a given stimulus (Kuppens and Tong, 2010; Scollon et al., 2011). Despite these individual differences, if we look at the survey data in aggregate, we would expect that a video in the DEAP dataset with a label of high valence would, on average, result in similar ratings on the SAM when our participants watched that video. For example, the song 'What a Wonderful World' by Louis Armstrong (Table 1) was labeled at high

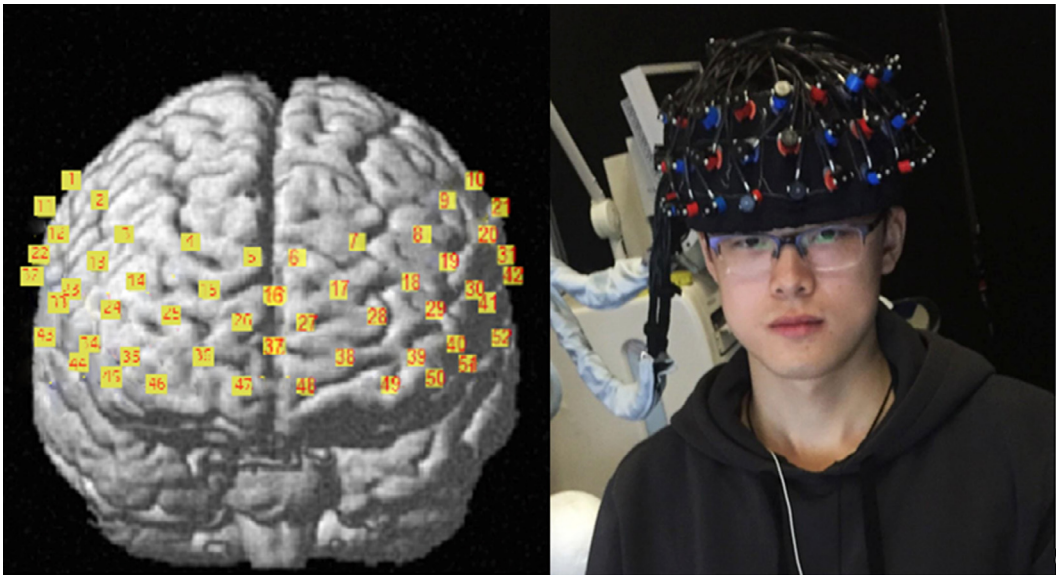


Fig. 3. A participant wearing the fNIRS sensors.

Fig. 4. Music video stimuli that was presented to the user and the rest screen that was shown in between trials.

Table 1
Block design of experiment. Videos within each block were randomized.

Music video block 1	
1.	A Fine Frenzy, Almost Lover: Low Valence Low Arousal (LVLA)
2.	Black Eyed Peas, My Humps: High Valence Low Arousal (HVLA)
3.	Blur, Song 2: High Valence High Arousal (HVHA)
4.	Smashing Pumpkins, 1979: Neutral (N)
5.	Stigmata, В отражении глаз: Low Valence High Arousal (LVHA)
Music video block 2	
1.	Sia, Breathe Me: Low Valence Low Arousal (LVLA)
2.	Christina Aguilera, Lil' Kim, Mya, Pink, Lady Marmalade: High Valence High Arousal (HVHA)
3.	Napalm Death, Procrastination on the Empty Vessel: Low Valence High Arousal (LVHA)
4.	Madonna, Rain: Neutral (N)
5.	Taylor Swift, Love Story: High Valence Low Arousal (HVLA)
Music video block 3	
1.	Glen Hansard, Falling Slowly: Neutral (N)
2.	White Stripes, Seven Nation Army (HVHA)
3.	Trapped Under Ice, Believe: Low Valence High Arousal (LVHA)
4.	Wilco, How to Fight Loneliness: Low Valence Low Arousal (LVLA)
5.	Louis Armstrong, What a Wonderful World: High Valence Low Arousal (HVLA)

valence low arousal in the DEAP dataset, and we would expect most participants to feel these pleasant and serene emotions while viewing the video. However, an individual who doesn't like that song, or who is in a hurry to complete the experiment and collect compensation, may experience the slow-paced song in a different way than others. This would result in slightly different emotional experiences due to individual differences. We were curious to see whether the subjective responses from our participants were, on average, in agreement with the labels from the DEAP dataset. So, we took all videos with a DEAP dataset label of HVHA and computed the average of our respondents' valence and arousal self-report scores reported after they saw that video. We did the same for the rest of the experimental conditions. Average results are shown in Table 2, with Fig. 5 depicting a more detailed view comparing our participants' survey responses on valence arousal and their agreement with the labels from the DEAP dataset.

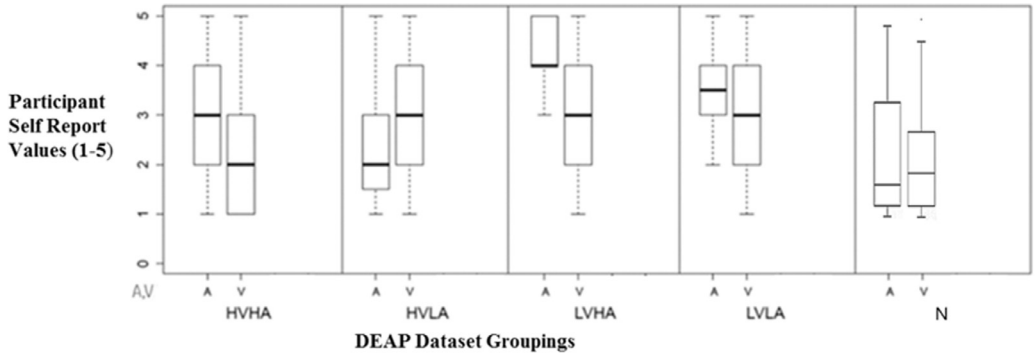


Fig. 5. The rating distribution of participants' responses using the labels provided by DEAP dataset (A = Arousal, V = Valence).

Table 2
Comparison of DEAP labels to self-report surveys from experiment.

DEAP grouping	Average Valence from survey	Average Arousal from survey
HVHA	3.38	2.95
HVLA	3.90	2.83
LVHA	2.35	3.15
LVLA	2.95	2.33
N	1.38	1.38

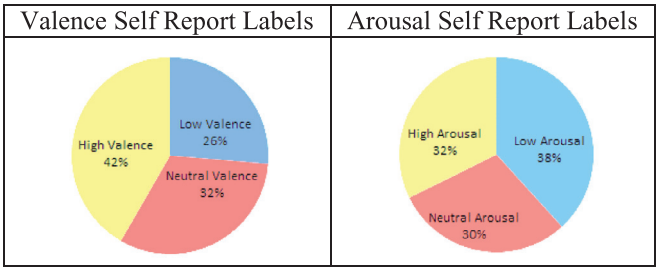


Fig. 6. Post task survey label distribution among the subjects. The numbers represent the percentage of trials, across participants, that were labeled as the respective arousal or valence with Likert scores of 1, 2 = low, 3 = neutral, 4, 5 = high.

Our survey results showed a moderate amount of agreement between the DEAP dataset labels and our participants perceived emotional responses. Notice the wide range of responses shown in the rating distribution (Fig. 5), showing there were varied responses to each video. In fact, the subjects in this experiment did not report complete agreement; neither amongst themselves, nor with the DEAP dataset.

Our participants disagreed with DEAP's labels in the HVHA condition –participants on average said they felt 2.95 (just under "neutral") for arousal, instead of above 3. Second, Neutral videos were reported to elicit Low Valence and Low Arousal (LVLA) instead of an average of '3' for a neutral response. In line with prior research on individual differences in emotional experiences (Kuppens and Tong, 2010; Scollon et al., 2011), this disagreement illustrates the fact that different individuals can have different, highly individualized, emotional reactions to stimuli, and it is essential to gauge each participant's self-reported reaction to stimuli, rather than assuming the stimuli will affect all individuals in the same way.

4.2. Label selection

The self-report valence scale was a five-item scale (1–5), and each participant's response was rounded to the closest integer and collapsed to Low (Likert scores of 1–2), High (Likert scores of 4–5) and Neutral (Likert scores of 3) Valence. The self-report arousal scale was also a five-item scale, and the same process was used to collapse each participant's response into Low, High, and Neutral Arousal. Fig. 6 shows the distribution of the resulting valence and arousal labels for all participants.

These collapsed labels were used as the labels for subsequent supervised machine learning. It is apparent from Fig. 6 that there is an uneven distribution between class labels, which will cause unbalanced datasets for machine learning. This is a common issue when conducting machine learning on participants' self-report data, as individuals are likely to experience different emotional reactions to various stimuli, as reflected in their self-reports. We account for this imbalance by reporting F1-scores, because F1-scores are commonly used in lieu of overall accuracy in the presence of small unbalanced datasets. The F1 score is denoted by the following equation:

$$F1 - score = 2 * (precision * recall)/(precision + recall)$$
$$precision = no.of\ true\ positives/(no.of\ true\ positives + no.of\ false\ positives)$$
$$recall = no.of\ true\ positives/(no.of\ true\ positives + no.of\ false\ negatives)$$

We used a Support Vector Machine (SVM) on our data, as that classifier is robust to changes in high variance (Boser et al., 1992), often unbalanced (Osuna et al., 1997), datasets.

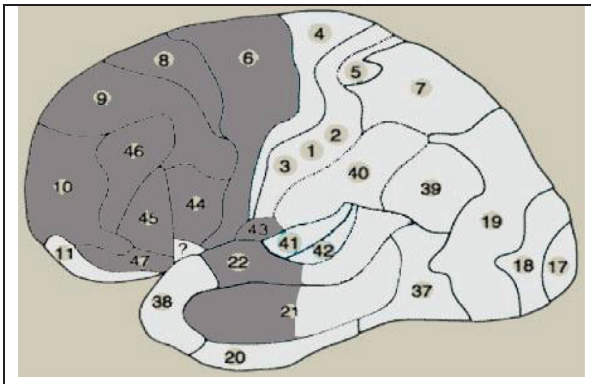
4.3. FNIRS data analysis and results

Data from three participants were removed from the analysis due to large motion artifacts throughout their datasets, with many channels reporting a value of 4.999, the default value used by the Hitachi-ETG when the source-detector channel has been oversaturated with light (Hitachi Medical Systems, 2016). Machine Learning was carried out on the remaining 17-subject dataset. We preprocessed each participant's raw light intensity data by first down sampling our data from 10 Hz to 2 Hz. Next, we used a band pass filter to remove noise from our data, saving the frequencies between 0.5 and 0.01 Hz. We then used the modified Beer-Lambert law to convert the resulting light intensity data into relative changes of oxy- and deoxy-hemoglobin. The data was then normalized in each channel using Z-score normalization.

4.4. Region of interest analysis and feature generation

Our preprocessed data from above included 52 channels of data, where each channel contained the rate of change in oxy- and deoxy-hemoglobin as measured at that location over time. We converted our 3d-digitzer data (which measured the positions of fNIRS optodes on the scalp in real-space) into MNI coordinates on the brain. Next, channels were averaged together into Regions of Interest (ROI) per Brodmann areas. The Brodmann areas covered by and accessible to the fNIRS channels are depicted in Fig. 7.

This resulted in 10 ROI's for analysis, where each ROI contained information about oxy- and deoxy-hemoglobin in that region. Next, for each ROI we computed several features of interest. These features were chosen because they have been successfully employed in prior machine learning research (Hirshfield et al., 2011) or because they were



Brodmann Region	Anatomical Regions Covered
6	Pre-Motor & Supplementary Motor Cortex
8	Frontal eye fields
9	Dorsolateral prefrontal cortex
10	Frontopolar cortex
21	Middle Temporal gyrus
22	Superior Temporal Gyrus
43	Subcentral area
45	pars triangularis Broca's area
46	Dorsolateral prefrontal cortex
47	Inferior prefrontal gyrus

Fig. 7. Brodmann regions covered by the fNIRS probes (21, 22 only partially covered).

employed by other researchers in recent fNIRS classification models (Karamzadeh et al., 2016). The features were generated for both the oxy and deoxy-hemoglobin time series data in the 10 ROIs noted above. We also generated the features separately for the (i) first half (ii) second half, and (iii) for the total of each 60 s task:

- ***Full-width-at-half* max** (*Karamzadeh et al., 2016*)*: The time difference between the two points where the signal is at half of its maximum value for the data from each 60-second-long video second.*
- ***Slope**: Slope calculated between the start and ending values of the signal.*
- ***Mean**: Average Signal Value.*
- ***Max**: Maximum Signal Value.*
- ***Min**: Minimum Signal Value.*

For each of the first half, second half, and total chunks of time series data noted above, we also further split that data into six equal segments of time and we took average values across those segments:

- ***Piecewise Mean 1**: Average Signal Value of Segment 1.*
- ***Piecewise Mean 2:** Average Signal Value of Segment 2.*
- ***Piecewise Mean 3:** Average Signal Value of Segment 3.*
- ***Piecewise Mean 4:** Average Signal Value of Segment 4.*
- ***Piecewise Mean 5:** Average Signal Value of Segment 5.*
- ***Piecewise Mean 6:** Average Signal Value of Segment 6.*

This resulted in: (10 ROIs * 2 types of data (oxy and deoxy) * 11 features (slope, min, max, etc.) * 3 time-segments (first, second half of task and total)] = **660 features to describe the brain activity during each 60-second-long video.**

4.5. Correlation tests on the fNIRS features vs the SAM survey labels

We were curious to see which brain regions were most highly correlated with the survey labels. Therefore, we took all participants' Average Oxygenated and Deoxygenated blood concentration data for each 60 s session and correlated it with the valence labels from the surveys for those sessions. Then we obtained the same for the arousal labels. The results of Pearson correlations are shown in Table 3, where positive correlations indicate a direct relationship between the relative change in oxy or deoxy- hemoglobin, and the survey label. Likewise, negative correlations indicate an inverse relationship between the relative change in oxy or deoxy-hemoglobin and the survey label. The *most significant correlations* (Different from 0 with a significance level alpha = 0.05) obtained from this test are marked by *.

It is notable that the DLPFC region had a high correlation with valence in both oxy and deoxy features. This is consistent with what Colibazzi et al. (2010) found in their fMRI study. Also, a larger number of deoxy features have significant correlations especially when it comes to valence.

4.6. Machine learning

Because some individual's datasets were imbalanced, we did not run a standard leave-one-participant-out cross validation. Instead, we grouped participants' data into four **folds**, while ensuring that each participants' data could never be split between train and test sets, as shown in Table 4 below. We ran a leave one-**fold** out cross validation to prevent any overfitting and biases that might affect the results when only using one person's data at a time for the test set. The grouping was decided simply by considering the order of participation in the study. After pulling out one-fold of data as the test set, the resulting training set was ranked using an information gain heuristic and the most predictive 15 features were selected for classification.

A Support Vector Machine (SVM) classifier was trained on these features and then tested on the participants that were initially left out as the test set. This was carried out for each of the folds. The average F1-scores achieved for each fold are shown in Table 4.

Since there was disagreement between the self-report labels and the original DEAP dataset labels as seen in Fig. 5, for a more complete comparison with the DEAP experiment, the same analysis was done using the DEAP labels instead of the self-report labels. The results of this analysis are shown in Table 5.

Composition of machine learning models: As noted previously, we created a feature vector with 660 features based on the fNIRS data acquired during each 60 s video. In this section, we describe the most predictive features and brain regions that contributed to our self-report label based SVM models' output. To demonstrate which brain regions were included in the feature selection process in our self-report valence and arousal models, the left side of Fig. 8 shows the frequency that each brain region was included as one of the top 15 features by the information gain heuristic employed by all the self-report valence models created during the leave-one-fold out cross validations. The same process was done for the self-report arousal models, with the right side of Fig. 8 showing the frequency that each brain region was included among the top 15 features created for each of the models created during the leave-one-fold out cross validations.

Looking beyond just the brain regions showing relevant emotional activation, we were also curious to better understand the type of features (see Section 4.4 for the type of features we generated) that were most predictive of the self-report valence and arousal class values across all our participants. To explore these features, we simply merged all participants' data and used the Weka 'Ranker Feature Selection' method to list the top 15 features using the information gain heuristic. We did this with all participants' data with the high and low self-report valence labels. We then repeated the process using all participants' data with their corresponding high and low self-report arousal labels. Summary data for these feature analyses is shown in Table 6. For the valence

Table 3
Pearson correlation between fNIRS data and the SAM survey labels for Valence and Arousal. A * is used to denote statistical significance.

Feature	Valence SAM label correlation with the average blood concentration data (Oxy and Deoxy)	Arousal SAM label correlation with the average blood concentration data (Oxy and Deoxy)
Premotor cortex average **oxy**	−0.0337	−0.0797
Frontal eye fields average **oxy**	−0.0672	0.0072
DLPFC average **oxy**	**−0.1247***	−0.0466
Frontopolar average **oxy**	−0.1014	−0.0335
Middle temporal Gyrus average **oxy**	−0.0748	−0.0376
Superior temporal Gyrus average **oxy**	−0.0485	−0.0098
Subcentral area average **oxy**	−0.1135	−0.0056
Broca's area average **oxy**	−0.0679	−0.0260
Inferior prefrontal Gyrus average **oxy**	−0.0805	0.0064
Premotor cortex average **Deoxy**	**−0.1227***	0.0124
Frontal eye eields average **Deoxy**	**−0.1905***	−0.0159
DLPFC average **Deoxy**	**−0.1251***	**−0.1446***
Frontopolar average **Deoxy**	−0.0854	−0.0398
Middle temporal Gyrus average **Deoxy**	−0.0796	−0.0399
Superior temporal Gyrus average **Deoxy**	−0.1177	−0.0002
Subcentral area average **Deoxy**	−0.0521	−0.0066
Broca's area average **Deoxy**	−0.0767	−0.0184
Inferior prefrontal Gyrus average **Deoxy**	−0.0241	−0.0479

Table 4
Average F1-Scores using the self-report survey labels for across subject classification for pairwise comparisons of high, neutral, and low valence, as well as high, neutral, and low arousal.

	Valence			Arousal		
	Low vs High	Neutral vs High	Low vs Neutral	Low vs High	Neutral vs High	Low vs Neutral
P1, P2, P3, P4	0.736	0.600	0.651	0.652	0.650	0.670
P5, P6, P7, P8	0.741	0.583	0.690	0.690	0.621	0.587
P9, P10, P12, P13	0.745	0.578	0.589	0.694	0.652	0.651
P14, P16, P17, P19, P20	0.737	0.733	0.691	0.590	0.630	0.700
Average F1-score with self-report labels	0.739	0.623	0.655	0.660	0.638	0.652

Table 5
Average F1-Scores using the DEAP dataset labels for across subject classification for pairwise comparisons of high, neutral, and low valence, as well as high, neutral, and low arousal.

	Valence			Arousal		
	Low vs High	Neutral vs High	Low vs Neutral	Low vs High	Neutral vs High	Low vs Neutral
P1, P2, P3, P4	0.670	0.520	0.560	0.667	0.680	0.561
P5, P6, P7, P8	0.660	0.703	0.441	0.650	0.710	0.730
P9, P10, P12, P13	0.682	0.523	0.600	0.670	0.732	0.634
P14, P16, P17, P19, P20	0.651	0.750	0.612	0.647	0.690	0.672
Average F1-score with DEAP labels	0.666	0.624	0.553	0.659	0.703	0.649

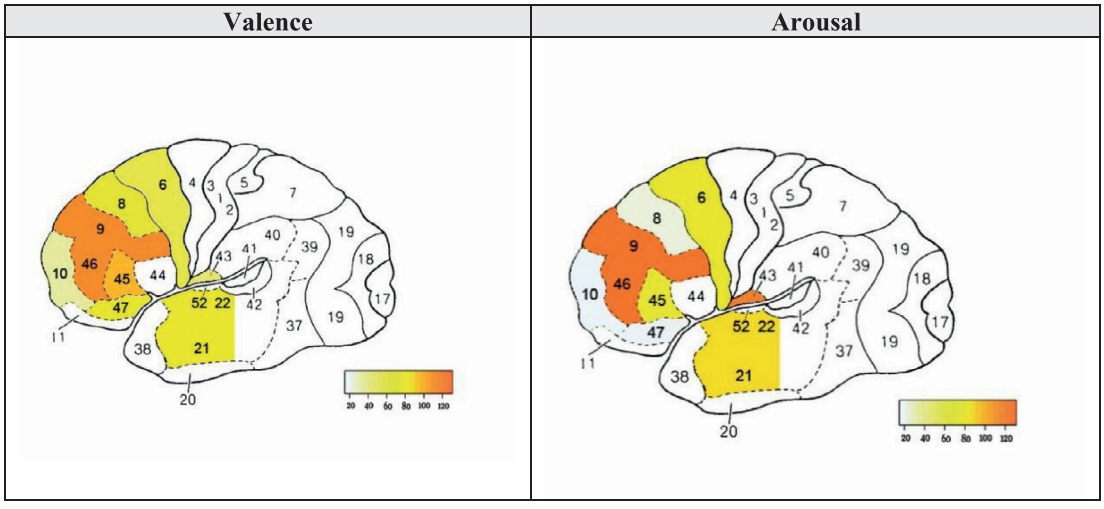


Fig. 8. Brain mapping of frequency counts of predictive Brodmann regions for valence and arousal classifications, across all participants.

Table 6
The most predictive features for self-report Valence and Arousal.

Top 15 Features for Predicting self-report Valence	Top 15 Features for Predicting self-report Arousal
• Deoxy Piecewise Mean 3 First Half (Frontal eye)	• Oxy Min Total (Premotor)
• Deoxy Piecewise Mean 3 First Half (Inferior Prefrontal gyrus)	• Deoxy Full Width at Half Max Second Half (DLPFC)
• Oxy Piecewise Mean 4 Second Half (DLPFC)	• Deoxy Max Second Half (Subcentral Area)
• Oxy Piecewise Mean 4 Second Half (Broca's Area)	• Oxy Piecewise Mean 3 First Half (Superior Temporal Gyrus)
• Deoxy Piecewise Mean 4 Total (Inferior Prefrontal gyrus)	• Deoxy Piecewise Mean 2 Total (DLPFC)
• Deoxy Min First Half (Inferior Prefrontal gyrus)	• Oxy Piecewise Mean 4 Second Half (Broca's Area)
• Deoxy Min Second Half (Premotor)	• Oxy Piecewise Mean 2 First Half (Premotor)
• Deoxy Piecewise Mean 4 Second Half (Frontopolar)	• Deoxy Max Total (Subcentral Area)
• Deoxy Piecewise Mean 5 Second Half (Subcentral area)	• Oxy Average First Half (Premotor)
• Oxy Min Total (Premotor)	• Oxy Piecewise Mean 3 Second Half (Inferior Prefrontal Gyrus)
• Oxy Piecewise Mean 4 Total (Broca's Area)	• Deoxy Average Second Half (Frontal Eye)
• Oxy Piecewise Mean 5 First Half (Premotor)	• Deoxy Full Width at Half Max Total (Frontal Eye)
• Oxy Piecewise Mean 5 Second Half (DLPFC)	• Deoxy Piecewise Mean 3 Total (Premotor)
• Oxy Slope Second Half (DLPFC)	• Deoxy Full Width at Half Max First Half (Frontal Eye)
• Oxy Piecewise Mean 5 First Half (Frontal Eye)	• Deoxy Piecewise Mean 3 Second Half (Middle Temporal gyrus)

comparisons, about half (8/15) of the selected features were based on oxy-hemoglobin, while the other half were based on deoxy-hemoglobin. For the arousal data, 2/3 of the top features were based on deoxy-hemoglobin. This is notable as many fNIRS studies only look at oxy-hemoglobin data, but in our analyses the deoxy-hemoglobin seems to have played an important role in the distinction of self-report valence and arousal classifications. It is not surprising to see the piecewise mean features listed often in the top 15 features, as they represented a large portion of the 660 features generated per instance (Section 4.4).

5. Discussion

As shown in Table 4, the self-report label based SVM model achieved average F1-scores of 0.739, 0.623, and 0.655 at distinguishing low vs high, neutral vs high, and low vs neutral valence, respectively. The arousal model achieved 0.66, 0.638, and 0.652 average F1-scores when predicting low vs high, neutral vs high, and low vs neutral arousal, respectively. Using the DEAP dataset labels (Table 5) achieved F1-scores of 0.666, 0.624 and 0.553 at distinguishing low vs high, neutral vs high, and low vs neutral valence, respectively. And 0.659, 0.703 and 0.649 average F1 scores when predicting low vs high, neutral vs high, and low vs neutral arousal, respectively. When comparing the use of self-report labels vs the DEAP dataset labels, the self-report label based model provided higher F1 scores than DEAP label based model in the case of comparing Low vs High Valence, Low vs Neutral Valence, Low vs High Arousal, and Low vs Neutral Arousal. Even though there is general consensus between the F1-score results from using DEAP labels and self-report labels, the self-report label based classification performed significantly better at classifying Low vs High Valence than the DEAP label based model. This could be due to individual differences in perception of emotion as discussed in Section 4.1.

These results are promising, especially since the classifiers were trained across participants, highlighting the potential for creating models based on large datasets of labeled participant data for training classifiers. Considering that Koelstra et al. (2012) reported an F1-score of 0.61 at classifying between low and high valence, our self-report F1-score of 0.739 for low and high valence distinctions suggests that the fNIRS may acquire unique information relating to the measurement of valence in the brain. The fNIRS results showed that the DLPFC and Broca's region were particularly useful at distinguishing between valence levels, which is in line with prior fMRI research on valence, especially in the context of listening to music. We posit that fNIRS' high spatial resolution enables the device to measure specific brain regions (such as the DLPFC and Broca's area) which are essential in the measurement of valence. When distinguishing between high and low arousal, our self-report F1-score of 0.66 is comparable to the 0.62 F1 score achieved with the EEG in Koelstra et al. (2012) study. These results suggest that fNIRS could be used to complement other physiological data to measure emotion, especially when distinguishing between levels of valence. It is worth noting that our classifier was strong at distinguishing between high and low levels of valence, but F1-scores were lower when the comparisons included a 'neutral' level of valence. This makes sense as neutral valence lies between high and low valence, and it is likely harder to distinguish. Our classifiers were built from the data from seventeen participants; it would be interesting to see the accuracies from classifiers trained on a dataset with double, or even triple, the number of participants.

The high density fNIRS used in this experiment (with 52 regions of the brain measured) allows for localization of brain activation. Fig. 8 showed the brain regions that were most predictive for the valence and arousal classification models, and these findings shed light on the neural correlates of valence and arousal. It is notable that features generated from deoxy-hemoglobin played an important role in the distinction of valence and arousal, and we urge the fNIRS research community to include deoxy-hemoglobin in their machine learning models to benefit from this information rich data. It appears that the model's prediction of valence relied heavily on Brodmann regions 9 and 46, which correspond to the dorsolateral prefrontal cortex (DLPFC). As noted in the literature review, the DLPFC region has been repeatedly linked to emotion regulation. The second most predictive region for valence was Broca's area (Brodmann region 45). Broca's is involved with processing of language, and it has been found to play a role in the processing of music lyrics, as would be the case in the music video stimuli. Perhaps participants engaged Broca's area more when they felt emotionally connected to a song and were fully engaged in lyric interpretation.

The models for predicting arousal level also relied heavily on the DLFPC, which makes sense, given the region's strong link to emotional experience. When comparing the regions of the brain most predictive of valence with those regions most predictive of arousal, it is interesting to note that prediction of valence seems to engage an interconnected network of brain regions, whereas the prediction of arousal does not require such a large region of brain real estate. For example, valence engaged the frontopolar region (BA10) and the Frontal Eye Fields (BA8), while the arousal models did not rely as heavily on those regions. The frontopolar region is heavily involved in processing of information, and it is interesting that this region has a stronger tie to the valence dimension than the arousal dimension of emotion.

The Frontal Eye Fields have been found to play a role in visual attention, and recent research from fMRI has suggested that emotionally charged visual information does influence activation in this region (Vuilleumier and Driver, 2007). Perhaps the music videos that were more emotionally charged on the valence dimension also drew more visual attention from participants as they watched the videos. It's also interesting to note that both valence and arousal engaged the pre-motor cortex (BA6), which has been linked not only to conscious planning of movement that will be executed, but to more subconscious thoughts about movements. As suggested by Warren et al. (2006), the emotion-

ally charged stimuli may have engaged participants' pre-motor cortex as their brains subconsciously prepared to make a facial gesture, such as producing a smile. This also dovetails with the previously noted activation in the DLPFC, as participants would have engaged that region while regulating their reaction to the emotional content.

These results are encouraging, and they rely upon localization of functional brain regions, which, before the introduction of fNIRS, could only be done with fMRI scanners. Thus, we claim that fNIRS is a strong choice for non-invasive brain measurement of emotional states, when there is a need for fMRI-quality measurements made under normal working conditions. Of course, fNIRS is limited by the fact that it has low temporal resolution and it does not measure the entire brain, making it unable to directly measure deep brain regions like the amygdala, which are heavily involved in emotional processing. We do not propose that fNIRS is the only modality for this type of measurement, but that fNIRS can be incorporated into experiments using EEG and other multi-modal sensors for measurement of emotion. It would be interesting to run an experiment with multiple sensors such as fNIRS, EEG, ECG, and GSR. Each sensor modality provides a different physiological measure, and when combined, they may provide enough pieces to the puzzle that is emotion; to get a full and accurate picture of one's emotional state. Particularly, the combination of the EEG and fNIRS (Salvatori et al., 2006; Savran et al., 2006; Roche-Labarbe et al., 2008; Sun et al., 2015), can take advantage of the best aspects of each system. For instance, a combined system benefits from both the high temporal resolution of the EEG and the high spatial resolution of fNIRS. This would enable researchers to pinpoint which parts of the brain are activated by a task while also measuring quick changes in neural activity.

6. Conclusion and future work

In this paper, we demonstrated the capability of classifying and distinguishing between affective states on the valence and arousal dimensions using fNIRS, a practical non-invasive device that can localize activation in functional brain regions with spatial resolution comparable to fMRI and superior to EEG, the most common non-invasive brain measurement modality. Our fNIRS results show that specific functional brain regions are recruited during changes in valence and arousal, and these regions are in line with prior fMRI research on emotion. We use our fNIRS data to build models to make predictions across subjects, rather than training each model individually per participant, which limits the data available for model training and testing, and can lead to overfitting. Developing accurate across-subject machine learning models is necessary to build the large datasets of data necessary for training robust classifiers that can be applied across different participants and task types.

Our stimulus materials and protocol build off prior research by Koelstra et al. (2012) where music videos were used as stimulus material for inducing emotional states. They found that fusing data from multiple sensors (EEG and other physiological measures) improved classification accuracy. We propose fNIRS as an additional measurement modality to further improve the predictive accuracy of emotion models. Future work should run this protocol while participants wear fNIRS, EEG, and other physiological sensors, to determine the improvement in accuracy achieved with the addition of the fNIRS modality. This line of future work would complement the growing body of literature using fNIRS in the measurement of cognitive states.

Acknowledgements

This work has been funded in part by National Science Foundation (NSF) under CAREER grant CNS-1206291 and NSF Grant CNS-1302559. We would also like to thank the Air Force Office of Sponsored Research (award FA9550-15-1-0021) for their support of this research.

References

Bailey, R., Wise, K., Bolls, P., 2009. How avatar customizability affects children's arousal and subjective presence during junk food–sponsored online video games. Cyber Psychol. Behav. 12 (3), 277–283.
Ball, G., Breese, J., 1999. Modeling the emotional state of computer users. In: Proceedings of the Workshop on Personality and Emotion in User Modelling.
Baumgartner, T., Esslen, M., Jäncke, L., 2006. From emotion perception to emotion experience: emotions evoked by pictures and classical music. Int. J. Psychophysiol. 60 (1), 34–43.
Berka, C., Levendowski, D.J., Lumicao, M.N., Yau, A., Davis, G., Zivkovic, V.T., Craven, P.L., 2007. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. Aviat. Space Environ. Med. 78 (5), B231–B244.
Bendall, R.C., Eachus, P., Thompson, C., 2016. A brief review of research using near-infrared spectroscopy to measure activation of the prefrontal cortex during emotional processing: the importance of experimental design. Front. Human Neurosci. 10.
Boser, B.E., Guyon, I.M., Vapnik, V.N., 1992. A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. ACM, pp. 144–152.
Bosma, W., André, E., 2004, January. Exploiting emotions to disambiguate dialogue acts. In: Proceedings of the 9th International Conference on Intelligent User Interfaces. ACM, pp. 85–92.
Bradley, M.M., Lang, P.J., 1994. Measuring emotion: the self-assessment manikin and the semantic differential. J. Behav. Ther. Exp. Psychiatry 25 (1), 49–59.
Bradley, M., Lang, P.J., 1999. The International Affective Digitized Sounds (IADS): Stimuli, Instruction Manual and Affective Ratings. NIMH Center for the Study of Emotion and Attention.
Brattico, E., Alluri, V., Bogert, B., Jacobsen, T., Vartiainen, N., Nieminen, S.K., Tervaniemi, M., 2011. A functional MRI study of happy and sad emotions in music with and without lyrics. Front. Psychol. 2, 308.
Brown, S.L., Schwartz, G.E., 1980. Relationships between facial electromyography and subjective experience during affective imagery. Biol. Psychol. 11 (1), 49–62.
Cacioppo, J.T., Gardner, W.L., Berntson, G.G., 1999. The affect system has parallel and integrative processing components: form follows function. J. Pers. Social Psychol. 76 (5), 839.
Cannon, W.B., 1927. The James-Lange theory of emotions: a critical examination and an alternative theory. Am. J. Psychol. 39 (1/4), 106–124.
Chance, B., Zhuang, Z., UnAh, C., Alter, C., Lipton, L., 1993. Cognition-activated low-frequency modulation of light absorption in human brain. Proc. Natl. Acad. Sci. 90 (8), 3770–3774.
Churchill, N.W., Yourganov, G., Strother, S.C., 2014. Comparing within-subject classification and regularization methods in fMRI for large and small sample sizes. Human Brain Mapp. 35 (9), 4499–4517.
Codispoti, M., Surcinelli, P., Baldaro, B., 2008. Watching emotional movies: affective reactions and gender differences. Int. J. Psychophysiol. 69 (2), 90–95.
Colibazzi, T., Posner, J., Wang, Z., Gorman, D., Gerber, A., Yu, S., Peterson, B.S., 2010. Neural systems subserving valence and arousal during the experience of induced emotions. Emotion 10 (3), 377.
Crosson, B., Ford, A., McGregor, K.M., Meinzer, M., Cheshkov, S., Li, X., Briggs, R.W., 2010. Functional imaging and related techniques: an introduction for rehabilitation researchers. J. Rehabil. Res. Dev. 47 (2) vii.
Cui, X., Bray, S., Bryant, D.M., Glover, G.H., Reiss, A.L., 2011. A quantitative comparison of NIRS and fMRI across multiple cognitive tasks. Neuroimage 54 (4), 2808–2821.
Darwin, C., 1965. The Expression of the Emotions in Man and Animals, vol. 526. University of Chicago press.
Davidson, R.J., Jackson, D.C., Kalin, N.H., 2000. Emotion, plasticity, context, and regulation: perspectives from affective neuroscience. Psychol. Bull. 126 (6), 890.
Dimberg, U., 1988. Facial electromyography and the experience of emotion. J. Psychophysiol.
Doi, H., Nishitani, S., Shinohara, K., 2013. NIRS as a tool for assaying emotional function in the prefrontal cortex. Front. Human Neurosci. 7.
Ehring, T., Tuschen-Caffier, B., Schnülle, J., Fischer, S., Gross, J.J., 2010. Emotion regulation and vulnerability to depression: spontaneous versus instructed use of emotion suppression and reappraisal. Emotion 10 (4), 563.
Ekman, P., Friesen, W.V., 1971. Constants across cultures in the face and emotion. J. Pers. Social Psychol. 17 (2), 124.
Ekman, P., Levenson, R.W., Friesen, W.V., 1983. Autonomic nervous system activity distinguishes among emotions. Science 221 (4616), 1208–1210.
Fadiga, L., Craighero, L., D'Ausilio, A., 2009. Broca's area in language, action, and music. Ann. New York Acad. Sci. 1169 (1), 448–458.
Feldman, B.L., Russell, J.A., 1998. Independence and bipolarity in the structure of current affect. J. Pers. Social Psychol. 74 (4), 967.
Fridlund, A.J., Schwartz, G.E., Fowler, S.C., 1984. Pattern recognition of self-reported emotional state from multiple-site facial EMG activity during affective imagery. Psychophysiology 21 (6), 622–637.
Girouard, A., Solovey, E., Hirshfield, L., Chauncey, K., Sassaroli, A., Fantini, S., Jacob, R., 2009. Distinguishing difficulty levels with non-invasive brain activity measurements. In: Human-Computer Interaction–INTERACT 2009, pp. 440–452.
Glotzbach, E., Mühlberger, A., Gschwendtner, K., Fallgatter, A.J., Pauli, P., Herrmann, M.J., 2011. Prefrontal brain activation during emotional processing: a functional near infrared spectroscopy study (fNIRS). Open Neuroimag. J. 5, 33.
Grimes, D., Tan, D.S., Hudson, S.E., Shenoy, P., Rao, R.P., 2008, April. Feasibility and pragmatics of classifying working memory load with an electroencephalograph. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 835–844.

Heilman, R.M., Crişan, L.G., Houser, D., Miclea, M., Miu, A.C., 2010. Emotion regulation and decision making under risk and uncertainty. Emotion 10 (2), 257.
Hirshfield, L.M., Bobko, P., Barelka, A., Hirshfield, S.H., Farrington, M.T., Gulbronson, S., Paverman, D., 2014. Using noninvasive brain measurement to explore the psychological effects of computer malfunctions on users during human-computer interactions. Adv. Hum. Comput. Interact. 2014, 2.
Hirshfield, L.M., Chauncey, K., Gulotta, R., Girouard, A., Solovey, E., Jacob, R., Fantini, S., 2009. Combining electroencephalograph and functional near infrared spectroscopy to explore users' mental workload. In: Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience, pp. 239–247.
Hirshfield, L.M., Costa, M., Bandara, D., Bratt, S., 2015, August. Measuring situational awareness aptitude using functional near-infrared spectroscopy. In: International Conference on Augmented Cognition. Springer International Publishing, pp. 244–255.
Hirshfield, L.M., Girouard, A., Solovey, E.T., Jacob, R.J.K., Sassaroli, A., Tong, Y., Fantini, S., 2007. Human-computer interaction and brain measurement using functional near-infrared spectroscopy. In: Proceedings of the ACM UIST'07 Symposium on User Interface Software and Technology. ACM Press.
Hirshfield, L.M., Gulotta, R., Hirshfield, S., Hincks, S., Russell, M., Ward, R., Jacob, R., 2011 May. This is your brain on interfaces: enhancing usability testing with functional near-infrared spectroscopy. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 373–382.
Hitachi, Ltd, 2016. ETG-4000: Options Retrieved May 20, 2017, from http://www.hitachi.com/businesses/healthcare/products-support/opt/etg4000/contents1.html.
Hoshi, Y., Huang, J., Kohri, S., Iguchi, Y., Naya, M., Okamoto, T., Ono, S., 2011. Recognition of human emotions from cerebral blood flow changes in the frontal region: a study with event-related near-infrared spectroscopy. J. Neuroimag. 21 (2).
Izzetoglu, K., Bunce, S., Onaral, B., Pourrezaei, K., Chance, B., 2004. Functional optical brain imaging using near-infrared during cognitive tasks. Int. J. Hum. Comput. Interact. 17 (2), 211–227.
Karamzadeh, N., Amyot, F., Kenney, K., Anderson, A., Chowdhry, F., Dashtestani, H., Diaz-Arrastia, R., 2016. A machine learning approach to identify functional biomarkers in human prefrontal cortex for individuals with traumatic brain injury using functional near-infrared spectroscopy. Brain Behav. 6 (11).
Kober, H., Barrett, L.F., Joseph, J., Bliss-Moreau, E., Lindquist, K., Wager, T.D., 2008. Functional grouping and cortical–subcortical interactions in emotion: a meta-analysis of neuroimaging studies. Neuroimage 42 (2), 998–1031.
Koelstra, S., Muhl, C., Soleymani, M., Lee, J.S., Yazdani, A., Ebrahimi, T., Patras, I., 2012. Deap: a database for emotion analysis; using physiological signals. IEEE Trans. Affect. Comput. 3 (1), 18–31.
Kring, A.M., Barrett, L.F., Gard, D.E., 2003. On the broad applicability of the affective circumplex: representations of affective knowledge among schizophrenia patients. Psychol. Sci. 14 (3), 207–214.
Kuppens, P., Tong, E., 2010. An appraisal account of individual differences in emotional experience. Soc. Person. Psychol. Compass 4 (12), 1138–1150.
Lane, R.D., Nadel, L. (Eds.), 1999, Cognitive Neuroscience of Emotion. Oxford University Press.
Lang, P.J., Bradley, M.M., Cuthbert, B.N., 1999. International affective picture system (IAPS): instruction manual and affective ratings. The Center for Research in Psychophysiology. University of Florida.
Lang, P.J., Greenwald, M.K., Bradley, M.M., Hamm, A.O., 1993. Looking at pictures: affective, facial, visceral, and behavioral reactions. Psychophysiology 30 (3), 261–273.
Lee, J.C., Tan, D.S., 2006 October. Using a low-cost electroencephalograph for task classification in HCI research. In: Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology. ACM, pp. 81–90.
León-Carrión, J., Martín-Rodríguez, J.F., Damas-López, J., Pourrezai, K., Izzetoglu, K., y Martin, J.M.B., Domínguez-Morales, M.R., 2007. A lasting post-stimulus activation on dorsolateral prefrontal cortex is produced when processing valence and arousal in visual affective stimuli. Neurosci. Lett. 422 (3), 147–152.
Massey, D.S., 2002. A brief history of human society: the origin and role of emotion in social life. Am. Sociol. Rev. 67 (1), 1.
Ortony, A., Clore, G.L., Collins, A., 1990. The Cognitive Structure of Emotions. Cambridge university press.
Osuna, E., Freund, R., Girosi, F., 1997. Support Vector Machines: Training and Applications Technical Report AIM-1602, 1997.
Parasuraman, R., Rizzo, M. (Eds.), 2008, Neuroergonomics: The Brain at Work. Oxford University Press.
Picard, R.W., Picard, R., 1997. Affective Computing, vol. 252. MIT Press, Cambridge.
Plutchik, R., 2001. The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. Am. Sci. 89 (4), 344–350.
Plutchik, R., 1980. A general psychoevolutionary theory of emotion. Theor. Emotion 1 (3-31), 4.
Posner, J., Russell, J.A., Peterson, B.S., 2005. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology. Dev. Psychopathol. 17 (03), 715–734.
Pressman, S.D., Cohen, S., 2005. Does positive affect influence health? Psychol. Bull. 131 (6), 925.
Roche-Labarbe, N., Zaaimi, B., Berquin, P., Nehlig, A., Grebe, R., Wallois, F., 2008. NIRS-measured oxy-and deoxyhemoglobin changes associated with EEG spike-and-wave discharges in children. Epilepsia 49 (11), 1871–1880.
Rodrigo, A.H., Ayaz, H., Ruocco, A.C., 2016 July. Examining the neural correlates of incidental facial emotion encoding within the prefrontal cortex using functional near-infrared spectroscopy. In: International Conference on Augmented Cognition. Springer International Publishing, pp. 102–112.
Russell, J.A., 1980. A circumplex model of affect. J. Pers. Social Psychol. 39 (6), 1161–1178.
Russell, J.A., 1991. Culture and the categorization of emotions. Psychol. Bull. 110 (3), 426.
Salovey, P., Rothman, A.J., Detweiler, J.B., Steward, W.T., 2000. Emotional states and physical health. Am. Psychol. 55 (1), 110.
Salvatori, G., Bulf, L., Fonda, S., Rovati, L., 2006, April. Combining near-infrared spectroscopy and electroencephalography to monitor brain function. In: Instrumentation and Measurement Technology Conference, IMTC 2006. Proceedings of the IEEE. IEEE, pp. 32–36.
Savran, A., Ciftci, K., Chanel, G., Mota, J., Viet, L., Sankur, B., Akarun, L., Caplier, A., Rombaut, M., 2006. Emotion detection in the loop from brain signals and facial images. In: Proceeding of the eNTERFACE 2006.
Schachter, S., Singer, J., 1962. Cognitive, social, and physiological determinants of emotional state. Psychol. Rev. 69 (5), 379.
Scherer, K.R., 2005. What are emotions? And how can they be measured? Soc. Sci. Inf. 44 (4), 695–729.
Schwartz, G.E., Fair, P.L., Salt, P., Mandel, M.R., Klerman, G.L., 1976. Facial muscle patterning to affective imagery in depressed and nondepressed subjects. Science 192 (4238), 489–491.
Scollon, C.N., Koh, S., Au, E.W., 2011. Cultural differences in the subjective experience of emotion: when and why they occur. Soc. Personal. Psychol. Compass 5 (11), 853–864.
Son, I.Y., Guhe, M., Gray, W.D., Yazici, B., Schoelles, M.J., 2005 May. Human performance assessment using fNIR. In: Defense and Security. International Society for Optics and Photonics, pp. 158–169.
Sun, Y., Ayaz, H., Akansu, A.N., 2015 December. Neural correlates of affective context in facial expression analysis: a simultaneous EEG-fNIRS study. In: Signal and Information Processing (GlobalSIP), 2015 IEEE Global Conference on. IEEE, pp. 820–824.
Viinikainen, M., Jääskeläinen, I.P., Alexandrov, Y., Balk, M.H., Autti, T., Sams, M., 2010. Nonlinear relationship between emotional valence and brain activity: evidence of separate negative and positive valence dimensions. Hum. Brain Mapp. 31 (7), 1030–1040.
Vuilleumier, P., Driver, J., 2007. Modulation of visual processing by attention and emotion: windows on causal interactions between human brain regions. Philosoph. Trans. R. Soc. Lond. B 362 (1481), 837–855.
Warren, J.E., Sauter, D.A., Eisner, F., Wiland, J., Dresner, M.A., Wise, R.J., Scott, S.K., 2006. Positive emotions preferentially engage an auditory–motor "mirror" system. J. Neurosci. 26 (50), 13067–13075.
Watson, D., Clark, L.A., Tellegen, A., 1988. Development and validation of brief measures of positive and negative affect: the PANAS scales. J. Pers. Social Psychol. 54 (6), 1063.