UNIVERSITY OF CALIFORNIA SAN DIEGO

Utilizing Multi-modal Bio-sensing Toward Affective Computing in Real-world Scenarios

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering
(Intelligent Systems, Robotics, and Control)

by

Siddharth Siddharth

Committee in charge:

Professor Mohan M. Trivedi, Chair
Professor Tzyy-Ping Jung, Co-Chair
Professor Vikash Gilja
Professor Patrick P. Mercier
Professor Terrence J. Sejnowski

2020

The dissertation of Siddharth Siddharth is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Co-Chair

Chair

University of California San Diego

2020

DEDICATION

To Daadi and Babaji—my teachers, friends, babies.

EPIGRAPH

Prajnanam Brahma

—Rig Veda

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

As I write this thesis dissertation while being homebound due to the COVID-19 pandemic, I am perhaps prone to reflect more than usual on the time I have spent as a college student. Like every journey, this too has come to an end. I have always believed that the aim of traveling is not to reach the destination but rather to enjoy the journey itself. I had wanted to pursue a Ph.D. since I commenced my undergraduate studies in 2011 and what a joyous journey this one has been.

It is generally foolhardy to write succinctly about so many people and institutions that have nurtured an ignoramus like me but then not doing so would be a gross injustice of massive proportions.

A favorite topic of discussion between doctoral candidates is their uneasy relationship with their advisers. I have witnessed this often in the university as well as during internships in the industry where many friends contemplated quitting their graduate programs due to disagreements with their respective advisers. I fondly recount now that I have always been the odd one out in such discussions since my academic marriage with Professor Tzyy-Ping Jung has been a very successful one. I am yet to meet a more generous, cheerful, and caring human being. The same goes for my relationship with Professor Terry Sejnowski and Professor Mohan Trivedi who were like an extended family. I have been very fortunate to be closely advised by three faculty members and have learned a lot from them about leadership and mentoring. I feel quite impatient to utilize this learning in my future endeavors. Professor Vikash Gilja and Professor Patrick Mercier provided insightful comments on my research plan from time to time and I am grateful for their feedback.

My research family at the Swartz Center for Computational Neuroscience (SCCN) has given me so much that I am afraid I will never be able to pay back even a small portion of the same. Daily tea time discussions and evening football games were the key sources of my energy. Scott, John, Rhonda, Ramón, Johanna, Makoto, Young, Masaki, Shawn, Chi-Yuan, Bob, Yu-Te, Poyuan, Clement, Nicole, and so many others at SCCN were ever helpful and stimulants of my cheerfulness. I have no words to thank my dear friend, Roger, who despite his busy schedule and popularity played multiple roles for me—mentor, editor, debater, storyteller, and much more. Other collaborators in different labs——Julia, Gedeon, Eduardo, Jaime, Yelena, Akshay, Nachiket—also invariably supported me in ways more than professional. I am also grateful to my extended family at the Prospect Journal of International Affairs and South Asia Initiative who adopted me and nurtured me in other research directions.

Aashish, whether it was converting my office to Tony Stark's workshop, jogging sessions, sunrise (rarely) and sunset (often) walk toward the ocean, or afternoon strolls to the Salk Institute to fetch lunch, San Diego would not have been the same without you. I am sure we will continue to abbreviate our future project names to match those of exotic Australian animals (wombat, emu, koala). Thank you for teaching me so much including the fine art of slowly sipping coffee. Sabareesh, I would always look forward to continuing our conversations about everything from Politics to History to Economics in the years to come. It was great to have you as a friend with so many similar interests.

The members of my family from undergrad—Shubham, Tyagi, Tahir, Ishan, Akash, Kaushal—were ever supportive while it would be entirely foolish to undertake the impossible task of writing what Nikhil has meant to me since we first met in 2011. I am also heavily indebted and thankful to many other friends—Raavi, Sushil, Ben, and Kirti—who extended constant support to my endeavors. My family at Facebook—James, Ruta, Kartheek, Diar, Arash, Luca, Camille, Pedro—was also a constant source of motivation (dare I say fun?) during the two summers I spent there.

This journey would not have been joyous, indeed realizable, without my family elders. Vibha Bua, my chief protector, intimate adviser, and unflinching scolder, always made me feel at home in America. Manoj Chachu, Viji Uncle, Arj Uncle, Abha Bua, Bittu Chachu, Mukta, Seema Mausi, Ashok Uncle, always warmly supported me in every way possible. Ironically, my little sister, Princy, instead backed me by playing the role of a strict teacher. I would also like to profusely thank many institutions in India—academic, governmental, and financial—without whose nurture and support, I would not have reached here in the first place. This brings me to the last and the single most important institution—my parents—without whose love and sacrifices, nothing would have been possible. Mummy and Papa, without your endearing and careful watch, where would I be?

Chapter 3, in part, contains materials from “A Wearable Multi-modal Bio-sensing System Towards Real-world Applications” by Siddharth Siddharth, Aashish N. Patel, Tzyy-Ping Jung, and Terrence J. Sejnowski, which was published in the IEEE Transactions on Biomedical Engineering, 2018. The dissertation author was the first author of this paper.

Chapter 4, in part, contains materials from “Impact of Affective Multimedia Content on the Electroencephalogram and Facial Expressions” by Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski, which was published in Scientific Reports, 2019. The dissertation author was the first author of this paper.

Chapter 5, in part, contains materials from “Utilizing Deep Learning Towards Multi-modal Bio-sensing and Vision-based Affective Computing” by Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski, which was published in the IEEE Transactions on Affective Computing, 2019. The dissertation author was the first author of this paper.

Chapter 6, in part, contains materials from “On Assessing Driver Awareness of Situational Criticalities: Multi-modal Bio-Sensing and Vision-Based Analysis, Evaluations, and Insights” by Siddharth Siddharth and Mohan M. Trivedi, which was published in Brain Sciences, 2020. The dissertation author was the first author of this paper.

VITA

2011-2015 B. Tech in Electronics and Communications Engineering, Indian Institute of Information Technology, Allahabad

2015-2017 M. S. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego

2017-2020 Ph. D. in Electrical Engineering (Intelligent Systems, Robotics, and Control), University of California San Diego

PUBLICATIONS

Journal Articles

Siddharth Siddharth and Mohan M. Trivedi "On Assessing Driver Awareness of Situational Criticalities: Multi-modal Bio-Sensing and Vision-Based Analysis, Evaluations, and Insights", *Brain Sciences*, 10, 2020.

Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski "Impact of Affective Multimedia Content on the Electroencephalogram and Facial Expressions", *Scientific Reports*, 9, 2019.

Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski "Utilizing Deep Learning Towards Multi-modal Bio-sensing and Vision-based Affective Computing", *IEEE Transactions on Affective Computing*, 2019.

Siddharth Siddharth, Aashish N. Patel, Tzyy-Ping Jung, and Terrence J. Sejnowski, "A Wearable Multi-modal Bio-sensing System Towards Real-world Applications", *IEEE Transactions on Biomedical Engineering*, 66, 2018.

Conference Articles

Siddharth Siddharth and Mohan M. Trivedi, "Attention Monitoring and Hazard Assessment with Bio-Sensing and Vision: Empirical Analysis Utilizing CNNs on the KITTI Dataset", *IEEE Intelligent Vehicles Symposium (IV)*, pp. 1673-1678, 2019.

Julia Anna Adrian, Siddharth Siddharth, Syed Zain Ali Baquar, Tzyy-Ping Jung, Gedeon Deák, "Decision-Making in a Social Multi-Armed Bandit Task: Behavior, Electrophysiology and Pupillometry", *41st Annual Meeting of the Cognitive Science Society (CogSci)*, 2019.

Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski "Multi-modal Approach for Affective Computing", *IEEE 40th International Engineering in Medicine and Biology Conference (EMBC)*, pp. 291-294, 2018.

Siddharth Siddharth, Aashish N. Patel, Tzyy-Ping Jung, and Terrence J. Sejnowski, “An Affordable Bio-sensing and Activity Tagging Platform for HCI Research” *In International Conference on Augmented Cognition*, pp. 399-409. Springer, Cham, 2017.

Siddharth Siddharth, Akshay Rangesh, Eshed Ohn-Bar, and Mohan Trivedi, “Driver Hand Localization and Grasp Analysis: A Vision-based Real-time Approach”, *IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2545-2550, 2016.

Patents

Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski. “Bio-sensing and eye-tracking system.” U.S. Patent Application No. 16/325,567.

Siddharth Siddharth, Aashish Patel, Tzyy-Ping Jung, and Terrence J. Sejnowski, “Wearable multi-modal bio-sensing system,” U.S. Provisional Patent Application No. 62/656,890.

ABSTRACT OF THE DISSERTATION

Utilizing Multi-modal Bio-sensing Toward Affective Computing in Real-world Scenarios

by

Siddharth Siddharth

Doctor of Philosophy in Electrical Engineering
(Intelligent Systems, Robotics, and Control)

University of California San Diego, 2020

Professor Mohan M. Trivedi, Chair
Professor Tzyy-Ping Jung, Co-Chair

Recognition and continuous monitoring of human emotions is a key problem that is spread across multiple research disciplines like electrical engineering, computer science, neuroscience, and cognitive science. Instead of the widely used method of approaching this problem from the perspective of computer vision i.e. by tracking user's facial expressions, we employ a multi-modal approach. This approach utilizes various bio-sensing modalities in addition to computer vision for assessing and classifying human affects. In the process, we address various limitations—hardware and software—of such bio-sensing modalities. We evaluate the developed framework on real-world applications ranging from watching emotional multimedia content (such as videos) to driving an automobile. Our holistic framework contributes toward (1) the development of a novel multi-modal bio-sensing headset that is capable of recording, monitoring, and tracking various bio-signals in real-time, (2) the development of algorithms that utilize signal processing as well as deep learning tools for various bio-sensor and vision-based modalities for emotion recognition, (3) the utilization of such algorithms in a wide range of applications namely emotion classification, studying emotion elicitation, and monitoring driver's awareness, and (4) the performance comparison of various sensor modalities for the above applications when they are used individually or in fusion with each other. Although the above multi-modal sensor platform has been evaluated in this thesis dissertation on affective computing applications only, it is generalizable. Thus, this platform can work for various applications in the field of human-computer interaction that require the use of bio-sensing modalities.

Chapter 1

Outline and Contributions

This doctoral dissertation's overarching aim is the development of multi-modal bio-sensing hardware systems and algorithmic platform for emotion recognition. This is especially geared toward two real-world applications: watching affective multimedia videos and driving an automobile. Chapter 2 presents the introduction to this doctoral dissertation. It also details the organization of the dissertation. Chapter 3 details the development and evaluation of a multi-modal bio-sensing headset capable of detecting, tracking, and recording various bio-markers and physiological states. It also presents the evaluation of each modality as well as that of using multiple modalities simultaneously while being engaged in two real-world gaming tasks. Chapter 4 utilizes multiple sensor modalities to study the elicitation of human emotions by assessing the impacts of affective multimedia contents on users' electroencephalogram (EEG) and facial expressions. Subsequently, Chapters 5 and 6 present applications of the tools from the above chapters in emotion recognition and driver awareness contexts. Specifically, Chapter 5 utilizes sensors developed in Chapter 3 for the development of various software algorithms capable of performing emotion classification. These algorithms are evaluated on four publicly available emotion datasets in which participants watched various emotional videos. Chapter 6 utilizes the hardware and software frameworks designed in the previous chapters toward applications in driver awareness monitoring. It details the classification of a driver's mental states based on his/her attentiveness and responses to hazardous on-road situations. Finally, Chapter 7 concludes this dissertation by commenting on the importance of the research findings from Chapters 3-6.

With the ultimate goal of developing a multi-modal bio-sensing headset capable of monitoring and tracking human emotions in real-time continuously throughout the day, we present the following contributions:

- We present the development of a novel multi-modal bio-sensing headset that is capable of recording, monitoring, and tracking various bio-signals in real-time. This headset can synchronize such signals in real-time and can be used for various real-world applications.

- We develop algorithms that utilize signal processing as well as deep learning tools to integrate various bio-sensor and vision-based modalities for emotion recognition. These algorithms are developed in a robust, scalable, and modular manner, and thus can be used either individually or together.

- We utilize the above algorithms in a wide range of applications namely emotion classification, studying emotion elicitation, and monitoring driver's awareness. To perform the studies, we also collect novel multi-modal datasets.

- We analyze and compare the performance of various sensor modalities for the above applications when they are used individually or in fusion with each other. This analysis aimed at understanding which sensor modalities or their combinations could prove to be most useful in such applications.

- We introduce a hardware and software platform capable of studying human affective states through multiple sensor modalities capable of being fabricated into a compact and wearable form factor. Subsequently, this research is a step toward the development of smart devices capable of continuously monitoring the user's emotional and physiological health.

Chapter 2

Introduction

Imagine that you have just returned home from a long and exhausting day at work. You are feeling tired and low in spirit. However, you are wearing a smart band in your wrist, a "Fitbit for detecting emotions" if you will, which sends a signal to your home theater music system to play your favorite song. Your spirits are enlivened, you feel much better, and you start relaxing to the tune of the musical notes. This doctoral thesis is a step in the design, fabrication, and evaluation of the development of such a smart and compact device capable of detecting, classifying, and monitoring human emotions.

Probably as long as humans have been self-aware, they have wondered about the origin, essence, and utility of emotions. Though the word *emotion* was only coined (adapted from the French word *émouvoir* meaning to "stir up") in the early 1800s, philosophers across the world have discussed the concept for millennia. In the West, through Plato's writings, Socrates comes out as a cognitivist who believed that feelings like fear or dread are constituted by cognitive states [1]. Western philosophers, in general, held the view about making the distinction between cognition and emotion i.e. the destructive role that emotions can play in decision-making. On the other hand, in the East, there is a more subtle perception of emotions. While Hinduism refers to *bhāva* (feeling) as illusory and transitory, Buddhism propagates universal compassion and empathy hand-in-hand with a feeling of abnegation since emotions prevent liberation [2].

How emotions are generated has long been the cause of discussion between philosophers, cognitive scientists, neuroscientists, psychologists, psychiatrists, and social scientists. Tens of hundreds of studies have been performed by researchers to study the onset of emotions. These studies are now characterized under the banner of Affective Computing. Affective computing has emerged as an interdisciplinary field spanning computer science, psychology, and cognitive science to study and develop systems that can detect, emulate, simulate, process, and interpret emotions [3].

For centuries, two methods have been predominantly utilized to assess somebody's emotional health. The first method has been to ask for feedback from the person directly or indirectly through a questionnaire. The second method has been to observe their bodily cues such as facial expressions and posture. While the first method is highly subjective and is prone to error because of the person's mood, the second method is indirect since it only monitors the products of the cognitive processes generating human emotions and is prone to deliberate misrepresentation. Physiological cues such as a change in heart-beats and sweating that are the immediate and involuntary processes accompanying human emotions have been difficult to measure throughout human history. Fortunately, in recent years, for the first time in history, with the development of bio-sensors capable of monitoring, recording, and tracking a person's physiology in real-time and in a real-world manner, it has become possible to detect and classify human emotions directly from the bodily signals tied with the cognitive processes generating human emotions.

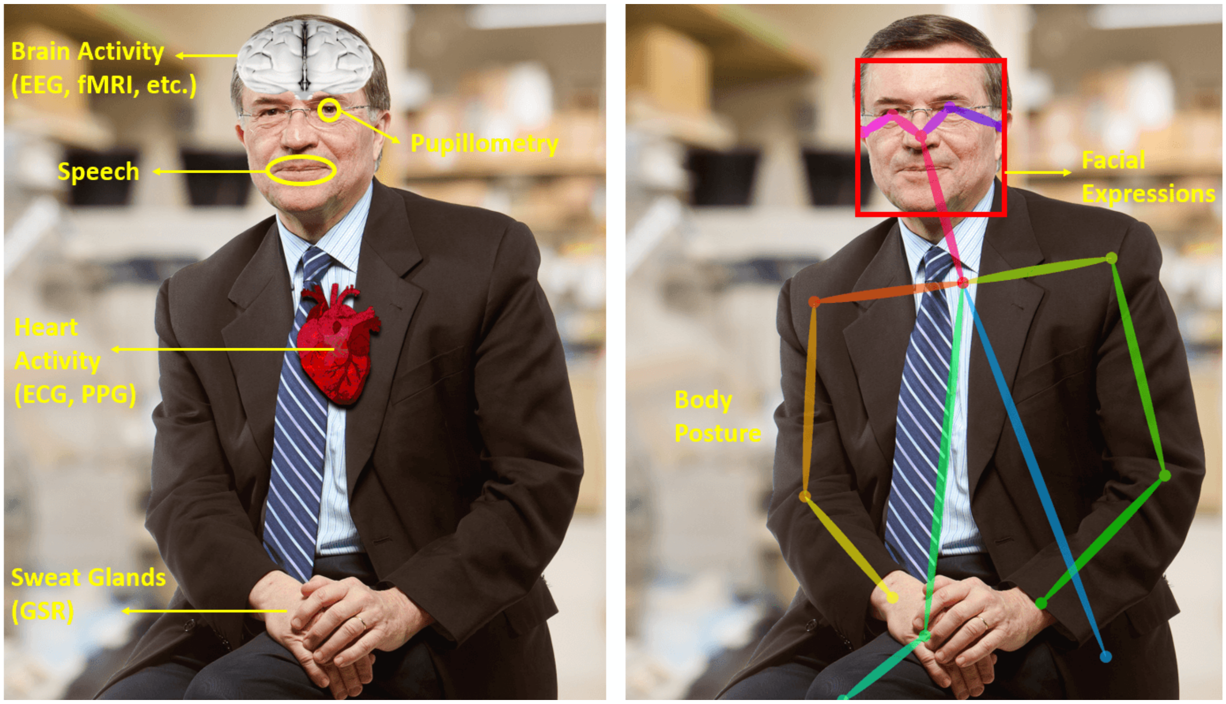


Figure 2.1: Various bio-sensing and physiological cues used for emotion recognition

"Follow your heart and not your mind," has been a much-quoted sentiment from centuries. Research in this field in recent times has pointed out that there may be scientific truth in the above adage. Neurocardiology has fast emerged as a new research avenue studying the role of the heart in emotion generation and perception [4]. Focusing on brain imaging and facial expressions to study emotions is still the dominant trend in affective computing. These modalities also have many limitations such as electroencephalogram (EEG) brain imaging technique being highly prone to noise due to any kind of motion. Utilizing facial expressions for emotion classification has its own limitations. For example, a computer vision-based modality for this application requires good illumination conditions, constant tracking of the human face i.e. camera positioning in front of the face, and raises privacy concerns for others in the vicinity who may be inadvertently being recorded by the camera. One way to address all of these concerns is to utilize bio-sensing modalities continuously strapped on the user's body. Many such real-world bio-sensors have been developed in recent times. Such sensors used to monitor the user's pupillometry, heart rate, temperature, skin conductivity, etc., have created an incentive to explore these physiological cues for emotion classification. Fig. 2.1 shows various such bio-sensing and physiological cues used for emotion recognition.

This incentive of utilizing multiple bio-sensing modalities has led to the collection of several multi-modal datasets for emotion classification, including DEAP [5], AMIGOS [6], MAHNOB-HCI [7], and DREAMER [8]. These datasets have been collected when users were watching multimedia content like movies and song clips while strapped with multiple sensor modalities including electroencephalogram (EEG), color camera, depth camera, galvanic skin response (GSR), electrocardiogram (ECG), photoplethysmogram (PPG), and eye-gaze. In addition to the datasets collected in a multimedia content setting to classify human emotions, the use of multiple sensor modalities has also made its way in the development of safe automobiles focused on assessing driver's awareness [9].

Most previous research studies have not adopted a multi-modal approach to affective computing and if they have, then the bio-sensing modalities used have not been wearable and capable of working in real-world settings. For example, the studies [10, 11, 12, 13, 14] utilizes only EEG in the context of driver awareness analysis and emotion classification while watching videos. On the other hand, studies like [7, 8, 6, 15, 16] do apply multiple bio-sensing and computer vision modalities for these applications but the sensors used in those datasets are incapable of data collection when the user is mobile. These sensor modalities do not come in a compact and integrated form factor but instead have to individually be placed on the user. It is because of such limitations that we developed a novel multi-modal bio-sensing headset with an integrated and compact form factor. We also utilized multiple bio-sensing modalities for various applications as detailed below.

A skeptic might ask what is the need for designing an automated system capable of recognizing human emotions? There are many applications of emotion recognition systems that are already being used worldwide. These can be broadly classified into three categories.

The first of these concerns health-related applications. These may constitute utilizing emotion recognition to detect how a patient is responding to various mental tasks or recovering from a traumatic accident. Such systems like facial expression monitoring ones can be used to track the patient's recovery over a period of time. Second, software usability testing and content rating. A huge market has come up in recent years in the advertising industry to assess products such as advertisements, video games, etc. based on tracking test users' physiology and facial expressions (bio-markers of human emotions) to gauge their feedback. This feedback can then be used to refine the product or make decisions about its rollback, etc. Third, such emotion recognition systems can be used to monitor and track students' responses to various types of learning content. By monitoring students' physiology in real-time and associating it with the learning contents, it may be possible to refine the latter based on automated feedback from the former. This application is in the early stages of research and may revolutionize how we learn and receive education in schools [17].

Ironically, the simple question, "What emotion are you experiencing at this moment?" is the greatest obstacle in the field of affective computing. This is because it is very hard to categorize emotions neatly in a disjoint manner.

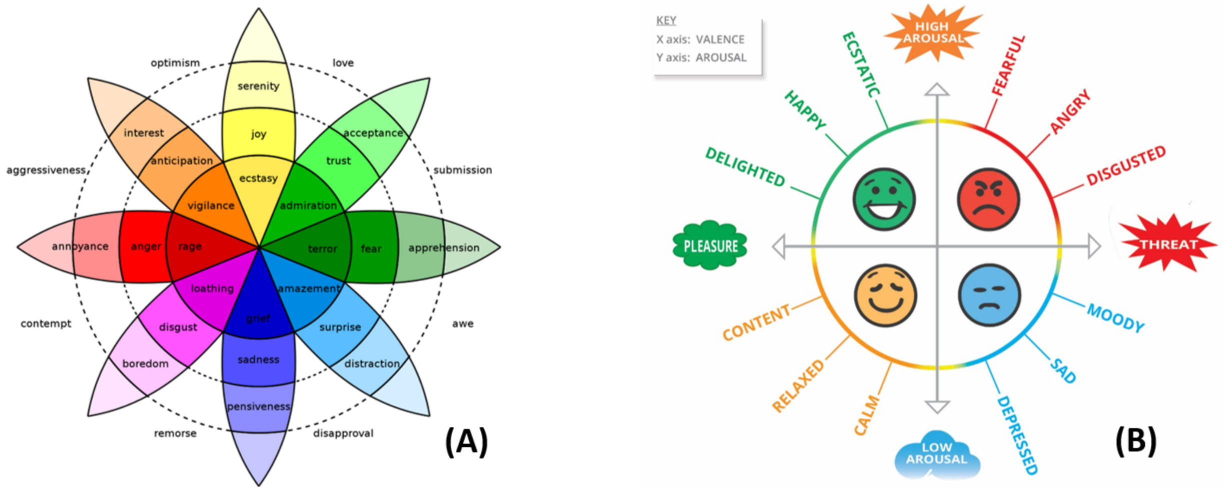


Figure 2.2: (A) Plutchik's Wheel of Emotions Model and (B) Russell's Circumplex Model

There are more than ten emotion classification models. While some classify emotions dimensionally, others do it discretely and they often do not agree with each other [18]. Figure 2.2 shows two of the most widely used emotion classification models. The Plutchik's Wheel of Emotions Model shown in Figure 2.2(A) suggests eight primary bipolar emotions whose interaction with each other generate other secondary emotions. This is akin to the RGB color model in computer vision. On the other hand, Russell's Circumplex Model shown in Figure 2.2(B) approaches emotion classification through affective states like valence and arousal and do not as such define any primary or secondary emotions. These states loosely define the mood and strength of a particular emotion. Utilizing valence and arousal it may be possible to map various emotions as shown in the figure. With so much discrepancy between various emotion classification models, it is imperative to monitor user's physiology to gauge their feelings rather than blindly asking them to rate their emotions on such complex measurement scales. It is to address this need to utilizing multiple bio-sensing modalities for emotion classification that this work fits itself. This work encompasses the designing of novel multi-modal bio-sensing systems, algorithms for classifying emotions, and studying the relationship between various physiological cues reflecting human emotions.

Any such multi-modal bio-sensing system for affective computing application must be capable of achieving five objectives:

- Being able to detect, classify, and monitor emotional and mental states.
- Infer emotional state using a minimal number of and the most comfortable sensors.
- Able to infer the context (such as what the user is doing) in real-time to make use of the processed data to perform an activity.
- Make recommendations, show/hide information, or take action based on the information from above (such as triggering music, modifying the environment in a game, etc.).

- Execute all the above in real-world settings throughout the day.

To encompass all five objectives into one doctoral thesis would have been impracticable. Thus, this doctoral dissertation focuses on the first three objectives in two real-world settings: watching affective multimedia content and driving a vehicle. These two real-world tasks are ubiquitous across all regions of the world. Detecting and classifying the emotions that are evoked while watching a video on YouTube or assessing a driver's state of awareness when s/he is behind the steering wheel are the primary focal points of this doctoral dissertation. The rest of this doctoral dissertation is organized in the following way:

Chapter 3 describes the state-of-art bio-sensing systems and their limitations. It then details the design, fabrication, and evaluation of a multi-modal bio-sensing system developed by us. This multi-modal bio-sensing headset is capable of monitoring, recording, and tracking a person's physiology in real-time and in a real-world manner. The comparison of the signal from each developed sensor is carried out with commercially available sensors. The last part of the chapter presents the evaluation of the headset through the examples of two real-world games.

Chapter 4 utilizes bio-sensing modalities such as those developed in Chapter 3 to understand the correlation between affective multimedia content and the user's brain state (monitored by electroencephalogram) and facial expressions. This chapter studies which regions of the brain and what kind of facial expressions are most affected by and are correlated to different types of emotional stimuli. Thus, this chapter takes forward our understanding of the elicitation of human emotions.

Chapter 5 takes the research forward from Chapter 4 to address the requirement of the multi-modal emotion classification algorithms in the real-world application of watching affective multimedia content. The chapter details the design of algorithms utilizing signal processing and deep learning techniques for various bio-sensing and camera-based modalities. These algorithms are then evaluated using four publicly available multi-modal affective datasets. The results of this study are shown to outperform the previously reported ones for these four datasets.

Chapter 6 utilizes the tools developed above in the context of a real-world driving scenario. The chapter details the collection of a multi-modal bio-sensing dataset as well as a vision dataset while the automobile was being driven in autonomous mode. The study expands on driver awareness based on how attentive the driver is in a particular setting and his/her reaction to various hazardous events that might occur during the drive. This application correlates with the study of human affects such mental states like awareness and attention are closely associated with a person’s emotions.

Chapter 7 summarizes the main contributions of this dissertation and discusses some future research avenues.

Chapter 3

Fabrication and Evaluation of a Real-world Multi-modal Bio-sensing Headset

This chapter describes the development and integration of various components of a novel multi-modal bio-sensing system. This system is capable of collecting, synchronizing, recording, and transmitting data from multiple bio-sensors: photoplethysmogram (PPG), electroencephalogram (EEG), eye-gaze headset, body motion capture, galvanic skin response (GSR), etc., while also providing task modulation features including visual-stimulus tagging. The developed sensors are evaluated by comparing their measurements to those obtained by standard research-grade bio-sensors. The earlobe-based photoplethysmogram (PPG) module with motion-noise canceling is evaluated against electrocardiogram (ECG) during the heart-beat calculation. The steady-state visually evoked potentials (SSVEP) measured by the novel shielded dry EEG sensors are compared against the potentials obtained by commercially available dry EEG sensors. The effect of head movements on the accuracy and precision of the wearable eye-gaze system is also investigated. Furthermore, two practical tasks are carried out to demonstrate the applications of using multiple sensor modalities for exploring previously unanswerable questions in bio-sensing. Specifically, bio-sensing is utilized to assess which strategy works best for playing "Where is Waldo?" visual-search game by studying the changes in EEG corresponding to true vs. false target fixations. Similarly, the performance of bio-sensing modalities is compared to classify loss/draw/win states in a "Rock-Paper-Scissors" game. The applications of the sensors developed in this chapter concerning affective computing will be discussed in the succeeding chapters.

3.1 Introduction and Related Research

In recent years, there have been many advances in the field of wearable bio-sensing. This trend has led to the development of multiple wearable bio-sensors capable of measuring galvanic skin response (GSR), photoplethysmogram (PPG), etc. integrated into portable form-factors such as smartwatches. The use of bio-signals for various applications such as robotics [19, 20], mental health [21], affective computing [22], human-computer interaction [23, 24] etc. has been expanding throughout the past decade. At the same time, the concept of using more than one bio-sensing modality has also gained popularity. This is primarily driven by the assumption that the limitations of a bio-sensor can be compensated by using another for specific applications. For example, since EEG provides good temporal resolution but poor spatial resolution it might be possible to use other modalities such as PPG and GSR to augment the performance in an emotion classification task rather than using EEG alone [5, 25].

Unfortunately, the integration of the above-mentioned bio-sensing modalities is usually overlooked for more naturalistic research studies due to cost, bulk, and technical difficulty [26]. A typical strategy used to measure multiple bio-signals in the real-world is to buy various sensors and then extract data from each of them separately. This setup, however, leads to unwieldy subject preparation and increased post-processing synchronization effort, both of which add sources of noise and inconvenience. Specifically, no integrated headset has been proposed to measure multiple bio-signals simultaneously in a synchronized manner. Without the possibility of simultaneous recordings from multiple modalities, it is difficult, if not impossible, to explore questions corresponding to changes in physiology while performing actions in the real world.

The problem of not being able to collect data in real-world environments is compounded by the lack of techniques to automatically recognize and tag real-life events or stimuli. The standard process employed for stimulus tagging requires an individual (experimenter) to manually tag the various stimuli from frame to frame in a video stream. This process is cumbersome, time-consuming, and laborious. Furthermore, the stimulus onset is not measured with fine-resolution or is ill-defined in such setups [27, 28]. A solution is to track eye-gaze to infer the stimulus onsets. This allows pinpointing of the visual region but still requires stimulus tagging.

Additionally, there is a design element associated with the bio-sensors, which needs to be optimized for compactness and cost for any multi-modal bio-sensing system. Any such device has to be capable of synchronizing multiple data streams and should be packaged in a compact form factor for easy use. For example, the use of wet electrodes for measuring EEG or electrocardiogram (ECG), which may require placing sensors over the chest, is undesirable for real-world setups. This research proposes a novel multi-modal bio-sensing system to address the above limitations.

3.1.1 Contributions

The main contributions of this work are as follows:

- This work presents the design of a novel research-grade bio-sensors capable of measuring physiological measurements in real-world environments with automatic visual tagging and integrating the sensors in the form of a compact wearable headset.

- This work also presents the evaluation of the developed sensors on two “real-world” experiment setups through the use of gaming towards utilizing multiple sensor modalities.

- This work shows conclusively that previously unexplored physiological questions can be addressed using multiple sensor modalities.

The above contributions were made by designing and evaluating a novel earlobe-based, high-resolution PPG sensor capable of measuring heart rate and heart-rate variability as well as providing raw PPG data from the earlobe. Using adaptive noise cancellation and intentional placement at the earlobe to minimize sensor movement, the PPG sensor is capable of minimizing motion noise. Novel dry EEG sensors capable of actively filtering the EEG signals by shielding themselves from ambient electrostatic noise were also designed and evaluated. These EEG sensors are used with a high-sampling and ultra-low-noise analog to digital converter (ADC) module. Subsequently, a dual-camera-based eyeglass capable of measuring eye-gaze (overlaid on the subject's field of view), pupillometry, fixations, and saccades was fabricated and evaluated. Data acquisition and synchronization from all these sensors are done using an embedded system. These data streams can then be saved on the device or wirelessly transmitted in real-time for visualization and analysis. The framework is designed to automatically tag visual stimuli in real-world scenarios with the user's eye-gaze over the various bio-sensing modalities. Finally, the framework is scalable such that it can be expanded to support other bio-sensing modalities from the market.

3.1.2 Related Research Studies

Table 3.1 compares this system with many existing state-of-the-art bio-sensing systems. Clearly, in all categories, this system is more comprehensive and flexible than all the existing bio-sensing systems. The lack of cost comparison in the Table is because many of the systems including this one have not been commercialized into products. Hence, it is not a fair comparison to evaluate the retail prices of select systems with the fabrication costs of others.

In real-world applications, PPG has been substituted for ECG due to the ease it offers in measuring heart rate. It does not require using wet electrodes over the chest and can easily

Table 3.1: Comparison of Multi-modal Bio-sensing Systems

Features	**This System**	iMotions [26]	Microsoft [29]	Biovotion [30]	Teaergo [31]	OpenBCI [32]	Imperial College [33]
Sensing Modalities	**EEG, PPG, Eye-Tracking, Pupillometry**	EEG, ECG, GSR, EMG, Eye-Tracking	PPG, GSR, Skin Temperature	PPG, GSR, Skin Temperature	EEG, EMG, GSR, Eye-Gaze, Motion Tracking	EEG, EMG, ECG	EEG, ECG
Fully integrated, self-contained module*	**Yes**	No, multiple modules	**Yes**	**Yes**	No, multiple modules	**Yes**	No, multiple modules
Wireless Synchronization†	**Yes**	**Yes**	No	No	**Yes**	No	No
Automatic Visual Stimuli Tagging	**Yes**	No	No	No	No	No	No
Noise canceling measures	**Yes (ANC, Shielding)**	No	No	No	No	No	No
Research grade‡	**Yes**	**Yes**	No	No	**Yes**	**Yes**	**Yes**
Performance Evaluation while in motion	**Yes**	No	**Yes**	**Yes**	**Yes**	No	No

*Containing data acquisition, noise filtering, digitizing, transmission circuitry and battery.
†Capable of synchronizing with any external sensor too while transmitting wirelessly.
‡Capable of acquiring data with high bit resolution and high sampling rate.

be integrated onto watches or armbands [34]. But, it has its own limitations. First, most of the available PPG sensors do not have a sampling rate high enough and fine ADC resolution to measure heart-rate variability (HRV) in addition to heart rate (HR). HRV has been shown to be a good measure of emotional valence and physiological activity. Secondly, PPG sensors over the arm or wrist tend to be noisy because of the constant movements of the limbs while performing real-world tasks. On the other hand, PPG systems designed for the earlobe also suffer from noise due to walking or other head/neck movements [34]. In the rare case when noise filtering is used in

PPG, the hardware design is bulky due to the large size of the circuit board used in the setup [35].

EEG sensors come in dry or wet-electrode based configurations. The wet electrodes either require the application of gel or saline water during the experiment and hence are not ideal outside laboratory environments [36]. The dry electrodes usually do not have a long service life since they are generally made of Ag/AgCl or gold (Au) coating over metal, plastic, or polymer, which tend to wear off [37, 38]. Furthermore, coating Ag/AgCl is a costly electrochemical process.

Eye-gaze tracking systems tend to be bulky and may even require the subject to place his/her face on a chin rest [39, 40]. Even when they are compact, these systems are not mobile and the subject has to be constantly in its field of view [41]. These limitations restrict their use outside well-controlled laboratories, where illumination varies and the subject is mobile at all times. Furthermore, all such systems only measure eye-gaze as being pointed over a display monitor and not in the real world. They are unable to overlay the gaze over the subject's view if the display screen is not in his/her field of view. The solution is to use head-mounted eye-gaze systems but they tend to use a laptop instead of a small embedded system for processing and viewing the camera streams [42]. Thus, the laptop has to be carried in a bag, restricting the subject's freedom of movement.

To tag the stimuli with various bio-sensing modalities, the norm has been to use a key/button press, fix the onset and order of stimuli on a display, or time it with a particular event, etc. [27, 28] But, in real-world scenarios, such methods either cannot be used due to the mobile nature of the setup or induce a sense of uncertainty which has to be removed by manual tagging. Such manual tagging is laborious and time-consuming. The only viable solution is to tag stimuli automatically after recognizing them in the subject's field of view. However, it lacks the knowledge of whether the subject was actually focusing on the stimuli or rather was looking at some other areas in his/her field of view.

The existing multi-modal experimental setups are tethered, are not compact, and tend to just attach various sensors on the subject, which are then connected to one or more data acquisition systems [22, 26]. This further reduces the mobility for experiments outside laboratories. The use of an independent clock for each of the different modality complicates the issue of synchronizing the various modalities. For real-time display, transmitting data streams from these sensors over Wi-Fi or Bluetooth may introduce varying latency. Thus, the only solution is to design a closely packed hardware system [43], which synchronizes the various data streams while acquiring them in a wired manner and using only one clock (that of the embedded system itself). The synchronized streams can then be either recorded or sent to a display screen which does not affect either the compact nature of hardware or synchronization in software framework. The next sections present the various sensors and methods developed in this study to address the above limitations.

3.2 System Overview

This section details the development of each of the sensor modules incorporated in the developed system with its features and the embedded system used for their integration.

3.2.1 Earlobe-based PPG Sensor

The developed earlobe-based PPG sensor module is very compact (1.6 x 1.6 x 0.6 cm) and sandwiched to the earlobe using two small neodymium magnets. The PPG sensor module (Fig. 3.1) houses an Infrared (IR) emitter-detector (Vishay TCRT 1000) for measuring PPG, a 3-axis accelerometer (Analog Devices ADXL 335), a high-precision (16-bit) and a high sampling rate (100 Hz.) ADC (Texas Instruments ADS 1115), and a third-order analog high-gain band-pass filter (BPF, cutoff 0.8 - 4 Hz using three Microchip MCP6001 op-amps).

This PPG signal is then amplified using the high-gain band-pass filter and a relevant frequency band is extracted. The filtered PPG data along with the accelerometer's data are digitized using the ADC before transmission. Thus, the PPG module is capable of filtering the

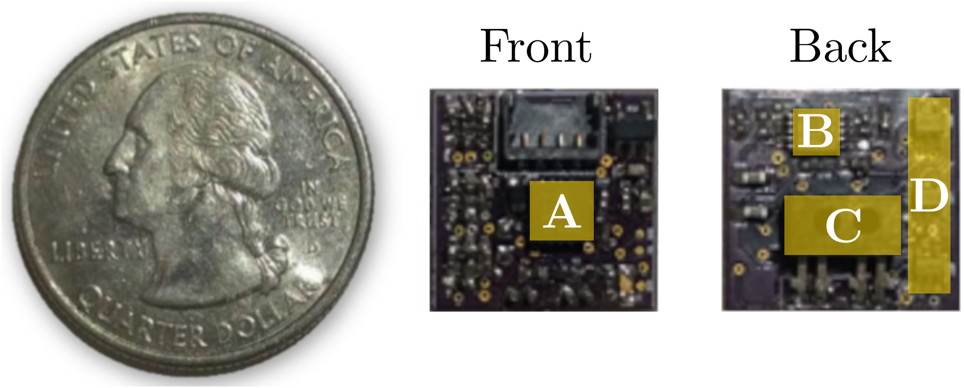


Figure 3.1: The miniaturized PPG sensor with a scale reference. (A) 3-axis accelerometer, (B) 100 Hz 16-bit ADC, (C) IR emitter and receiver, and (D) a third-order filter bank.

signal on-board (despite being so minuscule in size) and converting the signal to a digital format for the calculation of heart rate and heart-rate variability. The on-board accelerometer serves two purposes. First, it can be used to measure and monitor head movements because the sensor is fixed on the earlobe with reference to the position of the subject's face. Secondly, the accelerometer provides a measure of noise due to motion and removes it from the PPG signal using an adaptive noise-cancellation (ANC) filter. The filter [44] can be implemented inside the embedded system (Section 3.2.5) in real-time. The filter works by constructing a model of noise due to motion (such as while walking) from the reading of the accelerometer and reconstructing the noise-removed PPG signal.

3.2.2 EEG Sensors and Data Acquisition

The novel dry EEG sensors (Fig. 3.2) developed in this research can be easily adjusted under the hairs to measure EEG signals from the scalp. These EEG sensors consist of a highly conductive element made from silver (Ag) epoxy (electrical resistivity 0.007 Ω·cm). This silver-epoxy-based conductive element provides the sensor a long life since the silver does not wear off as fast as it does on EEG sensors coated with Ag/AgCl. The sensor also has an on-board OpAmp (Texas Instruments TLV 2211) in a voltage-follower configuration to shield the EEG signal from noise by increasing the signal-to-noise ratio (SNR) of the EEG signal. Furthermore, the sensor is enclosed in a Faraday cage made of conductive copper foil tape. This shielding is used to remove external noise from the environment before the signal is digitized. For subjects with dense hair, a drop of saline water can be added to increase the conductance between the sensing element and the subject's scalp.

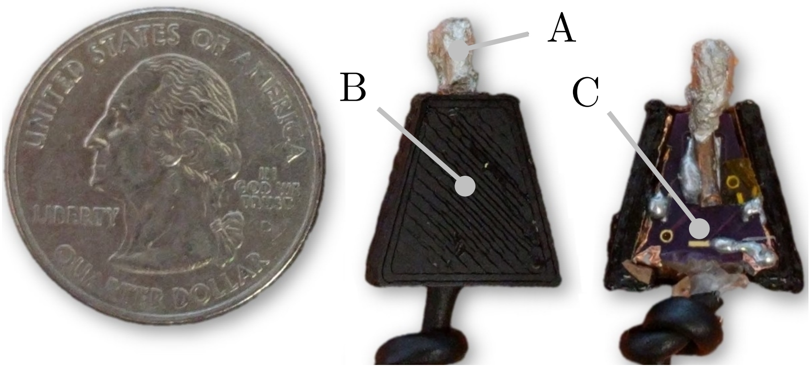


Figure 3.2: The EEG sensor with a scale reference. (A) Silver (Ag) based conductive element, (B) 3D printed case housing a conductive element for shielding, and (C) The amplifier circuitry.

For converting the analog noise-removed EEG signal to a digital format, an assembly for fine resolution (24-bit), high-sampling rate (up to 16k samples/second), ultra-low input referred noise (1 µV) ADC (Texas Instruments ADS 1299) was designed. This assembly employs a low-pass filter before the signal goes into the ADC, whose parameters such as sampling rate, bias calculation, internal source current amplitude for impedance measurement, etc. can be controlled by the software [45, 46]. The assembly can support up to eight EEG channels (Fig. 3.5F) whereas the design of the board is such that multiple boards can be stacked to accommodate more EEG channels. Two such boards were used in the headset to support 16 EEG channels (Fp1, Fp2, F7, F3, Fz, F4, F8, C3, Cz, C4, P3, Pz, P4, O1, Oz, and O2 according to the International 10-20 EEG placement). Continuous impedance monitoring is made for each electrode in real-time to assess the quality of the EEG signal and electrode placement. Furthermore, using Independent Component Analysis (ICA) [47, 48], various independent components of the signal can be separated to identify noise due to blink, eye movement, EMG, etc.

3.2.3 Eye-Gaze and Pupillometry sensors

Two miniature cameras (Fig. 3.6) are used to assess the subject's eye-gaze location and pupillometry parameters such as the diameter of the pupil, fixations, saccades, etc. The eye camera consists of two infrared (IR) LED's (970nm wavelength), which are used to illuminate the region around the eye. Because the LEDs are IR-based, the eye camera can detect the pupil under a wide variety of illumination conditions. PupilLabs' pupil-detection algorithm and eye-gaze calibration software have been modified to detect the pupil and calibrate the subject's gaze [42]. A display screen (i.e. laptop) is needed only for the initial eye-gaze calibration step, which is done using a manual selection of natural features in the field of view. The gaze is then superimposed on the subject's view from the world camera. Both cameras stream at 30 frames-per-second (fps) while the resolution can be adjusted as per the need of study.

3.2.4 Stimulus Tagging

You Only Look Once (YOLO) deep learning algorithm was used to automatically tag various stimuli in the feed from the world camera in real-time [49]. The algorithm can be trained for custom object classes using large image databases with multiple classes depending on experimental needs (for example 80 object categories in COCO dataset with 300K images). Whenever the subject's gaze falls inside the bounding box of one of the object classes (stimuli), the bio-sensing modalities are automatically tagged. Therefore, instead of manually tagging the stimuli during the experiment, the software automatically tags the salient information. Thus, for example, if the subject is looking at a person's face, his/her EEG can be time-synchronized to the gaze and analyzed to detect the level of arousal (Fig. 3.3). Due to the significant computational requirements of using YOLO, the stimulus tagging is done on a laptop (rather than the embedded system) in real-time or processed post-hoc. OpenPose [50] is used to automatically tag the human pose by detecting the positions of the body joints. Lab Streaming Layer (LSL) [51] library is used to synchronize the data streams from the camera on the embedded system and stimulus tagging on the laptop.

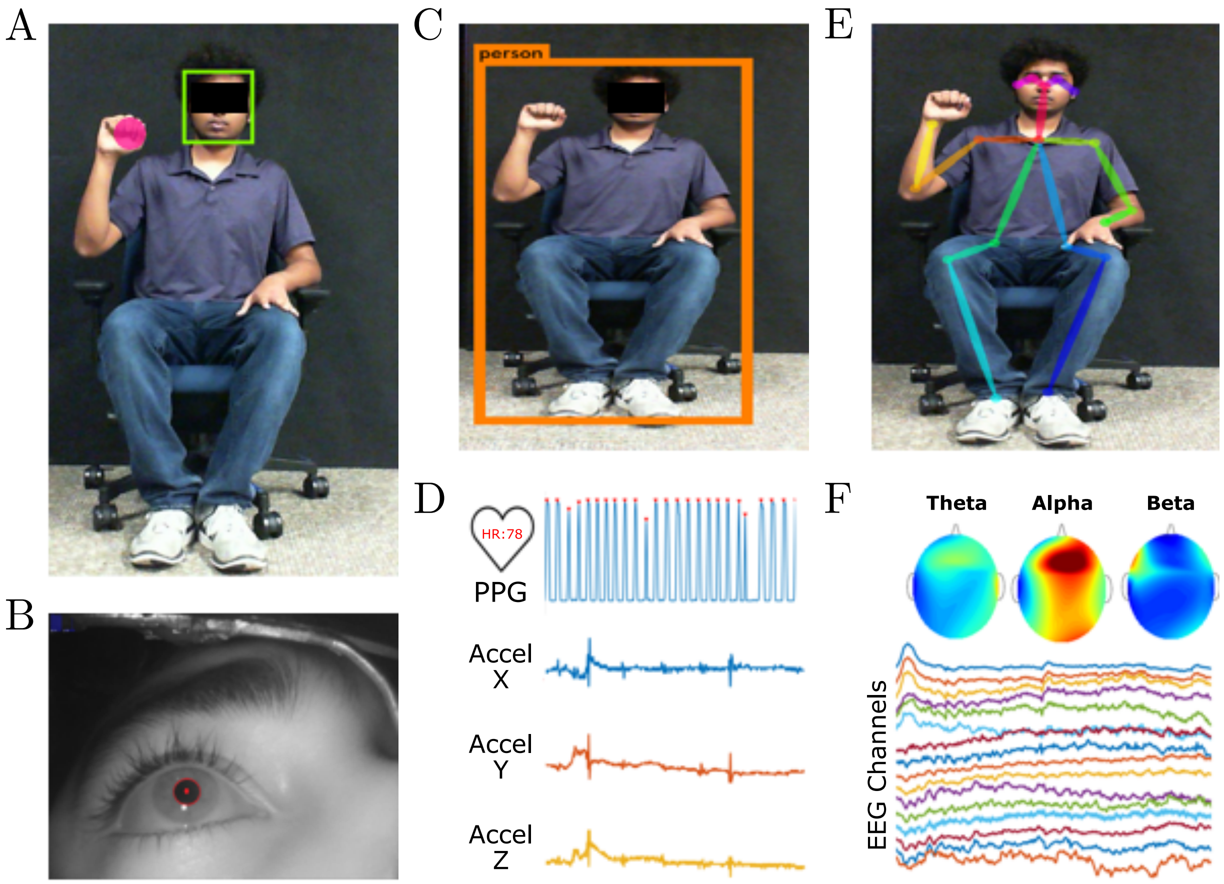

[redacted]
[redacted]

Figure 3.3: Visualization of the software. (A) Eye-gaze overlaid on world view, (B) Detected pupil in IR eye camera image, (C) Object detection network, (D) PPG and accelerometer signals with noise canceling and heart-rate computation, (E) Human pose detection network, and (F) EEG signals and power spectral density of three EEG bands.

3.2.5 Embedded System Framework

Each of the above modalities is wired to a custom development board (Fig. 3.5), which uses an embedded system containing the Broadcom BCM2837 processor. The board has the

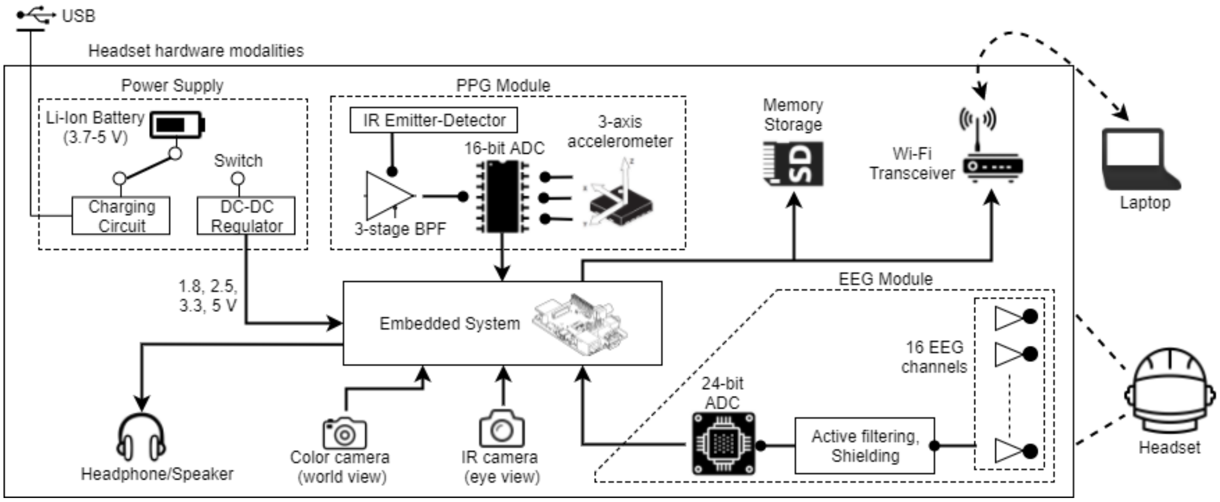


Figure 3.4: Overview of the integrated system architecture.

capability to attach a world camera, eye camera, PPG module, and EEG module. Additionally, the board houses a headphone jack which can be used for playing audio recordings during experiments. The clock on the embedded system is common for all modalities helping to ensure data synchronization from independent streams using LSL. This library allows for the spawning of a global clock which takes into account the relative difference between local clocks on the embedded system and laptop for synchronizing various data streams from the two devices in real-time. The video streams were compressed using MJPEG compression while transmitting them wirelessly. Fig. 3.4 shows the block diagram of the complete working architecture of the system, sensor components, data processing, and transmission by Wi-Fi (using Realtek RTL 8723BS module). The system is powered using a small Li-Ion battery (Panasonic NCR18650B 3400mAh), which lasts for approximately three hours when all sensor modalities are enabled. However, the system can also be powered by any compact 5V-output mobile power bank for more than eight hours of continuous use.

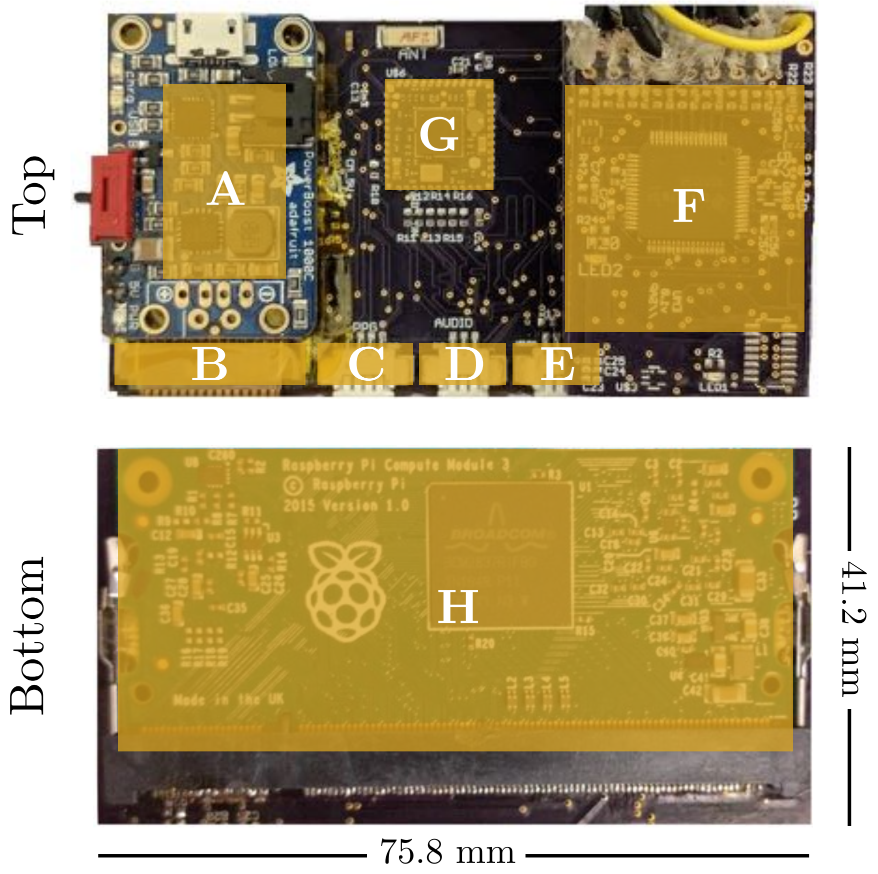


Figure 3.5: Embedded System (A) Power circuitry, (B) World camera connector, (C) PPG connector, (D) Audio jack connector, (E) Eye camera connector, (F) EEG sensors connector and ADC module, (G) Wi-Fi module, and (H) Microprocessor module

3.3 Evaluation of Individual Sensor Modalities

To evaluate the efficacy of the integrated headset (Fig. 3.6), the evaluation of the individual components on multiple subjects was carried out. Two experiment scenarios were designed to this end using commonly played games: “Where is Waldo?” and “Rock-Paper-Scissors” during which multi-modal data was collected using this system. The results for the fusion of information from individual modalities are provided below with new insights into human physiology. The human trial portion of this study was reviewed and approved by an Institutional Review Board (IRB) of the University of California San Diego, and all subjects provided informed consent.

Multiple benchmarks were used to evaluate each of the sensors developed in this study. The apparatus designed for evaluating these sensors also include various types of head and body movements to assess the effect of the above-mentioned noise-canceling techniques.

3.3.1 Earlobe PPG Evaluation

The earlobe PPG module was evaluated during rest and active conditions. In particular, the feature of interest, heart rate, was measured while subjects were sitting and walking in place. The PPG sensor was placed on the earlobe as in Fig. 3.6 and measured the changes in blood volume at the sampling rate of 100 Hz. Simultaneously, the baseline was collected using an EEG/ECG acquisition system sampled at 1 KHz from the Institute of Neural Engineering, Tsinghua University, China. Three electrodes were placed on the subjects' chest over the heart, and on either side of the ribs.

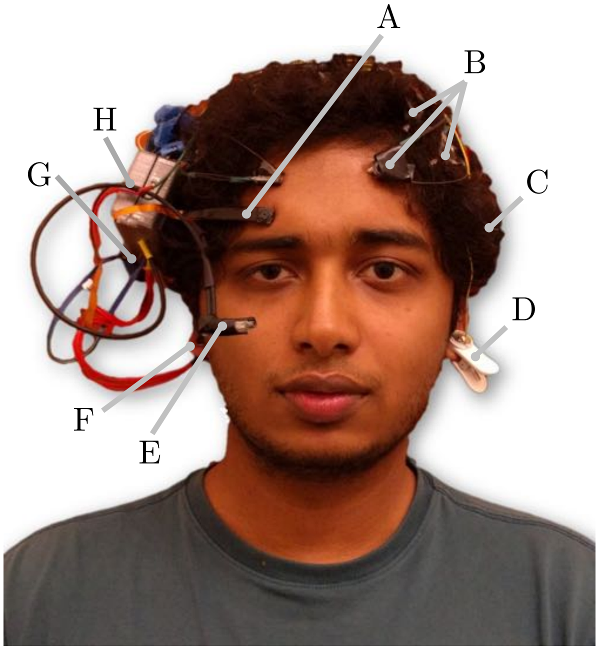


Figure 3.6: Integrated Headset (A) World camera, (B) EEG Sensors, (C) Battery, (D) EEG Reference Electrode, (E) Eye Camera, (F) Earlobe PPG Sensor, (G) Headphone/speaker connector, and (H) Embedded System (The subject gave consent to use his face for publication)

Six subjects participated in eight different trials: four sitting and four walking, during which their ECG and PPG data were simultaneously measured. In each trial, two minutes of data were collected. For the walking condition, subjects were instructed to walk-in-place at a regular rate and ANC was performed to remove motion noise. A peak detection algorithm was used to find the heart beats in both signals for counting the heart rate.

Fig. 3.7 shows the working of the 10th order ANC filter utilized on a 10-second interval of PPG data while walking. The original PPG data (in blue) in Fig. 3.7A is clipped at the top because a third order high-gain band-pass filter was used thus amplifying the signal and making it easier to distinguish the peaks in PPG. Fig. 3.7A shows that the number of peaks in the original waveform is computed to be 20, which is incorrect as the waveform is distorted. Measurement of the noise from vertical acceleration (Fig. 3.7B) was used for the ANC filter. Noise-removed PPG waveform with the use of the ANC filter is shown in Fig. 3.7A and as expected the erroneous peaks are eliminated, giving the total number of peaks as 17.

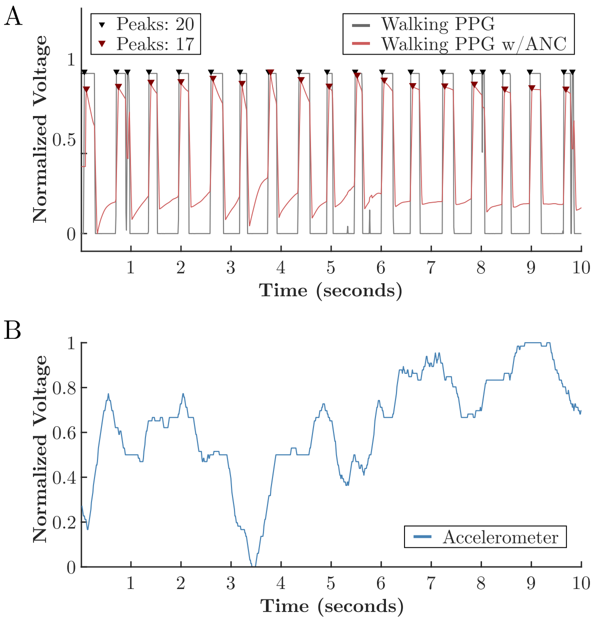


Figure 3.7: Comparison of the 10-second waveforms from the earlobe PPG sensor before and after ANC during walking. (A) PPG without and after noise-cancellation (B) Vertical acceleration used as the noise measure.

Bland-Altman analysis [52] was then performed which is a general and effective statistical method for assessing the agreement between two clinical measurements to compare the heart rate obtained by the developed PPG module to the true heart rate computed using the high-resolution ECG signal. Fifteen-second trials were used to calculate the HR using the peak-detection algorithm. Fig. 3.8A shows the result of the Bland-Altman analysis while the subjects were sitting. As the figure shows most of the trials are between the Mean ± 1.96SD agreement threshold for both, with and without using ANC. Further, using ANC decreases the agreement threshold, making the two signals adhere to more conformity. Similar results were observed for the trials when subjects were walking (Fig. 3.8B) and again using ANC makes the HR measures from the two signals more agreeable. Furthermore, for both cases, the trials from the two signals were almost always in agreement, indicating that the earlobe PPG module is capable of measuring heart rate with high accuracy.

3.3.2 Eye Gaze Evaluation

The performance of the paired eye-pupil monitoring and world-view cameras in measuring eye gaze were evaluated using a structured visual task to measure precision and accuracy during use. The gaze accuracy and precision were measured for subjects following calibration (an ideal setting) and after head movements (a real-world use). In this task, the subjects were asked to calibrate their eye gaze using nine targets which appeared on a screen 2.5 feet away from them (such that >90% of the camera's field of view was composed of the task-screen). For six subjects, a series of 20 unique targets randomly distributed on the screen were used to account for the majority of their field of view. Thus, this composed the accuracy and precision measurements just after calibration. The subjects were then asked to move their head naturally for 30 seconds without removing the headset. This action was designed to simulate the active head-movement scenarios when wearing the headset because usually the gaze performance is not reported after the subject has moved from his/her position. Similar to the above task, the subjects were asked

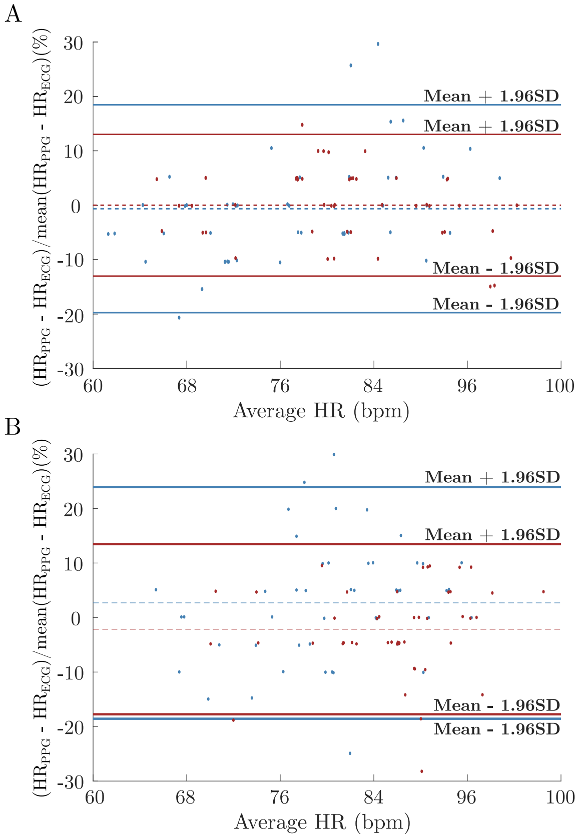


Figure 3.8: PPG vs. ECG Bland-Altman Evaluation (Blue- Before ANC, Red- After using ANC) (A) While sitting and (B) While walking. Heart-rate computed by ANC filtered PPG conforms more closely to the true heart-rate.

to again gaze at 20 different points appearing on the screen to assess the gaze performance after head movements. The above process was repeated three times for each subject. Importantly, no chin rest was used during or after calibration so that gaze performance is measured with natural head and body movements.

The accuracy is measured as the average angular offset - distance in degrees of the visual angle - between fixation locations and the corresponding fixation targets. Fig. 3.9A shows the gaze accuracy obtained before and after head movements. The mean gaze accuracy over all the trials was found to be 1.21 degrees without and 1.63 degrees after head movements. The decrease in gaze accuracy after head movements is expected because the headset's position is displaced albeit by a small value. For all the subjects, the mean gaze accuracy was mostly less than 2

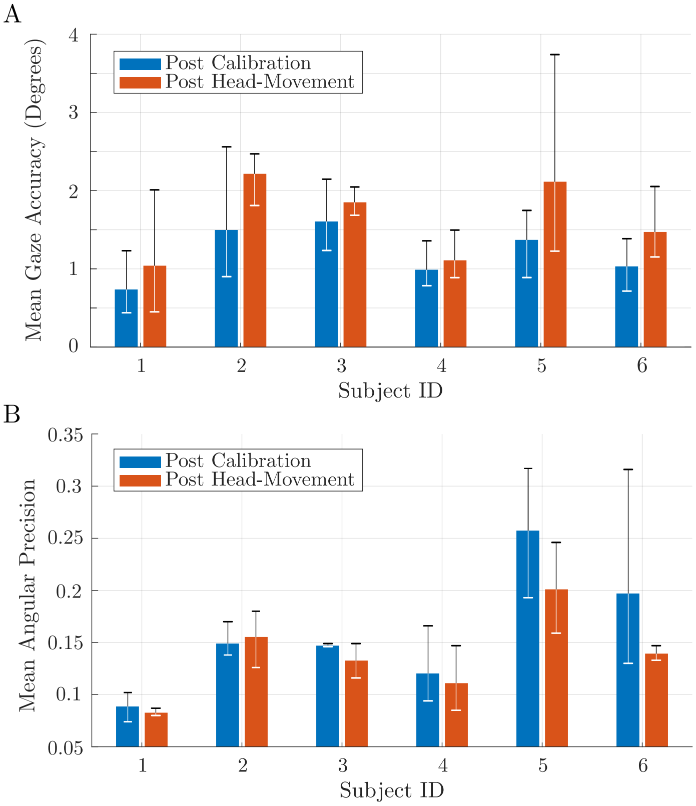


Figure 3.9: (A) Gaze accuracy (average angular offset between fixation locations and corresponding targets) evaluation and (B) Gaze precision (root-mean-square of the angular distance between successive samples during fixation) evaluation. For both metrics, this system performs as well or better than such existing ones.

degrees and the mean performance drifts only 0.42 degree, which is significantly less than 1-2 degree drift in commercially available eye-gaze systems [39].

The precision is measured as the root-mean-square of the angular distance between successive samples during a fixation. Fig. 3.9B shows the results of the angular precision for all the subjects. The mean angular precision was found to be 0.16 and 0.14 before and after head movements respectively. As is clear from the figure, the degree of visual angle is almost always within the range of ±0.15. Furthermore, the precision has a mean shift post head movement of only 0.2, indicating a minimal angular distance shift comparable to existing systems.

3.3.3 EEG Sensors Evaluation

For the evaluation of EEG sensors, two-fold was performed. First, the developed EEG sensors were compared with the state-of-the-art dry EEG sensors by Cognionics [38] to evaluate the signal correlation achieved using the two types of sensors. This also proves to be a test of whether the developed EEG sensors are actually acquiring EEG as opposed to just electromagnetic noise, and if they are able to shield themselves from ambient noise in the environment. Second, the developed sensors were evaluated on a steady-state visually evoked potentials (SSVEP) BCI task to evaluate their performance in measuring various frequencies during standard use.

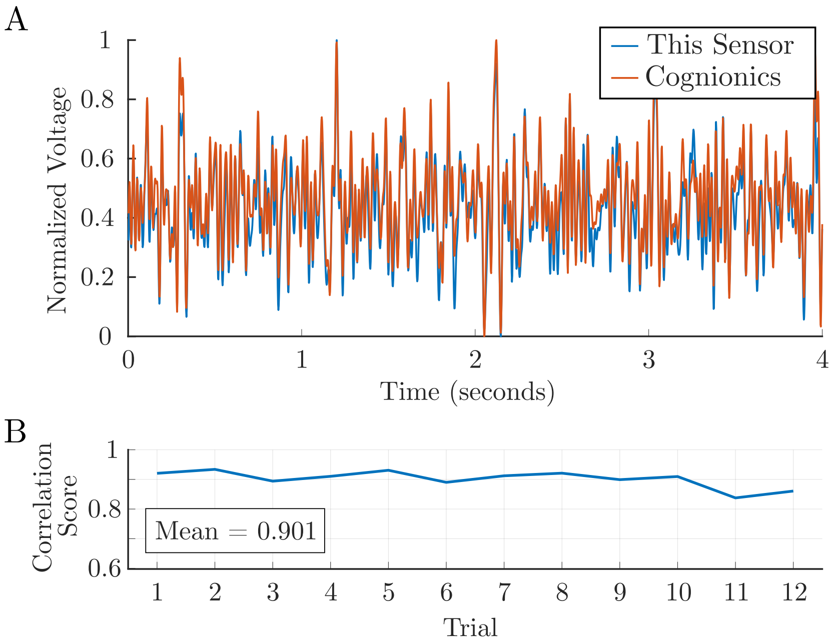


Figure 3.10: (A) A comparison of 4-second EEG signals acquired by the proposed and Cognionics dry EEG sensors and (B) Correlation score between the EEG recorded from the two sensors. Recordings from both the sensors show very high correlation.

For the SSVEP testing, five of the developed EEG sensors were placed at T5, O1, Oz, O2, and T6 sites according to the standard EEG 10-20 system. The location over and near the occipital lobe was chosen to evaluate the performance of the developed EEG sensors because the SSVEPs in response to repetitive visual stimuli of different frequencies are strongest over the occipital lobe.

Ten subjects participated in this experiment constituting three trials of ten random numbers each to be typed using an SSVEP-based keypad on a mobile tablet (Samsung Galaxy S2) with an EEG sampling rate of 500Hz. The frequencies of the 12 stimuli on the keypad (BCI speller) varied between 9-11.75 Hz with increments of 0.25Hz. This fine resolution in increment was chosen to assess the capability of sensors in distinguishing between minutely varying frequencies. The stimulus presentation time was 4 seconds with an interval of 1 second of a blank screen between two consecutive stimuli. Only the middle 2 seconds of data was used from each trial for SSVEP analysis. To compare the signal quality obtained from the two types of sensors, Cognionics sleep headband was used to acquire EEG from one Cognionics sensor at the temporal lobe and one of the developed sensors next to it. The location was chosen so that hairs on the scalp are present around the sensors.

Fig. 3.10A plots 4-second of EEG data acquired by the two sensors where a high correlation between the two signals is evident and almost always they follow a pattern. Fig. 3.10B plots the correlation of a subset of 12 of the total trials. The correlations between the EEG signals acquired by the two different sensors were very high (the mean correlation reached 0.901), indicating that the dry EEG sensor developed in this study is capable of measuring EEG signals from the hair-covered scalp areas.

As mentioned above, each subject needed to ‘type’ ten digits in each of the three trials. The SSVEP classification performance (Fig. 3.11) was computed using the filter-bank correlation analysis [53, 54]. This method does not require any training and is capable of working in real-time. As mentioned above, only the middle two seconds of EEG data was used during the 4-second stimulus presentation for evaluation. For almost all the subjects, the performance of SSVEP accuracy was very good ($\sim$80% accuracy). There were some expected variations because it is well known that the signal-to-noise ratio of SSVEPs varies among individuals. The mean performance across all the subjects was 74.23%.

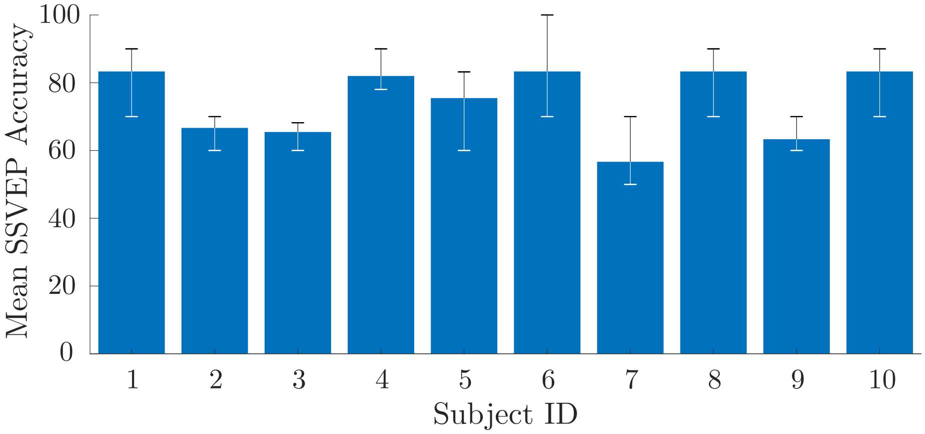


Figure 3.11: SSVEP accuracy plot for 0.25Hz increments in stimulus frequencies. Even for small increments of 0.25 Hz, high SSVEP accuracy across subjects was observed.

3.4 Multi-modal evaluation in "real-world" scenarios

Two experiments were designed in which ten subjects participated to play two games. For the "Where is Waldo?" game, the 30th anniversary book of the series was used which contains thirteen different scenes (trials) in which the target ("Waldo") has to be searched for. This experiment scenario was chosen because it allows for analyzing gaze-related fixations and patterns. While searching for the target, many non-targets are present and hence target vs. non-target event-related potential (ERP) can also be assessed. The book was placed at a distance of about 20 inches from the subjects who were seated and equipped with the sensor modalities mentioned above. The subjects were asked to search for the target without any time constraints. The researcher pressed a button before each trial and asked the subjects to start searching for the target. The subject conveyed that s/he has found the target verbally by saying "Found". Between each trial, the researcher flipped the page of the book to the next scene.

In the "Rock-Paper-Scissors" game, the researcher was seated in front of the subject (see Fig. 3.3A). A sound beep every 17 seconds signaled the start of a trial. For each subject, the game was played for 50 trials. The first two seconds after the sound beep were used to play the game whereas the next fifteen seconds the subject was asked to rest. The subject was instructed to only play the game from his/her wrist and not twist the whole arm to avoid any headset movement. During the game-play, motivational phrases such as “Come on! You can still win.”, “Watch out! I will now make a comeback”, etc. were utilized to ensure continual subject motivation. After each trial, the researcher marked the outcome (win/loss/draw) for the subject by pressing a button. The choice of this game was made to analyze the changes in human physiology, particularly EEG and cardiac signals during the perception of winning/losing a game. These changes might be closer to positive/negative valence in emotional states but much more reliable since winning/losing are independent of one’s likes and dislikes that influences his/her emotions.

3.4.1 Gaze-fixation pattern vs. time taken in finding the target

Different subjects used different strategies while searching for the target and hence the aim was to study that which strategy works best for this type of gaming paradigm by investigating the subjects’ eye-gaze fixations across the trials. The average time taken by the subjects is plotted (in ascending order) to find the target across the thirteen trials for the first visual-search experiment in Fig. 3.12A. Clearly, three distinct groups of subject strategies were observed and have been marked by different colors based on the average time taken by them. These groups were formed statistically taking 33^{rd} and 66^{th} percentile of the data as boundaries. Such a wide distribution is understandable since subjects use different strategies to find the visual target. Some subjects start looking for the target in small portions of the whole page while moving slowly towards previously unexplored parts whereas others tend to randomly scan the available page with longer distances between successive fixations.

To exclude the data associated with eye blinks (i.e. eye closures) a confidence threshold of 70% was imposed in the pupil detection algorithm. Subsequently, Euclidean distance of 25 pixels was used as the maximum inter-sample distance and 500 milliseconds as the minimum fixation duration to find all the fixations associated with the trials for all the subjects. The median (more robust than mean since a single large distance between successive fixations would skew

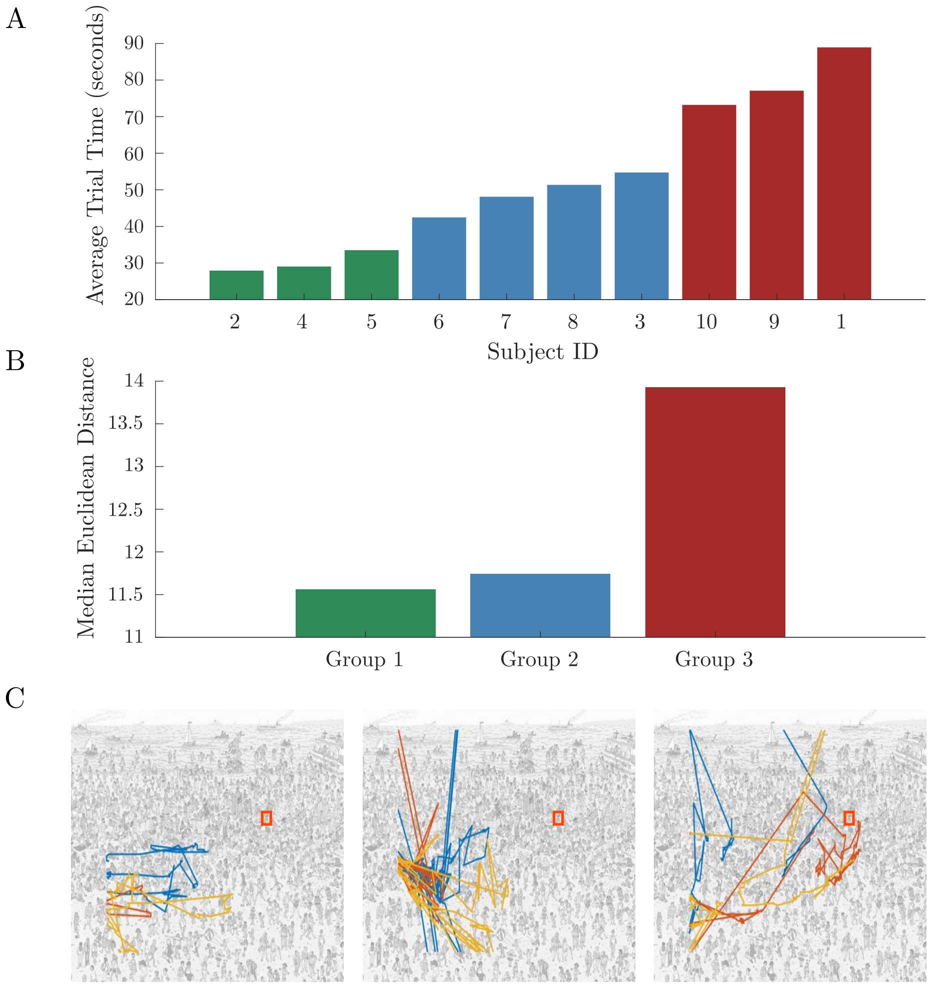


Figure 3.12: (A) Average time taken by subjects across trials, (B) Median distance between fixations for the three subject groups, and (C) Example of 30-seconds of gaze fixation data from one subject in each group (Red, blue and yellow colors each represent 10-seconds of successive gaze data while target is marked with a box in the background). Subjects who traverse the page in small steps (as can be seen from the gaze fixation locations) are able to find the target faster.

the mean) distance between successive fixations for the three groups of subjects was computed and plotted in Fig. 3.12B. The median distance between successive fixations tends to increase in the same manner for the three groups as the average time taken by them increases. Fig. 3.12C shows an example of 30-seconds of gaze data of one subject each from the three groups. As is clear from these figures, the subjects in Group 1 tend to search for the target in small sections of the page whereas the subjects in Group 3 search for the target randomly across the whole page.

Because subjects in Group 1 were able to find the target in the least time (on average), for "Where is Waldo?" game the best-observed strategy was to focus on small portions of the page rather than searching for the target randomly across the whole page. It would have been impossible to gauge this insight without the use of an eye-gaze tracker in which the subjects' gaze can be overlaid on the image from the world camera in addition to detecting the pupil's location with another camera.

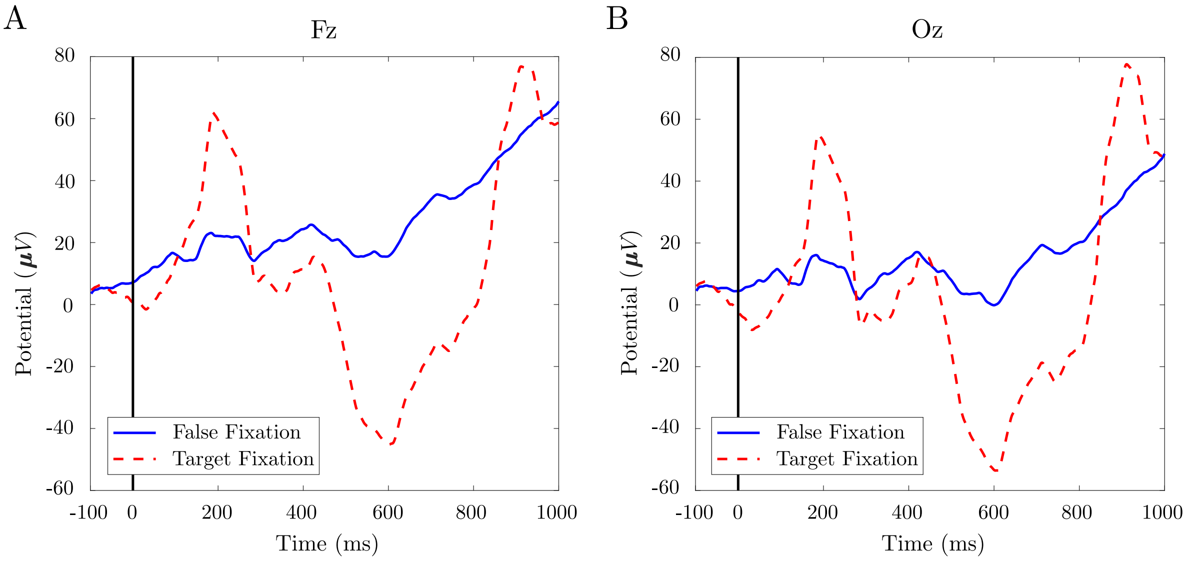


Figure 3.13: FRP plot of two EEG channel locations (A) Fz and (B) Oz. Target-fixated gaze FRP has much clear response than false-fixated gaze FRP.

3.4.2 Fixation-related Potentials (FRP) Analysis

EEG corresponding to fixations to the targets and non-targets while searching for the target in the "Where is Waldo?" game should represent distinct FRPs associated with the change in physiology. To discover this relationship, for each trial, the EEG data was band-pass filtered within the trial between 1-30 Hz. The mean of 200ms of data was taken before each fixation as the baseline and subtract it from one-second of post-fixation EEG data to remove the variations in amplitude. The FRP was calculated by averaging the data across the trials and subjects for all fixations greater than or equal to one second. Fig. 3.13 plots the averaged FRPs for Fz and Oz EEG channels. The FRPs show distinct variations after the onset of the fixations at zero seconds; the characteristic large peak at 200ms i.e. VPP and the trough between 200 and 400ms i.e. N2 are consistent with the earlier findings that VPP and N2 components are associated with the face stimuli [55, 56].

Furthermore, a large P3 response almost at the rightmost part of the plot is associated with decision making and is clearly much larger for the target than the non-targets [57]. This is understandable because while searching for the target's face, there are many non-targets with a similar face and clothing as the true target. When the subject first fixated on the true target, it takes time for him/her to assure it is indeed the true target. The slightly smeared nature of the P3 response is likely due to the fact that the latency of the P3 can vary across trials and individuals and the FPRs are time-locked to the onset of fixation, which is dependent on at what instant the fixation is detected by the algorithm since the pupil is continuously in motion. These results show that combining eye-gaze and EEG provides insight into the search patterns and their effects on the EEG in a visual-search task. Another significant aspect of this apparatus was to use the book in a naturalistic setup, which allows unconstrained head/body movements rather than asking the subjects to search the face in front of a computer screen with their head positioned on a chin-rest.

3.4.3 Variation across HRV during win/loss

HRV can be a reliable indicator of human emotions and mental states [25] and the aim of this study was to assess the variation across the subjects for the trials they won versus the ones they lost in the "Rock-Paper-Scissors" game. HRV was computed using the pNN50 algorithm [58] from 15-seconds of data for each trial. Fig. 3.14 shows the variation in HRV for all the subjects arranged by the number of trials won by them in ascending order. Based on the final score where Loss/Draw/Win corresponded to $-1/0/1$ points, subjects 4 and 5 lost the game, subject 8 tied the game and the remaining subjects won the game from the researcher. Fig. 3.14

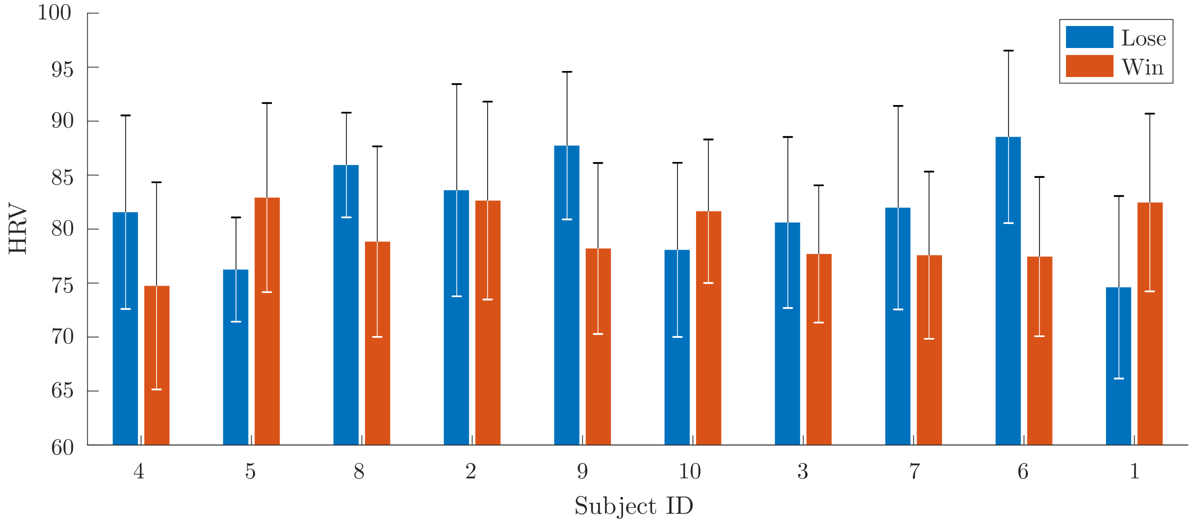


Figure 3.14: Mean HRV variation across subjects arranged in ascending order of number of trials won.

shows that for all the subjects except one (subject 2) there is an unmistakable difference between the values of averaged HRV for winning and for losing the trials, indicating that cardiac measures such as HRV can be helpful in distinguishing between physiological states corresponding to win vs. loss situations. Additionally, it was observed that different subjects respond differently to win/loss. For example, subjects 1, 5, and 10 show an increase in averaged HRV for the trials they won whereas others show a decrease in HRV. This might mean that different subjects react differently to the situation when they are winning or losing the game i.e. some might be enjoying the experience of the game whereas others might be under stress to come up with techniques to win the game.

3.4.4 Using machine learning to predict the results of the gaming

The aims of this evaluation are twofold. First, it was to study how well the bio-sensing modalities can predict i.e. classify the result of a game trial utilizing changes in physiology for a new subject, and what are their limitations in temporal frequency domain i.e. for how long the

Table 3.2: Modality Performance for Multi-modal Classification

Subject ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Mean**	**Max**	**Std.**
Classification Performance (Loss/Draw/Win) Chance Accuracy: 33%													
EEG (1-sec)	56	56	52	54	62	56	54	46	52	50	**53.80**	**62**	**4.26**
PPG (15-sec)	58	58	60	46	46	48	54	58	56	52	**53.60**	**60**	**5.32**
EEG + PPG (15-sec)	54	54	52	52	56	54	56	52	54	54	**53.80**	**56**	**1.48**
Classification Performance (Loss/Win) Chance Accuracy: 50%													
EEG (1-sec)	87.88	80.65	86.84	70.97	63.33	81.82	72.73	70.00	68.97	72.41	**75.56**	**87.88**	**8.21**
PPG (15-sec)	87.88	87.10	86.84	70.97	70.00	81.82	75.76	86.67	75.86	72.41	**79.53**	**87.88**	**7.30**
EEG + PPG (15-sec)	84.85	87.10	81.58	80.65	70.00	81.82	72.73	73.33	68.97	68.97	**77.00**	**87.10**	**6.92**

Leave one subject out validation was performed. All values denote percentage accuracy.

data from a modality is required. Second, there was a desire to document the potential benefits of combining features from different modalities in terms of accuracy and consistency. Thus, EEG and PPG were used to predict the outcome of a trial through these bio-sensing modalities. From the 500 total trials (50 each for 10 subjects), there were 137/183/180 trials for loss/draw/win respectively. The result of each trial was divided into two- (win/loss) and three (win/draw/loss) classes. Conditional entropy features [59] were calculated between each possible pairs of eight EEG electrodes. Hence, 28 EEG features were computed which were reduced to 20 components using Principal Component Analysis [60]. These features have been shown to work well for emotional valence and arousal classification [25]. Various time-durations were tried for the EEG data and it was found that 1-second of EEG data post-trial gave the best performance results. For PPG, in addition to the HRV features described above, six statistical features were computed[61]. The resultant seven features computed from 15 seconds of the PPG data were used for training the classifier. Leave-one-subject-out validation i.e. training data from 9 subjects and testing it for the remaining subject was then performed. Hence, since there are 50 game trials for each subject, 450 samples were used for training and 50 samples were used for testing. Extreme learning machines (ELM) with a single hidden layer were used for training the classification model [62].

Table 3.2 shows the results for 2-class and 3-class classification performance for each subject and both the sensor modalities. For all cases and subjects, it was observed that the classification performance i.e. mean accuracy is well above the chance accuracy level. The maximum accuracy goes up to 62% for three classes (loss/draw/win) and 87.88% for two classes (loss/win). Interestingly, PPG works as well as EEG for three classes and even better than EEG for two-class classification despite being only a single-channel signal with fewer features. However, PPG changes were slow (here 15 seconds of data being used) and thus do not provide as good temporal resolution as EEG. Hence, if the trials would have been spaced closer in time (say only 5 seconds apart), it would not have been possible to classify the trial result using PPG because of the inability to compute HRV and other cardiac features in such a short time window. Whereas, EEG can perform well on a shorter timescale but needs more channels. When both modalities were used (by taking 15-seconds of EEG data), it was found that the performance is not that much affected but the standard deviation significantly decreases. This means that using multiple modalities can help in producing consistent results across subjects because the fusion of features is able to compensate for the limitations of a single modality. Hence, using multiple modalities is good in both ways i.e. it gives the advantage to choose the modality as per the temporal resolution requirement or multiple modalities can be used together for more consistent performance across subjects.

3.5 Chapter Concluding Remarks

Bio-sensing technology is advancing rapidly both as a clinical research tool and applications in real-world settings. The existing bio-sensing systems are numerous and capable of measuring various physiological metrics in well-controlled laboratories. But, they are not practical for routine use by users in unconstrained real-world environments. Repeatedly, it has been shown that using multiple bio-sensing modalities improves the performance and robustness of decoding brain states and responses to cognitively meaningful real-life events. Hence, a research-grade wearable multi-modal bio-sensing system such as that developed in this research would allow researchers to study a wide range of previously unexplored research problems in real-world settings similar to the two gaming paradigms presented in this research work. Finally, this work presented a novel use of multiple bio-sensing modalities on “real-world” data while exploring previously unanswered questions in this area.

In the next chapter, multi-modal bio-sensors such as those designed in this chapter will be utilized to study the relationship between EEG, facial expressions, and affective multimedia content being consumed by a user. The chapter discusses the relationship between the physiological information provided by such biosensors and the audio-visual stimulus. Thus, the next part of the thesis moves on to study various physiological and stimuli-based emotional cues and is more closely connected with studying the initiation of human emotions.

This chapter is in part a reprint of material that has been accepted for publication in the journal IEEE Transactions on Biomedical Engineering (2018), by Siddharth Siddharth, Aashish N. Patel, Tzyy-Ping Jung, and Terrence J. Sejnowski. The dissertation author was the primary author of this paper.

Chapter 4

Assessing the Impact of Multimedia Content on the EEG and Facial Expressions

Having discussed the development of a multi-modal bio-sensing system, this chapter studies the impact of affective multimedia content on two of the most commonly-used modalities for emotion recognition, namely, EEG and Facial Expressions. This chapter presents insights from the correlation of affective features between three modalities namely, affective multimedia content, EEG, and facial expressions. Interestingly, low-level Audio-visual features such as contrast and homogeneity of the video and tone of the audio in the movie clips are most correlated with changes in facial expressions and EEG. The regions associated with the human face and the brain (in addition to the EEG frequency bands) that are most representative of affective responses respectively are also detected. Finally, the correlation between different layers of convolutional neural networks with EEG and Face images as input provides insights into human affection. Together, the findings of assist in (1) designing more effective multimedia contents to engage or influence the viewers, (2) understanding the brain/body bio-markers of affection, and (3) developing newer brain-computer interfaces as well as facial-expression-based algorithms to read emotional responses of the viewers.

4.1 Introduction

The design of multimedia content such as movies is driven by the director's assessment of the situation aimed at evoking a particular emotion. For example, action sequences evoking surprise tend to have frequent cuts between the scenes to keep the audience engaged while horror movies evoking fear are usually shot in dimly lighted surroundings to associate the circumstances with negativity. Such cues implicitly affect the audience's response to the content.

Overwhelmingly, past research aimed at emotion detection/classification has utilized electroencephalogram (EEG) and/or facial expressions [5, 6, 63]. In other cases, when such Audio-visual affective cues from the director's perspective are used in addition to the audience's facial expressions and/or EEG, the goal has been to boost the accuracy of emotion detection/classification framework [64]. Indeed, as a recent detailed survey about recognizing emotions using EEG demonstrates [65], even after the introduction of multi-modal affective datasets [5, 6, 7], the research has not translated from an emotion detection/classification problem to assess the correlation between the Audio-visual cues from the multimedia content and audience's associated response through changes in the EEG.

When it comes to understanding emotion elicitation and not just detection/classification, there are two parallel trends. First, past research has aimed at designing the optimal feature extraction techniques for mapping the audience's emotions to either audience's physiology or to facial expressions. For example, it has been shown that emotion elicitation can be assessed using frontal EEG asymmetry [66]. For facial expressions, there is a vast literature in assessing emotions using the facial action coding system (FACS)[67, 68] that weights the changes in different parts of the face. FACS is based on detecting the changes in facial action units (AUs) such as opening lips, widening eyes, raising eyebrows, etc., associating a score with each AU, and predicting the emotion using the sum of such scores by comparing to a predefined dictionary of possible scores. Such methods do not in any way account for the stimulus that the user is watching, this study goes a step further to associate AUs and other such features with the emotion-invoking Audio-visual cues present in the stimulus.

Second, researchers have studied emotion elicitation using films by recording and assessing the subjective feedback provided by the audience [69, 70]. However, the audience reports evoked emotions as per their perception of the multimedia stimulus. Their physiological recordings are not used to evaluate their responses. It is between these two trends that this research fits itself.

4.1.1 Contributions

- The Audio-visual cues that evoke the emotions in the audience are assessed and correlated with the audience's physiology. This is done by extracting feature cues from the multimedia content such as shot duration, visual excitement, color energy, audio tone, etc. that are known to evoke an emotional response in the audience. Canonical correlation analysis (CCA) [71] is then used between such cues and features from the audience's (neuro)physiology. In this way, a novel model is generated to represent the features of both worlds in a single domain.

- Subsequently, the analysis is performed on the model to provide insights into the regions from the face and brain (i.e. the audience's physiology) that are most correlated with emotions. The analysis is also performed to detect what type of Audio-visual cues are most effective in evoking changes in human physiology.

- Finally, this study correlates the components of the joint model generated from the audience's physiology and Audio-visual cues with the audience's recorded emotional response.

The insights obtained from this study will help in designing more affective multimedia contents to engage or influence the viewers, understanding the brain/body bio-markers of affection, and developing newer brain-computer interfaces as well as facial expression-based emotion classification algorithms to read emotional responses of the viewers.

4.2 Methods

This section details the dataset and the analytical tools used in the study.

4.2.1 Dataset

The publicly-available MAHNOB-HCI Dataset [7] was used for this study. The dataset contains multi-modal data i.e. frontal video, EEG, electrocardiogram (ECG). galvanic skin response (GSR), etc. from 27 users watching 20 short movie clips. These movie clips vary between 35 and 117 seconds in duration. Each clip is reported by the subject on a scale of integers from 1 to 9 to denote their arousal and valence according to the emotion circumplex model [72]. Additionally, the users also tag their felt emotion into one of the twelve categories such as anger, disgust, fear, joy, anxiety, etc. The clips have been chosen to evoke a certain emotional response in the audience. For example, clips from the *Mr. Bean* character and funny cat videos have been used to evoke amusement, clips from horror movies such as *Hannibal* and *Silent Hill* have been used to evoke disgust/fear response associated with horror, and clips from weather news coverage have been used to evoke a neural response in the audience.

The choice of this dataset was made because the dataset captures frontal videos at a high resolution of 780 × 580 pixels and EEG with a high-density headset (32 channels) at a high sampling rate (256 Hz). Furthermore, the dataset also provides the 20 movie clips that were used as stimuli for evoking the emotional responses in the audience. Since the dataset contains only about 540 trials, each fifteen-second section was windowed of the data and advanced it by one second. In this way, the dataset is augmented to get more than 34,000 trials for each data modality.

The sections below detail the cues that were extracted from each movie clip (stimulus) for the analysis with the audience's response. Eleven cues were extracted from the video and nineteen from the audio.

4.2.2 Audio Multimedia cues

Nineteen audio cues were extracted from each multimedia clip. Almost every such multimedia clip has background music playing with selected musical instruments to associate the scene with a particular emotional context. The presence of the human voice in the clip also provides another cue that some high-level information is being conveyed by communicating through the voice.

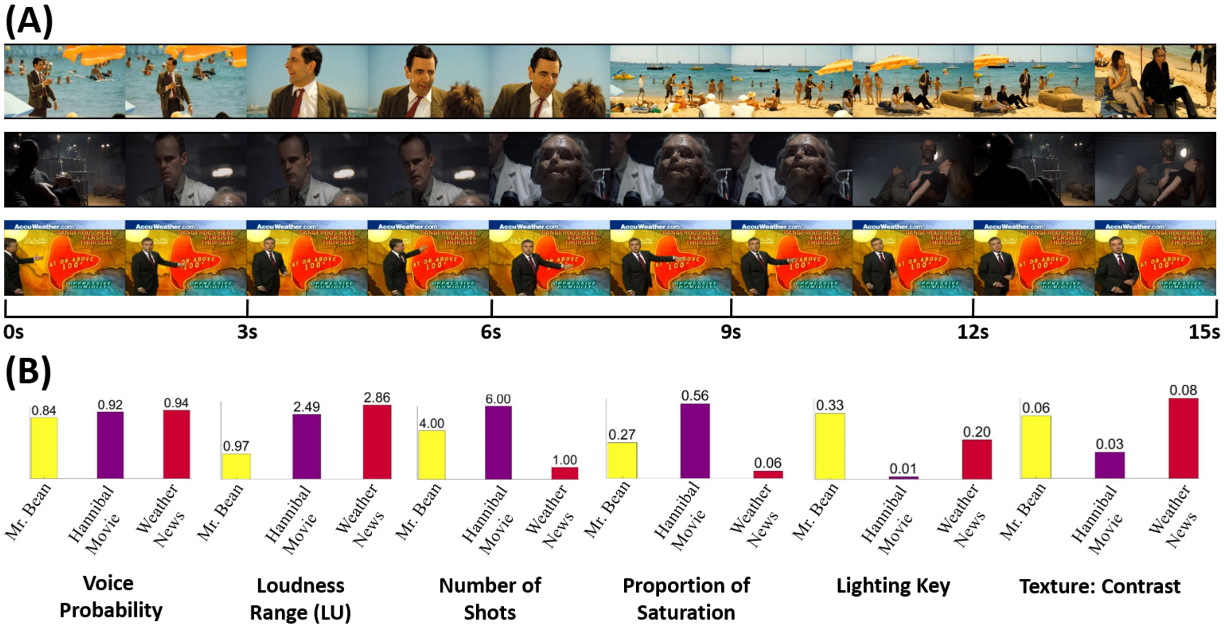


Figure 4.1: (A) 15-seconds examples (showing one frame for every 1.5 seconds) of three multimedia clips: *Mr. Bean* (above), *Hannibal Movie* (center), and *Weather News* (below) and (B) Six of the thirty Audio-visual cues are plotted to show the variation of features across the three clips.

MFCC Features: Acoustic characteristics in music are associated with the expression of emotions [73]. To assess spectro-temporal characteristics that represent the quality of sound making a particular type of music different from another, thirteen Mel-frequency cepstral coefficients (MFCC) were extracted as features. These features characterize the spectral shape of the sound and have been used in previous studies too for analysis with emotions [64]. To obtain these features, this study utilized the publicly available music information retrieval toolbox, MIRToolbox [74].

Loudness and Loudness Range: Loudness depends on the physical intensity of sound as well as the frequency and duration. Loudness, in general, is associated with emotional arousal as well as the perception of loudness itself can be influenced by the emotion [75]. Thus, the average loudness was computed for each audio clip as well as the loudness range (LU units).

Voice Probability: The presence of human speech in a clip is a high-level feature depicting that some information is being conveyed through communication in addition to the background sound. Thus, the probability of having human speech in each audio clip was calculated using a commonly-used statistical model.[76].

Pitch Features: To extract features related to the fundamental frequency of sound, pitch-based features were extracted. Pitch can model the harmonic as well as the melodic aspect of the music and is thus a good low-level feature. It has also been shown that the systematic changes in pitch level can affect the experience of pleasantness and arousal [77]. Thus, three features were extracted based on the pitch: keyclarity, mode, and harmonic flux [74].

4.2.3 Visual Multimedia cues

Eleven cues were extracted based on the video from each multimedia stimulus. These features capture low-level information such as texture and color as well as high-level information such as shot duration and visual excitement. These visual cues are generally used while producing movies and have been shown to capture affective information [78].

Visual Excitement: The amount of motion in a video plays a significant role in the human perception of the cinematic experience and affective response, particularly the amount of arousal [79]. Thus, visual excitement was calculated based on the average number of pixels that change between successive frames according to human perception [78].

Shot Duration Features: As mentioned before, the duration of the shot, also known as the pace of the movie, changes as per the type of content. Rapid changes in shots from multiple camera angles induce a degree of excitement far more effectively than a long duration shot [80]. Thus, the number of shots and average shot duration for each video clip were calculated using the open-source PySceneDetect tool [81].

Lighting Key Features: In the cinematographic perspective, lighting is an extremely powerful tool since it can be used to affect the audience's emotions by manipulating the mood of a scene. Dim lighting is generally used to convey negative themes in the content whereas an abundance of illumination generates joy and warm atmosphere around the scene [82]. The RGB color space frames were first converted to LUV color space for separating the illuminance information of each frame. Subsequently, the lighting key features were computed for each frame to measure the median general level of light and the proportion of shadow area in the clip [83].

Color Energy: Psychological studies on color have demonstrated a high correlation between valence and arousal with brightness and saturation of the colors respectively [84]. Color energy based on the brightness, saturation, and area occupied by the colors in a frame was calculated. This measure is defined as the product of the raw energy and color contrast [78, 82].

Texture-based Features: Visual detail has been shown to affect the audience's emotional distance to a scene such as by varying the camera distance of the shot [85]. Each video frame was transformed to generate the grey level co-occurrence matrix (GLCM) which models the distribution of co-occurring pixel values. This matrix can map the variations in texture such as gray-level contrast across the frame [86]. Four features on this GLCM matrix were then calculated and averaged across all the frames in a clip. These features were, namely, contrast, correlation, energy, and homogeneity. These features map local variations in GLCM, the joint probability of occurrence of the pixel pairs, the sum of squared elements of the GLCM, and closeness of elements distribution from the diagonal of the GLCM respectively. Finally, texture-based features were calculated based on the saturation of colors by calculating the proportion of pixels whose saturation in the normalized HSV color space exceeds 0.2 as the threshold.

Fig. 4.1 shows an example of calculating these Audio-visual cues through three 15s multimedia clips from the dataset. As one would expect, the voice probability is highest for *Weather News* clip but it has been filmed in only one shot whereas the lighting key is least (dim lighting) and saturation is the highest for the *Hannibal* movie clip i.e. horror scene.

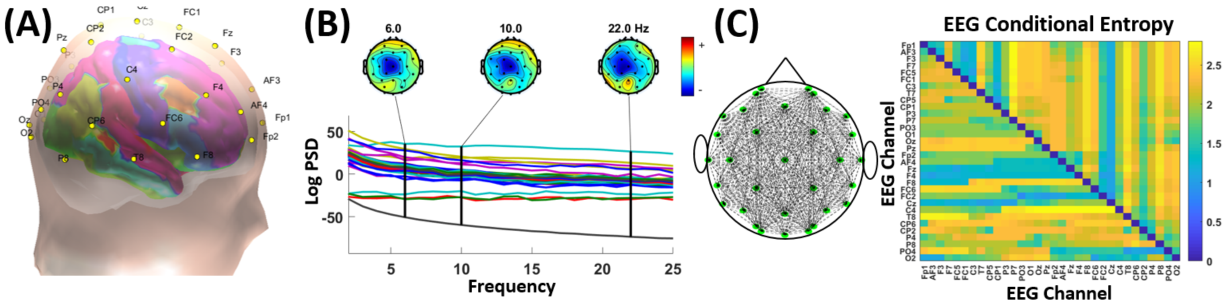


Figure 4.2: (A) EEG channel locations on the scalp used to compute local EEG-PSD Features, (B) An example of EEG-PSD heat maps calculated at three frequencies, and (C) Conditional Entropy pairs across the brain shown from the 2-D top-view of the scalp with the EEG Conditional Entropy feature matrix.

4.2.4 EEG Cues

The EEG data was first cleaned of various types of noise/artifacts such as due to eye blinks, muscle movements, electromagnetic disturbances, etc. using the Artifact Subspace Reconstruction (ASR) pipeline [87]. After cleaning the data, two distinct types of EEG features were extracted. The first kind of EEG features was the traditionally used power spectrum density (PSD) for each of the 32 EEG channels. The power spectrum density for three EEG frequency bands namely, theta (4-7 Hz), alpha (7-13 Hz), and beta (13-30 Hz) was then computed. These three EEG frequency bands were chosen because they account for most information towards human cognition [88] and thus have been most commonly used in affective research [65]. Other EEG bands such as the high-frequency gamma band (above 30 Hz range) were not used since the present literature does not support its association with human consciousness (highly associated with affective states) [89]. For each 15s clip, the EEG-PSD features were averaged across the whole clip. As a result, 96 EEG-PSD features were obtained for each multimedia clip.

The above EEG-PSD features are topographically localized and hence do not account for the variations in EEG across the human brain. Emotion elicitation is a complex process in which more than a single brain region may be involved. Previous research [90, 63] has shown that using features that capture changes in EEG across different brain regions can boost the performance for emotion classification. Thus, for each pair of EEG channels (32 EEG channels forming 496 such pairs), conditional entropy features [59, 90] were computed. These features were calculated based on the mutual information content between each sensor pair i.e. how much information is contained in one EEG channel given the information from another EEG channel. In this way, 496 features were extracted for each 15-second multimedia clip to represent changes across different regions of the brain.

Fig. 4.2 shows the 32 EEG channel locations that were used to calculate the localized EEG-PSD features. The 3-D head model was generated using the open-source headModel toolbox [91]. The figure also shows examples of PSD heat-map at three example frequencies, and the 496 channel pairs used to calculate conditional entropy and the conditional entropy matrix so calculated. To reduce redundancy, only the lower half of the matrix values i.e. one-way conditional entropy were used.

4.2.5 Facial Expressions Cues

For each face video clip, the user's face was first detected and extracted the face region from each camera frame using Haar-like features by utilizing the Viola-Jones object detector [92]. While running the face detection algorithm, the ends of the image were excluded and a threshold was placed of minimum face size to be 50 × 50 pixels to remove false positives. State-of-the-art automated Chehra algorithm [93] was then used to extract the face region to obtain 49 facial landmarks. These facial landmarks are located at the most expressive regions on the human face.

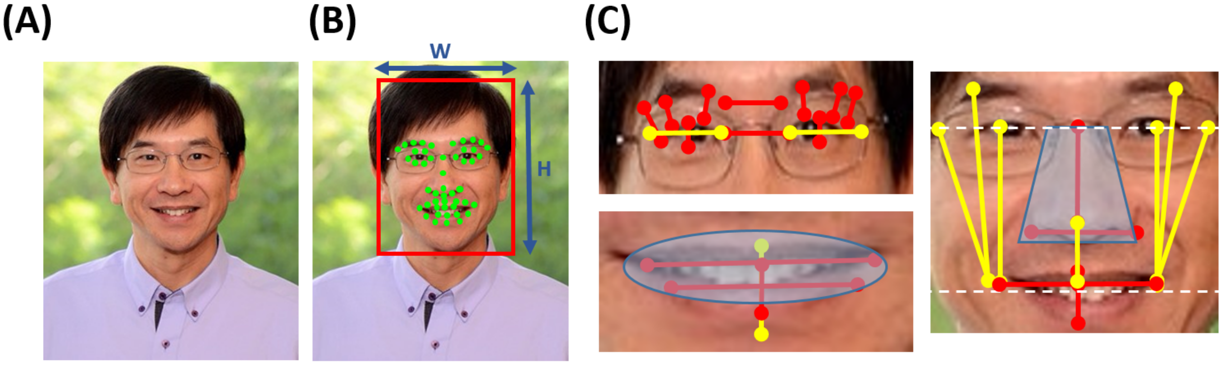


Figure 4.3: Illustrating the working of the developed algorithm through (A) a sample image frame, (B) extracted face region and locating facial landmarks, and (C) 30 distance/area features computed based on the facial landmarks. To account for different face sizes and distance from the camera, all features were re-scaled based on face width (W) and face height (H). A sample image was used to generate this figure in order to respect the privacy of the users who participated in the study.

These 49 facial landmarks were used to calculate 30 features based on the distance and area calculated using these landmarks such as the vertical distance of the eye-opening, the distance between the eyebrow to the lips, horizontal and vertical distances between the ends of the lips, the area of the mouth and nose, etc. Fig. 4.3 shows these facial landmarks and features. Most of these features are the same which are used in the calculation of facial action units (AUs) [68] and thus have been used for emotion recognition through facial expressions. Some other features were selected by hand. Because the shape of the face varies across human beings and the area taken by the facial region in an image varies with distance from the camera, all such features were normalized based on the width and height of the detected face. Since each multimedia clip contains many frames, wavelet decomposition across all face images [94] was first used to fuse all such face images to one face representation. It is on this image which contains the "representative" facial expression across the multimedia clip that the above features were computed.

4.3 Results

4.3.1 Relationship between Audio-visual Cues and EEG Cues

Canonical correlation analysis (CCA)[71] is a method for exploring the relationship between two multivariate sets of vectors. This method can be used to compute which variables and their combinations in a set are most correlated with those in another set. Since all the feature cues are correlated with human emotions, it helps in perceiving which of those cues contribute the most towards emotion elicitation and classification. CCA was then first used to find the relationship between changes in the EEG cues (theta, alpha, and beta EEG band power-spectrum density and conditional entropy across brain regions) and over $34,000$ instances of features from the Audio-visual cues representing human emotion. These EEG and Audio-visual cues are detailed in the Methods section below and they capture the low- and high-level information from the multimedia content. Specifically, low-level cues are those which are not directly perceived consciously by humans such as the texture of video content and pitch of the sound while high-level cues are the ones which are associated with human cognitive capabilities such as speech recognition and how visually exciting a scene is based on the temporal speed of the content. The method was applied to the data from each subject separately and the aim was to find the leading Audio-visual cues that modulate the EEG spectra. This would help in pinpointing the kinds of stimulus that strongly affects human emotions.

Fig. 4.4 shows the corresponding EEG coefficients for each of the three frequency bands for the first three most correlated Audio-visual cues. It is notable that among the 30 Audio-visual cues, the three most correlated cues are low-level ones i.e. based on video texture and audio pitch rather than high-level ones such as the number of shots and voice probability. These texture- and pitch-based features usually model the background imagery and sound in multimedia contents. The corresponding brain regions connected by 10 most correlated conditional entropy cues are also shown. It is notable that the most active areas in the brain are present across the C3, Cz, and

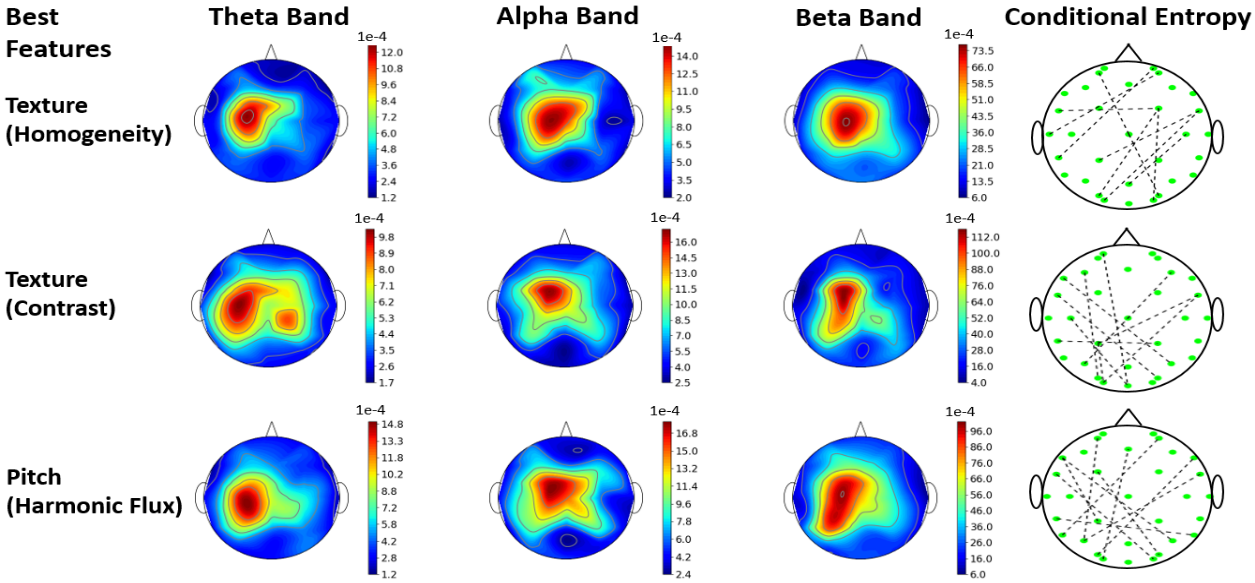


Figure 4.4: The three leading Audio-visual Cues, Texture-based Homogeneity, Texture-based Contrast, and Pitch-based Harmonic Flux, associated EEG spectral changes in the theta, alpha, and beta bands, and top 10 Conditional Entropy EEG channel pairs.

C4 electrodes i.e. in the central section of the 2D brain representation. Finally, looking at the values on the three color-bars, it is observed that beta-band was affected by the Audio-visual cues much higher than the theta and alpha bands. The p-values calculated by the paired-sample t-test for Texture (Homogeneity) for theta-alpha, alpha-beta, and theta-beta pairs at the Cz electrode (located at the center of the 2D brain image) were 0.16, 0.002, and 0.005 respectively, for Texture (Contrast) were 0.178, 0.00008, and 0.0001 respectively, and for Pitch (Harmonic Flux) were 0.16, 0.00015, and 0.00018 respectively. These statistical tests showed the significance of beta power being statistically different from theta and alpha ones whereas Fig. 4.4 shows the beta power to be significantly larger than the theta and alpha power. The results suggest that the beta-band power in EEG recordings might be used for distinguishing affective states to a significantly more extent than the theta and alpha EEG band power.

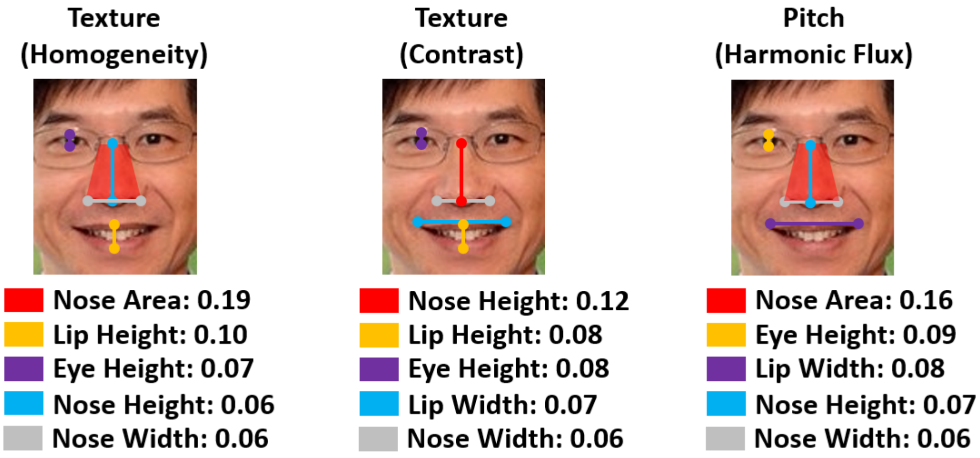


Figure 4.5: The same three Audio-visual Cues as with the EEG (Texture-based Homogeneity, Texture-based Contrast, and Pitch-based Harmonic Flux) were found to be most correlated with Facial Expressions. The above five best associated Facial Expressions Cues are plotted and weights of their associated CCA coefficients are denoted.

4.3.2 Relationship between Audio-visual Cues and Facial Expression Cues

Similar to the EEG cues above, CCA was performed between Audio-visual cues and Facial Expression Cues. By this process, the aim was to find which cues from each of these modalities are most correlated with each other. Similar to the EEG, it was found that among the 30 Audio-visual Cues, texture-based and pitch-based low-level cues affect facial expressions most. Fig. 4.5 shows the five most correlated Facial Expression cues corresponding to each of the leading Audio-visual Cues. In all three cases, the nasal region is the most representative part of the human face. This is because either the area of the nose or the height of the nose which modulates in a 2D camera image as per vertical bending of the human face contributes significantly more than other facial regions. After the nasal region, lips and eye-opening contribute most towards the Audio-visual Cues.

4.3.3 Relationship between EEG Cues and Facial Expressions Cues

Having found the correlation between the multimedia contents that the users are watching and their physiological responses, this study sought to find the relationship between their physi-

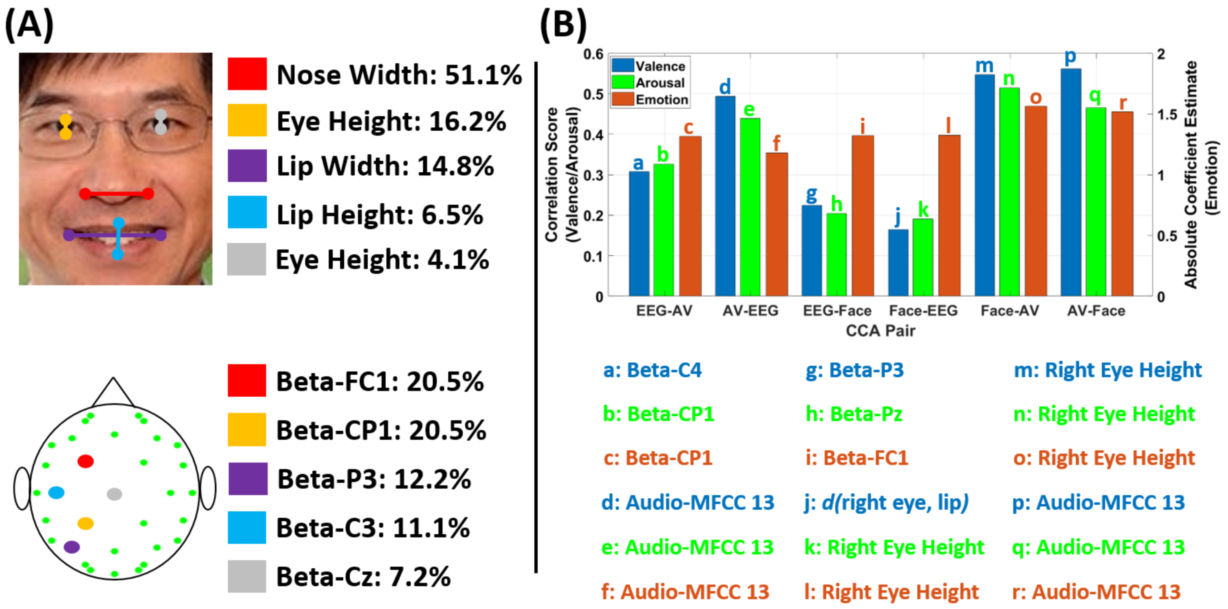


Figure 4.6: (A) (Top) Five most contributing Facial Expressions Cues (note that the width of nose alone contributes towards half of the total) and (Bottom) Five most contributing EEG Cues (note that all contributions are by EEG Beta-band Power only) and (B) (Top) Correlation score between each CCA pair (EEG-AV denotes the correlation between EEG projected from the joint EEG-AV space and Valence/Arousal (Left y-axis) or Emotion (Right y-axis)) and (Bottom) the cue with the largest correlation score for each bar graph. *d(a,b)* denotes the Euclidean distance function.

ological response (EEG) and behavior (Facial Expressions). To this end, CCA was applied to the EEG Cues and Facial Expression Cues. Fig. 4.6(A) shows the results of this analysis. As a compilation of the above two results, it was again observed that the nasal region contributes most towards the Facial Expression Cues, followed by eye height and lip width and height. Similarly, all five most contributing EEG Cues are from the Beta-band EEG power in the left central areas. This is also consistent with previous results which have shown that this brain region represents human valence and arousal very well. [63].

4.3.4 Correlation between the Audio-visual, EEG, and Facial Expressions Cues with subjective responses

As discussed in the Methods section, for each multimedia clip, the subject reported his/her valence (between 1 to 9), arousal (between 1 to 9), and emotional tag (one of twelve emotions). After calculating the CCA projection between each pair of the three kinds of feature cues (Audio-visual, EEG, and Facial Expressions), the joint CCA space was projected back to the feature cues. For Arousal/Valence, the Pearson correlation was then calculated between the CCA projections and behavior data (user-reported labels) while for Emotions, a linear regression model[95] was fitted between CCA projections and behavior data (user-reported labels). This is done differently for these three affective measures since Valence/Arousal are distributed on a scale of 1 to 9 denoting their intensity whereas Emotions were tagged by the users directly to be one of the many categories. Hence, a linear regression model was fitted on users' reported emotions and plot the highest absolute value of the coefficient estimate. In this way, for Valence/Arousal/Emotion, it was able to plot how closely did the CCA components method developed in this study correlated with users' behavior data.

Fig. 4.6(B) shows that the projections along with the top CCA components between Facial Expression Cues with Audio-visual (AV) Cues providing the highest correlation with the behavior data (two rightmost sets of bars). Below that plot, the figure shows which cue in each of the three modalities provides the highest correlation with the user-reported affective measures. Each set of bars such as Face-AV and AV-Face denote the projection of either Facial Expressions Cues and Audio-visual Cues from the joint Face-AV Cues space that was used to calculate the correlation with behavior data. For each such correlation, the figure also lists that which of the cues from that modality provided the highest correlation.

Thus, it was observed that EEG beta-band power near the centro-parietal region (e.g. C4, Cp1, etc.) has the highest correlation with behavior data (bars labeled as a, b, and c), while

MFCC feature 13 i.e. the last audio MFCC coefficient provides the highest correlation among the Audio-visual Cues with emotional responses (the bar labeled as d). This MFCC coefficient is related to very fine pitch and tone information since it is the highest coefficient among those generally used in speech processing research. Finally, among the Facial Expression Cues, the height of the right eye i.e. the vertical distance of the eye-opening (bars labeled as k-o) has the highest correlation with user-reported affective measures in most cases. Hence, this part of the analysis bridges the gap between the cues from the three modalities that were analyzed and users' self-reported affective measures (behavior data).

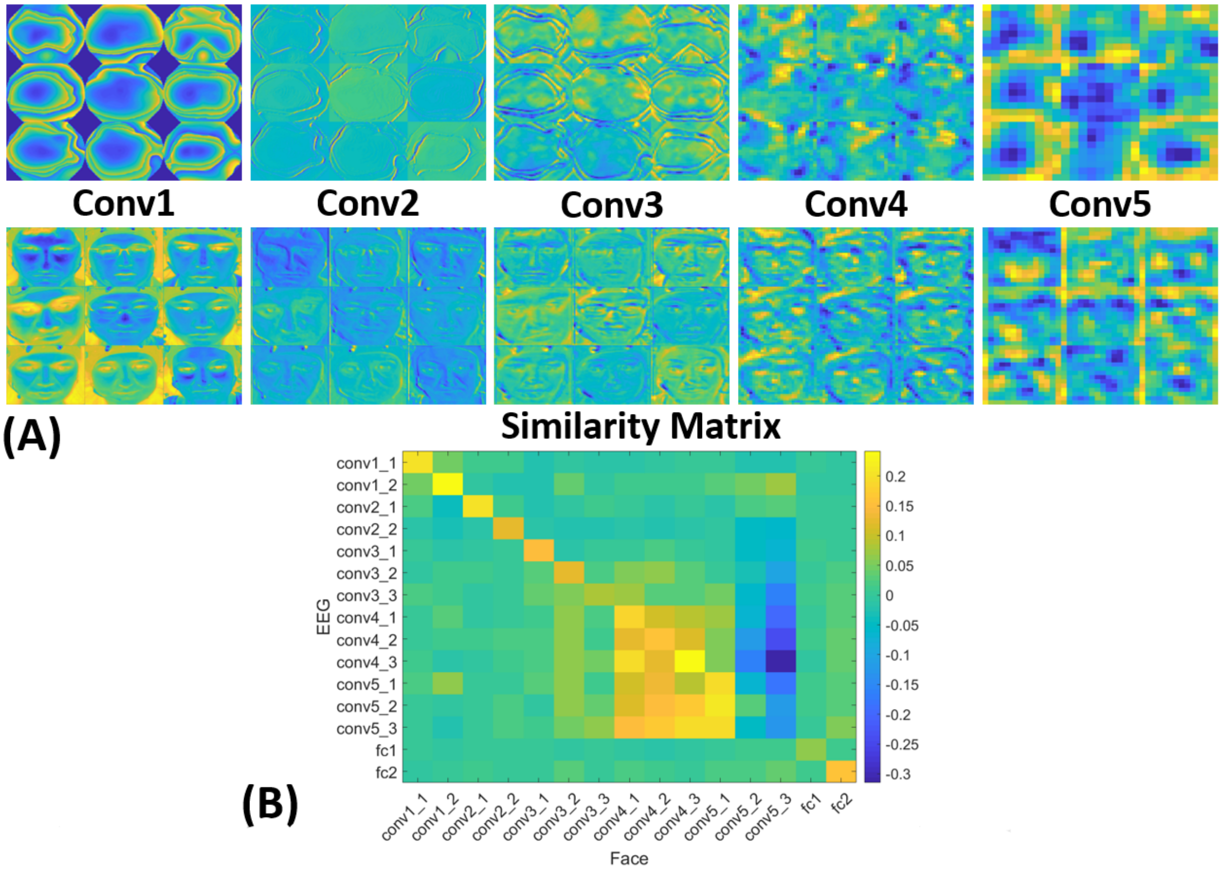


Figure 4.7: (A) A sample feature representation of different layers' activations in the VGG-16 network for nine example EEG and Face data inputs and (B) A similarity Matrix shows the correlation at each layer between EEG and Face representations in the VGG-16 network.

4.3.5 Correlation between EEG and Facial Expressions through Deep Learning Network Cues

All the Cues that were analyzed above were extracted by the methods reported in previous academic research or were designed by hand. But, with the advent of deep learning research in previous years, it is now possible to let the convolution networks decide which feature cues to extract and use for a particular application. To this end, a recently developed technique was utilized to convert time-domain EEG data into image-based representation[63]. Utilizing this method, a single (RGB) color image representative of EEG power in all three frequency bands (theta, alpha, and beta) was generated for each of the 15s multimedia clips. Similarly, an image fusion[94] method was used to represent the face region in successive frames with a single frame.

While deep learning research on computer vision data has a wide breadth of research, the knowledge from two very recent publications detailing the use of deep learning for EEG [11, 96] was utilized to train a VGG-16 [97] network-architecture-based model. This network consists of 16 weight layers with millions of parameters and has been shown to work very well for various image classification problems. This analysis randomly used 90% of the total trials for training the network and remaining for testing the activations. Similarly, another VGG-16 model was trained on face images from the affective dataset. The goal behind training these networks was to extract features varying from low to high level that is not possible for humans to visualize and design.

Fig. 4.7(A) shows nine sample inputs to the network and the activations of five successive convolution layers. It can be surmised from the figure that as the number of layers increases, the activations' representation inside the network becomes more low-level i.e. higher resolution. Thus, going through all layers of the network, it is possible to capture the full variation in such feature representation. For each such layer for all input images, a representation similarity matrix (Fig. 4.7(B)) was generated between EEG and Face images similar to what is done for Magnetic Resonance Imaging (MRI) data[98]. The figure shows the correlation between 15 of the 16 layers' activations of the VGG-16 matrix (since the last layer corresponds to classification and hence was discarded). The highest correlation in the similarity matrix is located in the middle layers, which means that those layers (conv4 and conv5) represent the activations that are highly correlated in the EEG- and face-representation learned by the deep learning network. It is also notable that because Conv4 and Conv5 layers are just before the fully connected ones, they contain the most discriminative features as one progresses from Conv1 to fc2 layers. This would mean that such low-level discriminative features could be more useful when utilized for affective analysis just like this study found that low-level Audio-visual features were more discriminative than high-level ones in influencing users' EEG and facial expressions.

4.4 Chapter Concluding Remarks

Past research in affective computing has mostly focused on emotion classification using various Audio-visual or bio-sensing modalities[5, 6]. A severe limitation of such research has been its inability to boost classification accuracy to human-like levels. Another limitation of previous research has been its inability to demonstrate which cues related to Facial Expressions or EEG are most correlated with human emotions. This is because FACS-based facial action units [68] are hand-coded cues that were chosen since they represent specific muscle movements or set of muscle movements and are by definition mapped to particular emotions. Similarly, the most commonly used EEG power spectrum features have been used blindly for emotion classification without any insight into which brain regions or frequency bands actually contain the most affective content. It was to address these limitations that this study sought to understand the relationships between multimedia content itself with users' physiological responses (facial expressions/EEG) to such content.

By the means of Canonical Correlation Analysis (CCA)[71], these three modalities were analyzed after extracting cues representative of affective information from each of them. It was found that for both EEG and facial expressions, the Audio-visual cues that are most correlated with them are those that represent low-level features found in the image background/background music such as texture and pitch. This is a useful insight into the design of such content to intentionally customize the content's image background and background music for invoking particular affective responses.

The other side of the coin from the above analysis is that most of the spectral features from the EEG are concentrated at/around the Cz electrode position (center of the 2D brain image) and are present in the beta frequency band. This is an insight that can help design the next generation of wearable EEG headsets for affective applications since focusing on these brain regions and frequency band can provide most of the affective information and thus sensors covering the whole brain may not be needed. Similarly, for facial expressions, it was found that the nasal region followed by eye-height i.e. how much the user opens his/her eyes, and lips are most representative of changes in affective multimedia contents. Hence, new features can be designed and added to the FACS method to utilize more information from these three facial regions only (for example, the nasal area and nose height representing vertical head tilt in a 2D image are not currently used as facial AUs). This might point to the fact that the changes in features across different analyses (e.g. nasal region vs right eye height as the most correlated features) take place because different kinds of affective contents may induce different affective states in human physiology. Thus, it is not possible for a single brain region or facial region to be universally representative of every affective state of the subject[99].

When the relationship between facial expressions and EEG was quantified, it was again found that the nasal region followed by eye-height and lips are most representative of changes in facial expressions while the beta-band power distributed at/near the central region is most representative of changes in the EEG while the participants watch the multimedia contents. Similarly, previous research has shown that beta-band activity reflects emotional and cognitive processes very well[100]. In fact, the top five features in facial expressions and EEG alone constitute more than 90% and 70% weightage respectively compared to all possible 30 facial expression cues and 96 EEG cues. It suggests that utilizing only these five cues from both modalities can provide much of the affective information. Finally, the high EEG-activity at/near the center of the brain was also consistent with previous work that showed these brain regions are also most representative of changes in human valence and arousal[63].

When the joint CCA space of these modalities was projected back to reconstruct the cues from each modality and was correlated with users' subjective valence/arousal/emotion ratings, it was found the highest correlation (with values above 0.5 even when valence/arousal ratings have a high resolution since they are divided on 9/9 point scale respectively) for Audio-visual cues followed by that for facial expressions cues. In a way, this is intuitive since the Audio-visual cues extracted from the multimedia content itself should contain most of the affective information. Subsequently, this also demonstrated that the extracted Audio-visual cues do represent the affective information as perceived by the users. Facial expressions cues are second-most correlated whereas EEG cues are least correlated with users' responses. This is probably due to the fact that image data does not have as much noise/artifacts as are present in EEG data. It was again found that low-level Audio-visual cues contribute the most towards high correlations with users' subjective ratings. This shows that in such a multimedia stimuli context, human emotions are affected much more by "background" cues in this context such as pitch and texture of the scene rather than "foreground" human speech and how fast the video content changes from frame to frame.

Deep learning networks utilize many convolution layers to extract features that can be characterized by an increasing level of complexity. Such a deep convolution network[97] was used to extract features of various complexity from EEG and face data. A similarity matrix was then generated to represent the correlations between each pair of convolution layers between the modalities. It was found that the highest correlation was present at the fourth and fifth convolution layers. This points out that the information from raw EEG and face data has to be first sufficiently processed to generated these “optimal” low-level features since they are most representative of joint changes in the two modalities while watching affective multimedia contents. These results are presented as a starting point to take this research further into actually interpreting such low-level features extracted by the convolution network and then use them for emotion elicitation and classification problems.

To conclude, the findings of this study provide various insights into the affective cues of these three modalities, both for designing affective multimedia content as well as for designing the next generation of EEG- and vision-based emotion classification systems. Through the above-mentioned analysis, this study shows that low-level Audio-visual cues are most representative of human emotions, even when they may be subconsciously influencing human emotions.

The next chapter shows the applications of insights from this chapter toward emotion recognition when users are watching emotional multimedia content. In that chapter, various signal processing and deep learning-based algorithms are designed for different bio-sensors that were evaluated in this chapter. These algorithms are designed for applications in affective computing, specifically for the classification of human emotions while participants are engaged in a task with an audio-visual stimulus. Thus, from developing hardware multi-modal bio-sensing systems and studying the elicitation of emotions, we now move toward the applications of such hardware and software frameworks.

This chapter is in part a reprint of material that has been accepted for publication in the journal Nature Scientific Reports (2019), by Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski. The dissertation author was the primary author of this paper.

Chapter 5

Utilizing Deep Learning on Multi-modal Bio-sensing for Emotion Classification

Having detailed the relationship between EEG, facial expressions, and affective multimedia content in the domain of emotion elicitation in the previous chapter, this chapter presents the development of novel algorithms to various bio-sensing and video data of four publicly available multi-modal emotion datasets for emotion classification. For each dataset, the emotion-classification performance obtained by each modality is first evaluated. Subsequently, the performance obtained by fusing the features from these modalities is evaluated. It is then shown that these algorithms outperform the results reported by other studies for emotion/valence/arousal/liking classification on DEAP and MAHNOB-HCI datasets and set up benchmarks for the newer AMIGOS and DREAMER datasets. This work also evaluates the performance of the developed algorithms by combining the datasets and by using transfer learning to show that the proposed method overcomes the inconsistencies between the datasets. Hence, a thorough analysis of multi-modal affective data from more than 120 subjects and 2,800 trials has been undertaken. Finally, utilizing a convolution-deconvolution network, a new technique towards identifying salient brain regions corresponding to various affective states has been proposed.

5.1 Introduction and Related Research

In recent years, there has been growing interest towards approaching research in affective computing from a bio-sensing perspective. To be sure, it is not just in affective computing that research in bio-sensing has been gaining popularity. Other avenues of research such as health [21, 101], virtual reality [102], robotics [23, 103], content rating [104], etc. have also exploited bio-sensing as a research tool. Bio-sensing systems specifically those which are used to measure electrocardiogram (ECG), electroencephalogram (EEG), galvanic skin response (GSR), etc. have been around for decades. But, because of their bulkiness and complexity, they were restricted to controlled laboratory environments and hospitals. The current interest in utilizing bio-sensing systems for various applications has been motivated or driven by the development of wearable bio-sensing systems that make data collection faster and easier [36, 35, 105]. The advances in hardware have led to the further development of multi-modal bio-sensing systems i.e. those capable of monitoring and recording multiple bio-signals simultaneously [30, 26, 43, 90].

Many research studies have shown that it is possible to recognize human emotions by the use of facial expressions from images and videos [68, 106, 107, 108]. Advances in deep learning have also made it possible to train large neural networks on big datasets for research in affective computing [109, 110, 111] apart from other problems such as object detection and classification [49, 112, 113]. Compared to the amount of deep learning research that has translated towards solving problems involving images/videos, the deep learning research conducted on bio-sensing data has been sparse. A recent survey on using EEG for affective computing [65] suggests that in almost all cases the feature extraction and classification steps do not utilize deep neural networks.

There are chiefly three reasons limiting the use of deep learning to bio-sensing modalities. First, it is easier to create an image/video database by collecting a huge amount of image/video data with any decent camera (even that of a smartphone) whereas the data collection of bio-signals is often costly, time-consuming, and laborious. Second, the image/video datasets generated using different cameras are usually consistent or can easily be made so by changing the frame resolution or modifying the number of frames being captured per second without losing critical information in the process. On the other hand, commercially available bio-sensing devices vary widely in terms of sampling rate, analog to digital resolution, numbers of channels, and sensor positioning [105, 65]. Furthermore, there are differences in the signal profiles between different types of bio-sensing signals such as EEG vs ECG. Third, visualizing image data for object detection or assessing emotions by looking at faces/body postures in the images is much easier (such as for manual tagging) and intuitive. But, extracting meaningful knowledge about various features from bio-sensing signals requires pre-processing. Unlike image data, additional steps are required in bio-sensing data to first filter the data of any noise such as due to motion artifacts or unwanted muscle activity.

Using multiple bio-sensing modalities can be advantageous over using a singular one because the salient information in the respective modalities may be independent of and complementary to each other to some extent. Thus, together they may enhance the performance for a given classification task [114]. In most cases, the emotion-classification problem has been approached by measuring the arousal and valence as given by the emotion circumplex model [72]. It is evident from various studies [65, 14, 115] that a single modality may perform differently for arousal and valence classification. So, in theory, two modalities that show good performance independently for valence and arousal respectively may perform even better jointly for the emotion classification problem.

5.1.1 Contributions

- This study focuses on multi-modal data from both bio-sensing and vision-based perspectives. The features with deep learning-based methods and traditional algorithms are fused for all modalities on four different datasets.

- This study shows that using multi-modality is advantageous over singular modalities in various cases. It also shows that deep learning methods perform well even when the size of the dataset is small.

- For each of the four datasets, this study shows that the developed methods outperform previously reported results.

- The results of this study also demonstrate the applicability of deep learning-based methods to overcome the discrepancies between different modalities and even effectively fuse the information from them, as shown by results from combining the datasets and transfer learning.

5.1.2 Related Research Studies

The developed framework was designed and evaluated on four publicly available multi-modal bio-sensing and vision-based datasets namely DEAP [5], AMIGOS [6], MAHNOB-HCI [7], and DREAMER [8]. Table 5.1 briefly describes the four datasets, with a focus on the modalities that were used in this study. For DEAP and AMIGOS datasets, the preprocessed bio-sensing data that has been suitably re-sampled and filtered was used whereas for MAHNOB-HCI and DREAMER datasets, filtering and artifact removal were performed before extracting features. The trials in the DEAP and AMIGOS datasets have been tagged by subjects for valence, arousal, liking, and dominance on a continuous scale of 1 to 9. For MAHNOB-HCI and DREAMER datasets, the valence and arousal have been tagged on a discrete scale using integers from 1 to 9 and 1 to 5, respectively. This study used the emotion circumplex model [72] to divide the emotions into four categories namely, High-Valence High-Arousal (HVHA), Low-Valence High-Arousal (LVHA), Low-Valence Low-Arousal (LVLA), and High-Valence Low-Arousal (HVLA). These categories loosely map to happy/excited, annoying/angry, sad/bored, and calm/peaceful emotions,

Table 5.1: Table Highlighting the Inconsistencies Among the Datasets and Sensing Modalities

DEAP Dataset [5]	**AMIGOS Dataset [6]**	**MAHNOB-HCI Dataset [7]**	**DREAMER Dataset [8]**
32 subjects	40 subjects	27 subjects	23 subjects
40 trials using music videos (trial length fixed at 60 seconds)	16 trials using movie clips (trial length varying between 51 and 150 seconds)	20 trials using movie clips (trial length varying between 34.9 and 117 seconds)	18 trials using movie clips (trial length varying between 67 and 394 seconds)
Raw and pre-processed data available	Raw and pre-processed data available	Only raw data available	Only raw data available
32-channel EEG system (Two different EEG systems used. Channel locations: Fp1, AF3, F7, F3, FC1, FC5, T7, C3, CP1, CP5, P7, P3, Pz, PO3, O1, Oz, O2, PO4, P4, P8, CP6, CP2, C4, T8, FC6, FC2, F4, F8, AF4, Fp2, Fz, Cz)	14-channel EEG system (A single EEG system used for all subjects. Channel locations: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4)	32-channel EEG system (A single EEG system used for all subjects. Channel locations: Fp1, AF3, F7, F3, FC1, FC5, T7, C3, CP1, CP5, P7, P3, Pz, PO3, O1, Oz, O2, PO4, P4, P8, CP6, CP2, C4, T8, FC6, FC2, F4, F8, AF4, Fp2, Fz, Cz)	14-channel EEG system (A single EEG system used for all subjects. Channel locations: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4)
—	2-channel ECG system	3-channel ECG system	2-channel ECG system
1-channel PPG system	—	—	—
1-channel GSR system	1-channel GSR system	1-channel GSR system	—
Sampling rate 128 Hz	Sampling rate 128 Hz	Sampling rate 256 Hz	Sampling rate EEG/ECG: 128/256 Hz
Face video recorded for 22 of 32 subjects (EEG cap and EOG electrodes occludes parts of the forehead and cheeks)	Face video recorded for all subjects (Only a small portion of the forehead is occluded by the EEG system)	Face video recorded for all subjects (Only a small portion of the forehead is occluded by the EEG system)	—
3-seconds of pre-trial baseline data available.	No baseline data available.	30 seconds of pre-trial and post-trial baseline data available.	61 seconds of pre-trial baseline data available
Valence/Arousal/Liking rated using a continuous scale between 1 to 9	Valence/Arousal/Liking rated using a continuous scale between 1 to 9	Valence/Arousal rated using a discrete scale of integers from 1 to 9	Valence/Arousal rated using a discrete scale of integers from 1 to 5

respectively. For each dataset, the labels were self-reported by the subjects after the presentation of the video stimuli.

As shown in Table 5.1, the datasets differ in many aspects. Hence, many traditional algorithms cannot be generalized across datasets because of differences in the number and nature of extracted features. Apart from the types of the audio-visual stimulus (music videos vs. movie clips), the datasets vary in the trial duration and baseline data availability. The DEAP dataset has trial length fixed at 60 seconds whereas for the AMIGOS dataset, the trial length varies between 51 to 150 seconds. This varying trial length is significant since the longest video is about thrice the length of the shortest one. Hence, if a particular emotion of the subject is invoked for 25 seconds during a trial, it will appear in half of the trial in the shortest video but only in one-sixth of the trial in the longest one. Furthermore, the trial length variation of the DREAMER dataset is even greater than that in the AMIGOS dataset. There is also no baseline data present in the AMIGOS dataset to compensate for the subject's initial emotional power (defined as the distance from the origin in the emotion circumplex model). Different kinds of systems have been used to collect the EEG data in these datasets. 32-channel EEG in the DEAP and MAHNOB-HCI datasets may contain much more emotion-relevant information than the 14-channel EEG in the AMIGOS and DREAMER datasets.

Only the DEAP dataset uses photoplethysmogram (PPG) to measure heart rate instead of ECG. The use of PPG generally loses the information that is present in the ECG waveform such as QRS complex, PR, and ST segment lengths, etc.

The EEG electrodes introduce varying degrees of occlusion while capturing frontal videos of the subjects. This effect was found to be more problematic in the DEAP dataset because of the placements of the EOG electrodes on subjects' faces. Furthermore, some data are missing in some modalities in a subset of the trials of these datasets. Such entries were ignored in the evaluation pipeline.

Table 5.2: Classification Performance and Evaluation Performed by Various Reported Studies on Four Multi-modal Datasets

Study	Used Modalities	Extracted Features	Classifier	Evaluation
DEAP Dataset				
Liu et al. [14]	EEG	Fractal dimension (FD) based	SVM	Only 22 of the 32 subjects used. 50.8% Valence (4-classes) and 76.51% Arousal/Dominance.
Yin et al. [16]	EEG, ECG, EOG, GSR, EMG, Skin temperature, Blood volume, Respiration	Various	MESAE	77.19% Arousal and 76.17% Valence (2-classes) using fusion of all modalities.
Patras et al. [5]	EEG	PSD	Bayesian Classifier	62% Valence and 57.6% Arousal (2-classes)
Chung et al. [12]	EEG	Various	Bayesian weighted-log-posterior	70.9% Valence and 70.1% Arousal (2-classes)
Shang et al. [116]	EEG, EOG, EMG	Raw data	Deep Belief Network, Bayesian Classifier	51.2% Valence, 60.9% Arousal, and 68.4% Liking (2-classes)
Campos et al. [13]	EEG	Various	Genetic algorithms, SVM	73.14% Valence and 73.06% Arousal (2-classes)
AMIGOS Dataset				
Miranda et al. [6]	EEG, ECG, GSR	Various	SVM	*57.6/53.1/53.5/57 Valence and 59.2/54.8/55/58.5 Arousal (2-classes) using EEG/GSR/ECG alone/EEG, GSR, and ECG fusion.
MAHNOB-HCI Dataset				
Soleymani et al. [7]	EEG, ECG, GSR, Respiration, Skin Temperature	Various	SVM	57/45.5/68.8/76.1% Valence and 52.4/46.2/63.5/67.7% Arousal (2-classes) using EEG/Peripheral/Eye gaze/Fusion of EEG and gaze.
Koelstra et al. [15]	EEG, Faces	Various	Decision classifiers fusion	73% Valence and 68.5% Arousal (2-classes) using EEG and Faces fusion.
Alasaarela et al. [117]	ECG	Various	KNN	59.2% Valence and 58.7% Arousal (2-classes)
Zhu et al. [118]	EEG and Video stimulus	Various	SVM	55.72/58.16% Valence and 60.23/61.35% Arousal (2-classes) for EEG alone/Video stimulus as privileged information with EEG.
DREAMER Dataset				
Stamos et al. [8]	EEG, ECG	PSD, HRV	SVM	62.49/61.84% Valence and 62.17/62.32% Arousal (2-classes) using EEG alone/EEG and ECG fusion.

*Denotes mean F1-score. Accuracy value not available.

Table 5.2 shows that in almost all the cases EEG has been the preferred bio-sensing modality while vision modality i.e. the use of the frontal videos to analyze facial expressions has not been commonly used on these datasets. The classification accuracy for all emotion classes as per the circumplex model rather than only for arousal/valence is rarely reported. In other cases such as [115, 119], where the analysis of emotions is reported, the goal seems to be clustering the complete dataset into four classes rather than having a distinct training and testing partition for evaluation.

In terms of accuracy, it can be observed from Table 5.2 that using multiple sensor modalities, the best performance on the DEAP dataset is by [16] when utilizing data from multiple modalities. For the MAHNOB-HCI dataset, the best accuracy for valence and arousal is 73% and 68.5% respectively [15], which is again using multiple sensor modalities. The AMIGOS and DREAMER datasets were released recently and hence only baseline evaluation on these have been reported in Table 5.2.

This study will utilize complete datasets and not a subset of them, as in some previous studies. The developed methods were evaluated with disjoint partitions between training, validation, and test subsets of the complete datasets. This chapter reports the evaluation for all modalities separately (including using frontal videos that were ignored by other studies) and then combining them together.

5.2 Research Methods

This section details the various types of methods that were employed to extract features from each bio-sensing modality and frontal videos.

5.2.1 EEG feature extraction

For the DEAP and AMIGOS datasets, preprocessed EEG data are available, bandpass-filtered between 4-45 Hz, and corrected for eye-blink artifacts. For the MAHNOB-HCI and DREAMER datasets, bandpass filtering and artifact removal were performed using the Artifact Subspace Reconstruction (ASR) toolbox [87] in EEGLAB [120]. The processed EEG data were then converted into the frequency domain to extract both traditional and deep learning-based features (see below).

EEG-PSD features

For each EEG channel, the traditional power spectral density (PSD) in three EEG bands namely, theta (4-7 Hz), alpha (7-13 Hz), and beta (13-30 Hz) were extracted. These EEG bands were chosen since they account most towards human cognition. Half second overlapping windows were chosen for this procedure. The PSD was then averaged over the total trial length. Hence, because of the differences in the number of EEG channels, 96 features were obtained for trials in the DEAP and MAHNOB-HCI datasets and 42 features for trials in the AMIGOS and DREAMER datasets.

Conditional entropy features

To get information regarding the interplay between different brain regions, conditional entropy-based features were extracted. The conditional entropy between two random variables carries information about the uncertainty in one variable given the other. Hence, it acts as a measure of the amount of mutual information between the two random variables. The mutual information $I(X;Y)$ of discrete random variables X and Y is defined as

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) log \Big(\frac{p(x,y)}{p(x)p(y)} \Big) \tag{5.1}$$

The conditional entropy will be zero if the signal Y is completely determined by signal X. To calculate the conditional entropy, the mutual information $I(X;Y)$ between the two signals was first calculated, which requires the calculation of the approximate density function $\hat{p}(x)$ of the following form

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x^{(i)}, h) \tag{5.2}$$

where $\delta(.)$ is the Parzen window, h is the window width, N is the samples of variable x and $x^{(i)}$ is the ith sample. This approximate density function is calculated as an intermediary step to the calculation of true density $p(x)$, since when N goes to infinity it can converge to the true density if $\delta(.)$ and h are properly chosen [59]. $\delta(.)$ was chosen to be the Gaussian window.

$$\delta(z,h) = exp\Big(-\frac{z^T\Sigma^{-1}z}{2h^2}\Big)\Big/\Big\{(2\pi)^{d/2}h^d|\Sigma|^{1/2}\Big\} \tag{5.3}$$

where $z = x - x^{(i)}$, Σ is the covariance of z, and d is the dimension of x. By plugging the value of $d = 1,2$ the marginal density $p(x)$ and the density of the bivariate variable $(x,y), p(x,y)$ [59] were obtained. In this manner the mutual information $I(X;Y)$ was calculated which is related to conditional entropy $H(Y|X)$ by

$$I(X;Y) = H(Y) - H(Y|X) \tag{5.4}$$

The conditional entropy between all possible pairs of EEG channels was calculated over the complete trial length [25]. Hence, due to the differences in the number of EEG channels, 496 conditional entropy features were calculated for the DEAP and MAHNOB-HCI datasets and 91 features for the AMIGOS and DREAMER datasets.

EEG-PSD images-based Deep Learning features

This section proposes a novel method for feature extraction from the EEG data, which is based on deep convolution networks without requiring a large amount of training data. The method can also work in a similar manner for different types of EEG datasets i.e. datasets with different numbers and placements of electrodes, sampling rates, etc. The computed EEG-PSD features from the first method mentioned above were initially used to plot power spectrum heat maps for the three EEG bands using bicubic interpolation to calculate the values in the 2-D plane. These images now contain the topographical information for the three frequency bands according to the standard EEG 10-20 system (Fig. 5.1). It is worth noting that the commonly used EEG-PSD features in themselves do not take into account the EEG-topography i.e. the locations of EEG electrodes for a particular EEG band. Hence, this method tries to exploit EEG-topography to extract information regarding the interplay between different brain regions. It is for exploiting this information that EEG data was converted to an image-based representation and utilize pre-trained deep learning networks to extract such a relationship between various brain regions.

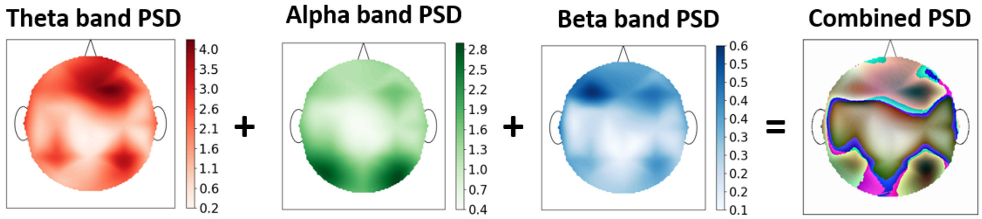


Figure 5.1: PSD heat-maps of theta (red), alpha (green), and beta (blue) EEG bands being added according to respective color-bar range to get combined RGB heat-map (Image border, nose, ears, and color-bars have been added for visualization only.)

As shown in Fig. 5.1, ‘Red’ color-map was used for the theta band, ‘Green’ color-map for the alpha band, and ‘Blue’ color-map for the beta band. The three colored images were then combined into an RGB colored image [121]. Based on the ranges and maximum values in the data for the three EEG bands, the ratio of alpha blending was used [122] to give weights to the three individual bands’ images before adding them together. This color image carries information about how the power in the three bands interacts with each other across the different brain regions. For example, a yellow-colored region has a higher amount of power in the theta (red) and alpha (blue) bands, whereas a pink-colored region has high power in the theta (red) and beta (blue) bands. In this manner, a single brain heat-map image can be used to represent spatial and spectral information in the three EEG bands. That is, one image was obtained representing the topographic PSD information for every trial. Because the images in the four datasets can be formed in a similar manner irrespective of the different numbers of EEG channels and positions, this method can be used across the four datasets easily.

A pre-trained VGG-16 network [97] was subsequently used to extract features from the combined RGB heat-map image. This network consists has been trained with more than a million images for 1,000 object categories using the Imagenet Database [123]. It has been shown that a pre-trained VGG-16 network can be utilized for feature extraction and classification for applications different than it was trained for [124, 125]. The RGB image is resized to $224\times224\times3$ size before submitting to the network. The last but one layer of this VGG network consists of 4,096 most significant features, which were extracted for emotion classification. Principal component analysis (PCA) was then applied to this feature space to reduce its dimension to 30 for each trial [60]. The features from this method and from the conditional entropy mentioned above were then combined for evaluation.

The resultant EEG-PSD images from the above method can be used to denote how the EEG spectral activity is distributed across various brain regions across time for a particular kind of stimulus. This can be done by sending such successive images (varying across time) to a reverse deep learning network and detect the most salient features i.e. activated regions of the brain across time for a particular stimulus. In the evaluation section below, these brain images were utilized to denote the brain regions that are most activated for different affective responses.

5.2.2 ECG/PPG-based feature extraction

Both ECG and PPG signals can be used to measure heart rate (HR) and heart-rate variability (HRV), though ECG can provide more useful information due to its greater ability to capture ECG t-wave etc. For consistency between the two types of signal measurements i.e. PPG and ECG, two methods were employed in the same manner on data from both of these modalities in the four datasets.

HRV features

HRV has shown to be a good metric for classifying emotional valence and arousal [126]. For every trial (whether PPG in the DEAP dataset or ECG in the AMIGOS, MAHNOB-HCI, and DREAMER datasets), a moving-average filter was first used with a window length of 0.25 seconds to filter out the noise in the data. Then, a peak-detection algorithm [127] was used after scaling the data between 0 and 1. The minimum distance between successive peaks as being at least 0.5 seconds apart was taken as the threshold to remove false positives as in Fig. 5.2.

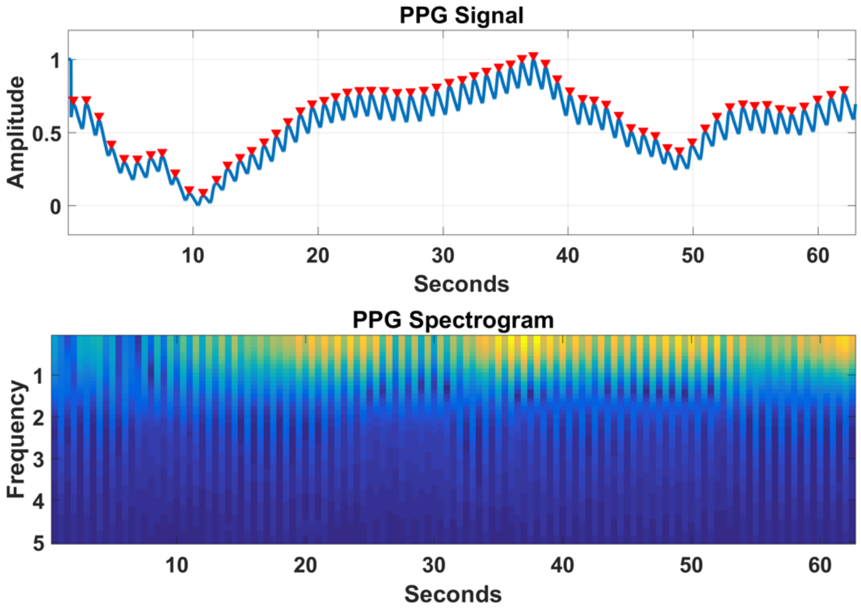


Figure 5.2: For a trial from DEAP dataset, PPG signal with peaks (in red) being detected for the calculation of RRs and HRV (above), and PPG spectrogram (below).

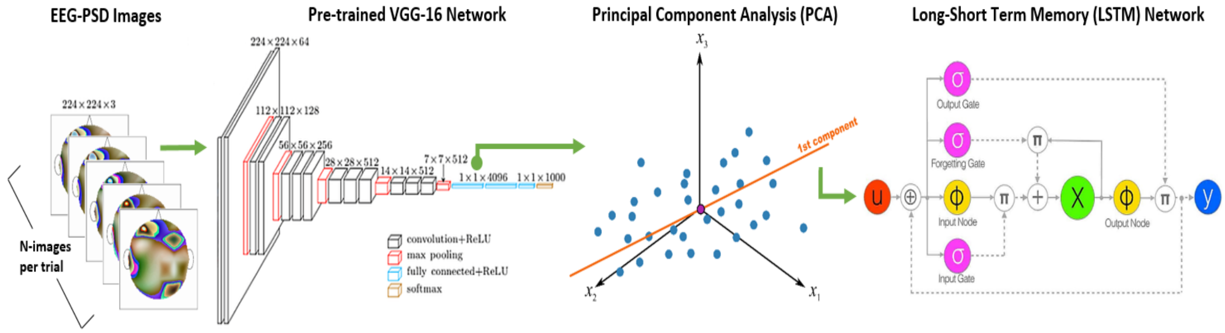


Figure 5.3: Network Architecture for EEG-PSD Trend-based Deep Learning Method.

The total number of peaks per minute represents the subject's heart rate. To calculate the HRV, the time differences between successive peaks were calculated to get inter-beat intervals (RRs). These RRs were then used to compute HRV using the pNN50 method [58]. This method of HRV calculation measures the percentage of successive RR intervals that differ by more than 50ms. This method has been shown to be correlated with the activity of the parasympathetic nervous system (PNS) [128]. For the datasets containing multiple ECG channels, the same procedure was performed for all the channels.

Extracting deep learning-based spectrogram features of ECG/PPG

Previous studies have reported that frequency-domain features in the ECG work well for tasks such as ECG beat discrimination [129]. To exploit time-frequency information from ECG/PPG, deep learning-based features were extracted on ECG/PPG by converting the time-series data to a frequency domain-based image representation. The frequency range of ECG/PPG signals is low and hence only 0-5 Hz range was taken into consideration. A spectrogram [130] was generated over the complete trial in this frequency range as in Fig. 5.2. To get a good amount of variations, the Parula color map was chosen for the spectrogram image. The frequency bins with various colors at different frequencies represent the signal across the trial length. The same procedure was employed to get the spectrogram images of ECG/PPG signals from the four datasets. The spectrogram images were resized to feed them into the VGG-16 network, and after which the resultant 4,096 extracted features were reduced to 30 features using PCA. The features from this method were concatenated with the HRV features from above for evaluation.

5.2.3 GSR-based feature extraction

Similar to ECG/PPG, two methods were employed to extract features from the GSR data, one in the time-domain and the other in the frequency domain.

Statistical features

A moving average filter was first used with a window length of 0.25 seconds to remove noise from the GSR signal. For each trial, eight statistical features were calculated from the time-domain GSR data. Two features are the number of peaks and the mean of absolute heights of the peaks in the signal. Six more statistical features based on n^{th} order moments of the GSR time-series data were calculated as shown in [131]. These features measure the trend i.e. variations in the GSR data in actual, and successive first and second differences of the signal.

Extracting deep learning-based spectrogram features of GSR

GSR signals change very slowly and hence this study focused only on the 0-2 Hz frequency range. Similar to ECG, the spectrogram image of GSR was generated for each trial in the above frequency range. VGG-16 network-based features were then extracted to characterize the most meaningful interactions between various edges in the spectrogram. These features were reduced to 30 using PCA and then concatenated with the time-domain GSR features from above.

5.2.4 Frontal video-based feature extraction

Unlike other studies in Table 5.2, this study also utilized the frontal videos of the subjects for emotion/valence/arousal/liking classification. Facial expressions can be very reliable indicators of one's emotions based on his/her personality i.e. willingness to show emotions by various facial expressions. For each frontal video trial, first, a single frame was extracted for every second in the trial by extracting the first frame for every second of the video. For this step, the extreme ends of the image were excluded and a threshold was placed on the minimum face size to be 50×50 pixels. This was done to reduce computational complexity and increase the face detector's accuracy. Face detection was performed using Haar-like features based on Viola-Jones object detector [92]. A small portion of images had a majority of the face occluded due to the subject putting his/her hand over their face. The face detector failed in these instances and hence these were discarded.

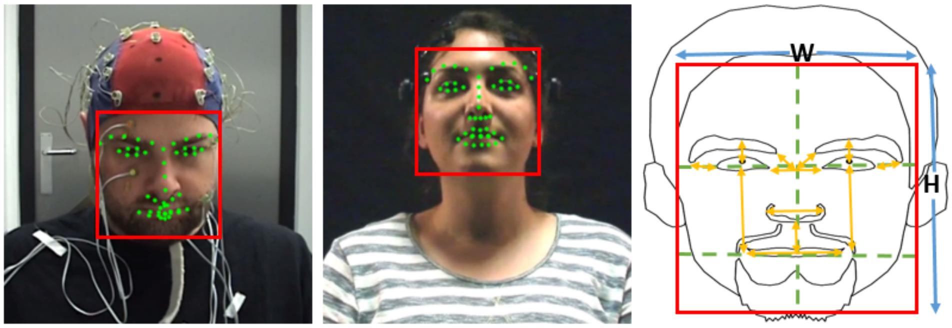


Figure 5.4: Detected face (marked in red) and face localized points (marked in green) in DEAP Dataset (left), AMIGOS Dataset (center), and subset of features (marked in yellow) computed using face localized points (right). The features are normalized using height (H) and width (W) of the detected face. These subjects' consent to use their face is marked in respective datasets.

Facial points localization based features

The state-of-the-art Chehra algorithm [93] was applied to the extracted facial regions to obtain 49 localized points on the face representing the significant parts as shown in Fig. 5.4. This algorithm does not need any human input or a dataset-specific training model for predicting the localized face points, making the process fully automated. Previous research studies have reported promising results using the face action units (AUs) based on such facial landmarks [68]. These 49 localized points were then used to calculate 30 features based on distances such as that between the center of the eyebrow to the center of the eye, between the nose and the middle part of the upper lip, between upper and lower lips, etc. Many of the 30 features are described in [68] while others by designed by hand. All such features were normalized based on the height and width of the detected face to remove variations due to the distance from the camera. The mean, 95^{th} percentile (more robust than maximum), and standard deviation of these 30 features across the frames in a single trial were then calculated. These 90 parameters were then used for evaluation.

Deep Learning-based features

The use of deep learning has transformed computer vision in multiple ways. This is because such deep networks are capable of extracting feature representations from images that capture both uniform (contrast etc.) and complex (small changes in texture etc.) types of features. Hence, these networks were utilized on face-images using a deep network pre-trained on VGG-faces dataset [132]. The extracted face region was resized to $224\times224\times3$ for each selected frame in the trial. Similar to the CNN-based deep learning method used above for the bio-sensing modalities, 4,096 most meaningful features were extracted on these resized images. But, unlike the bio-sensing method, a different VGG network that has been specifically trained on more than 2.6 million face images from more than 2,600 people [132] was employed. This was done to extract features that are more relevant to the face-dependent feature space. The mean, 95^{th} percentile, and standard deviation of the features across the images in every trial were computed and the subsequent features space was reduced to 30 using PCA.

5.2.5 Dynamics of the EEG/Face features using deep learning

The above-mentioned methods for extracting deep learning features from EEG/face-videos are special cases in which a single trial is represented by a single image (EEG-PSD image/Single feature space for face images in a video). But, these methods do not fully take into account the temporal dynamics of the features over time within the trial. Hence, this study proposes a new method in which such images (EEG-PSD or face region) are utilized for every second within a trial. Fig. 5.3 shows the network architecture for this method for the EEG-PSD images. Multiple EEG-PSD images were formed for each trial by generating one image for each second, all of which went through the pre-trained VGG network. The 4,096 features from the off-the-shelf deep learning network were then obtained for each image. In addition, the conditional entropy features for every second were also calculated. PCA was then used to reduce the dimensionality of the feature space comprising features from EEG and face-videos. The resultant feature space has 60 most representative features. These $60 \times N$ (N = trial length in seconds) features are then sent to a Long-Short Term Memory (LSTM) network [133]. However, this method could only be employed on the DEAP dataset since the AMIGOS and MAHNOB-HCI datasets have varying trial lengths and the DREAMER dataset does not contain any video data. The huge variations in the trial length in the AMIGOS and MAHNOB-HCI datasets meant that during the data preparation phase of LSTM, a large amount of padding was needed. This may be possible in data from physical sensors (like temperature, luminous, pressure, etc.) where interpolation is easy to perform. But, for bio-signals, this is not desirable because affective labels were not reported by the subject during the course of each video trial i.e. there are no labels to suggest which parts of the video contributed most towards the affective response. Hence, LSTM networks were not deployed on such datasets.

5.3 Evaluation

This section presents the evaluation of the various feature-extraction methods described above. First, the performance of the classification of affective states using the deep learning features from the pre-trained convolution network was compared with that using traditional EEG features. The evaluation also reports the classification performance when features from these modalities are fused together. Thereafter, the classification performance was evaluated using each modality individually on the four datasets, on combining the datasets together, and for transfer learning. Finally, results are presented for a novel deep learning-based technique to identify the most important brain regions associated with emotional responses.

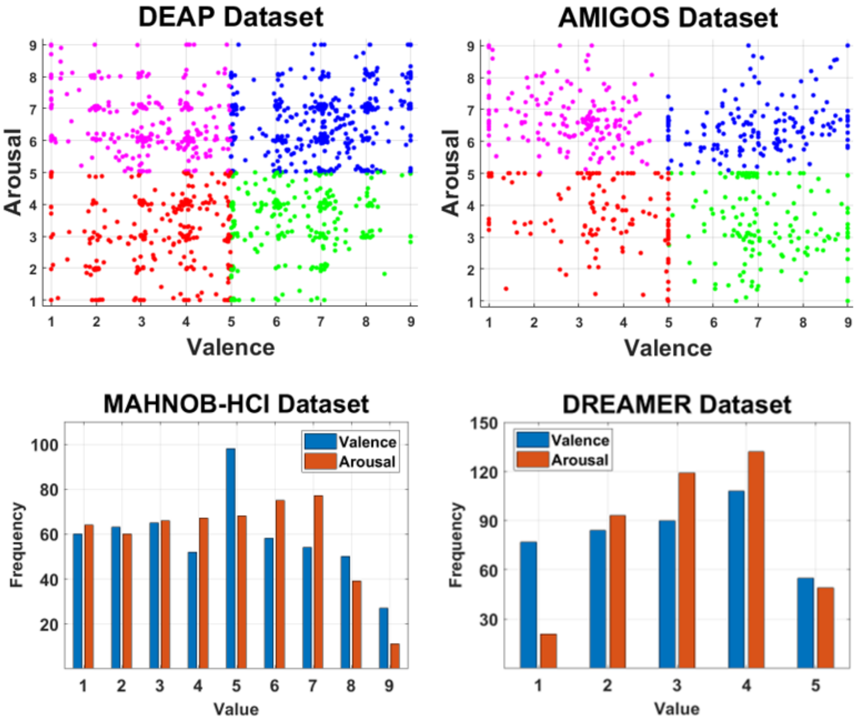


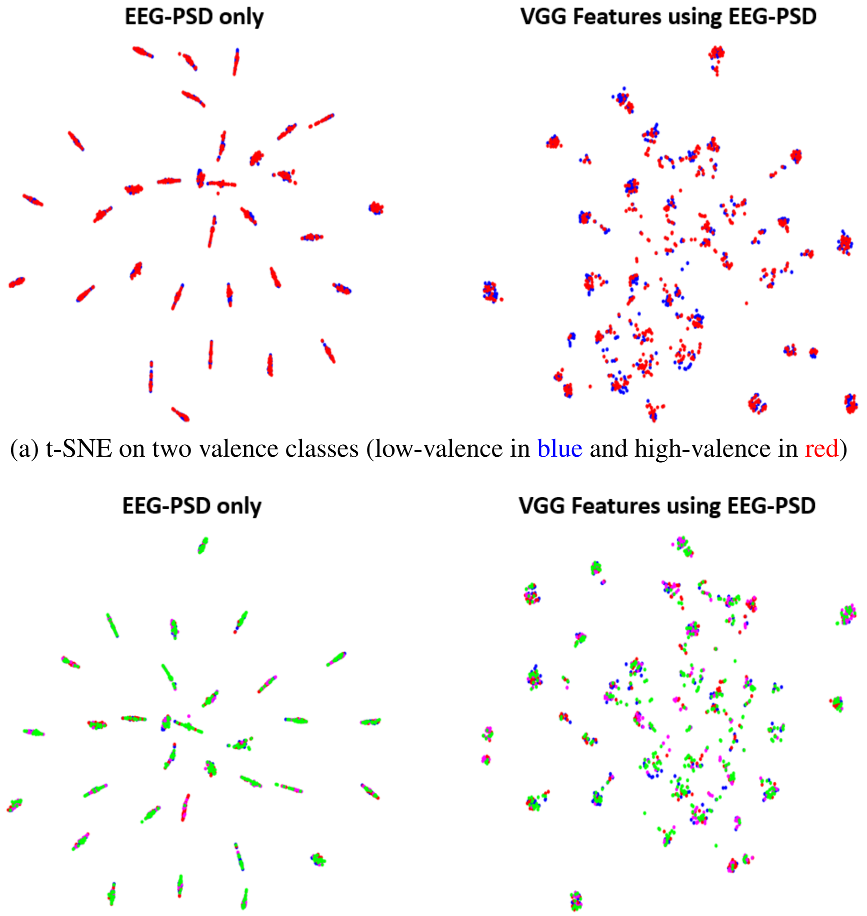
Figure 5.5: Distribution of emotion classes in the four datasets.

Fig. 5.5 shows the distribution of self-reported valence and arousal for the four datasets. It is evident that the DEAP dataset has a higher concentration of trials closer to neutral emotion i.e. near the center of the graph. For each individual dataset separately, leave-one-subject-out evaluation was performed and results for single modality classification are shown in Table 5.3 and for multi-modality classification in Table 5.4. Then, an evaluation was performed by combining datasets together and associated results are shown in Table 5.5. Finally, transfer learning was employed among the datasets i.e. training on one dataset and testing on another (Table 5.6). For these two latter evaluations of combining the datasets and using transfer learning, the datasets were randomly divided into two parts with an 80/20 ratio while 10-fold cross-validation was also performed. The classification was done using extreme learning machines (ELM) [62] with variable numbers of neurons, which has been shown to perform better than support vector machines (SVM) for various cases [134]. All the features were re-scaled between -1 and 1 before training the ELM. A single-layer ELM was used with a sigmoid activation function. For the trend-based deep learning method, two hidden layers in the LSTM were used with the number of neurons being 200 and 100, respectively. Stochastic gradient descent with a momentum (SGDM) optimizer was used to train the LSTM network.

5.3.1 Visualizing class-separability using the traditional vs. deep learning features

One of the hypotheses of this study is that the traditional methods for analyzing EEG can be improved by using deep learning-based features obtained from pre-trained convolution networks. This is important because training convolution networks requires huge datasets, which are usually unavailable in the bio-sensing domain. Hence, the Deep Learning method described in Section 5.2.1 should be able to extract more meaningful features from EEG-PSD features. t-SNE [135] was used to visualize the dimensionally reduced space using traditional EEG-PSD features for 2-class valence and 4-class emotions on the DEAP dataset with fixed trial length. Kullback-Leibler (KL) divergence was used for measuring similarity and Euclidean distance was used as the distance measure for the t-SNE implementation. The same approach was then applied to the features obtained by the VGG network, which were computed after using the EEG-PSD features to create a combined RGB image in Fig. 5.1.

Fig. 5.6 shows that trials in both 2-valence and 4-emotion classes can be separated to a better degree (although not optimal) when using the VGG features from the EEG-PSD combined image than directly using the EEG-PSD features. The EEG-PSD features only form distinct clusters for each subject and are unable to separate the valence/emotion classes whereas the VGG features allow for better separation.



(a) t-SNE on two valence classes (low-valence in blue and high-valence in red)

(b) t-SNE on four emotion classes (HVHA in blue, LVHA in red, LVLA in magenta, and HVLA in green)

Figure 5.6: Visualization of feature spaces using t-SNE [135] in trials from the DEAP dataset on the EEG-PSD features and the VGG features derived from the combined RGB image. The VGG features allow for better separation.

5.3.2 Evaluating individual modality performance

This section presents the classification results obtained by using individual modalities on the four datasets. Table 5.3 shows accuracy and mean F1-score results for individual modalities.

It is clear from Table 5.3 that the results in multiple categories for all the four datasets are better than those reported previously, as shown in Table 5.2. The CNN based features that were extracted for all modalities contribute most towards this classification improvement for all the modalities. Furthermore, for all four datasets and for all modalities, the performance is substantially greater than the chance accuracy. EEG proves to be the best performing bio-

Table 5.3: Individual Modality Performance Evaluation

Response	EEG	Cardiac	GSR	Face-1	Face-2
DEAP Dataset					
Valence	71.09/0.68	70.86/0.69	70.70/0.68	71.08/0.68	72.28/0.70
Arousal	72.58/0.65	71.09/0.63	71.64/0.65	72.21/0.65	74.47/0.68
Liking	74.77/0.65	74.77/0.64	75.23/0.64	75.60/0.62	76.69/0.62
Emotion	48.83/0.26	45.55/0.31	45.94/0.25	43.52/0.28	46.27/0.27
AMIGOS Dataset					
Valence	83.02/0.80	81.89/0.80	80.63/0.79	80.58/0.77	77.28/0.74
Arousal	79.13/0.74	82.74/0.76	80.94/0.74	83.10/0.76	77.28/0.72
Liking	85.27/0.81	82.53/0.77	80.47/0.72	80.27/0.72	79.81/0.72
Emotion	55.71/0.30	58.08/0.36	56.41/0.34	57.74/0.28	56.79/0.27
MAHNOB-HCI Dataset					
Valence	80.77/0.76	78.76/0.73	78.98/0.73	83.04/0.79	85.13/0.82
Arousal	80.42/0.72	78.76/0.74	81.84/0.75	82.15/0.77	81.57/0.76
Emotion	57.86/0.33	57.23/0.35	57.84/0.32	60.41/0.35	63.42/0.35
DREAMER Dataset					
Valence	78.99/0.75	80.43/0.78	—	—	—
Arousal	79.23/0.77	80.68/0.77	—	—	—
Emotion	54.83/0.33	57.73/0.36	—	—	—

Cardiac features refer to features extracted using PPG in the DEAP dataset and using ECG in the AMIGOS, MAHNOB-HCI, and DREAMER datasets. Face-1 and Face-2 refer to the methods in Sections 5.2.4 and 5.2.5 respectively. Valence, Arousal, and Liking have been classified into two classes (50% chance accuracy) whereas Emotion has been classified into four classes (25% chance accuracy). All values denote the mean percentage accuracy followed by the mean F1-score (separated by "/") whereas missing values represent missing modality data.

sensing modality whereas Cardiac and GSR features also perform very well despite containing fewer channels. Furthermore, the frontal-video-based showed high accuracy in the affective classification for the three datasets and surpassed the accuracy obtained by the bio-sensing modalities in many cases.

The classification performance using various modalities even for varying trial length is consistently better than that reported in previous studies (Table 5.2). The results of this study surpass the previous best results obtained by using only individual modalities for the DEAP and MAHNOB-HCI datasets and the baseline accuracies for the AMIGOS and DREAMER datasets. Furthermore, higher accuracy was observed for Liking classification than for Valence/Arousal for

DEAP and AMIGOS datasets, suggesting that it might be easier for subjects to rate their likeness for the video contents than rating valence and arousal. This is understandable since the latter terms are difficult to comprehend than Liking and depend highly on the physiological baseline of the subject at any particular time.

5.3.3 Evaluating multi-modality performance

This section presents the results of combining different modalities for affective state classification. Specifically, as shown in Table 5.4, the three bio-sensing modalities (only two for the DREAMER since it does not contain GSR data) are first combined to evaluate their joint performance and then the EEG and Face-video modalities through the CNN-VGG-extracted features. Finally, for the DEAP dataset, the results are presented of training an LSTM network with the time-varying features from the EEG and Face-video modalities (see Section 5.2.5).

In almost all cases, it was found that combining features from multiple modalities increases classification accuracy. The fusion of features from bio-sensing modalities increases the accuracy in many cases for all the four datasets. It is also notable that by training the LSTM network with the features from EEG and Face-video modalities not only increases the accuracy as compared to the individual modalities (from Table 5.3) but also outperform the best accuracy on the DEAP dataset reported in Table 5.2. For AMIGOS, MAHNOB-HCI, and DREAMER datasets, it was observed that using multiple modalities outperform single-modality accuracy in many cases and sets up new benchmarks by beating previous best results. Two-sample t-Test was also performed between Bio-Sensing and EEG plus Face multi-modal combinations for the datasets. The p-values of the t-Test analysis for the valence, arousal, liking, and emotion classification for the DEAP dataset were 0.676, 0.543, 0.939, and 0.347 respectively. The p-values for the valence, arousal, liking, and emotion classification for the AMIGOS dataset were 0.003, 0.266, 0.134, and 0.026 respectively. The p-values for the valence, arousal, and emotion classification for the MAHNOB-HCI Dataset were 0.0134, 0.293, and 0.149. Similar t-Test could not be performed on

Table 5.4: Multi-modality Performance Evaluation

Response	**Bio-sensing**	**EEG and Face**	**EEG and Face (LSTM)**	**Previous Best Accuracy**
DEAP Dataset				
Valence	71.87/0.68	73.94/0.69	79.52/0.70	77.19
Arousal	73.05/0.68	74.13/0.66	78.34/0.69	76.17
Liking	75.86/0.69	76.74/0.63	80.95/0.70	68.40
Emotion	49.53/0.27	48.11/0.28	54.22/0.31	50.80
AMIGOS Dataset				
Valence	83.94/0.82	78.23/0.74	—	—
Arousal	82.76/0.76	81.47/0.72	—	—
Liking	83.53/0.77	81.49/0.75	—	—
Emotion	58.56/0.40	58.02/0.29	—	—
MAHNOB-HCI Dataset				
Valence	80.36/0.75	85.49/0.82	—	73.00
Arousal	80.61/0.71	82.93/0.77	—	68.50
Emotion	58.07/0.30	62.07/0.35	—	—
DREAMER Dataset				
Valence	79.95/0.77	—	—	62.49
Arousal	79.95/0.77	—	—	62.32
Emotion	55.56/0.33	—	—	—

Bio-sensing refers to combining features from EEG, ECG/PPG, and GSR signals.
EEG + Face refers to combining features from EEG- and video-based modalities.
EEG + Face (LSTM) refers to combining features from EEG- and video-based modalities for every second in the trial to train an LSTM model. Due to the trial length varying widely in the AMIGOS and MAHNOB-HCI datasets, the LSTM-based method could not be applied to them. The DREAMER dataset does not have video data.

the DREAMER dataset since it does not contain Video (Face) modality data.

5.3.4 Evaluating the classification performance using combining datasets and transfer learning

To show that the proposed deep learning-based features are independent of the number of EEG channels, trial length, the image resolution of the video, ECG/PPG cardiac modality, etc., an ELM classifier was trained with data from more than one datasets. A transfer-learning

Table 5.5: Combined Dataset Performance Evaluation

Response	EEG	Cardiac	GSR	Face-1	Face-2
DEAP + AMIGOS Combined Dataset					
Valence	62.80/0.58	59.69/0.59	59.64/0.58	63.04/0.62	62.38/0.62
Arousal	62.27/0.61	63.61/0.61	61.98/0.62	67.66/0.65	68.65/0.66
Liking	69.13/0.59	69.27/0.61	69.27/0.55	67.99/0.64	68.65/0.64
Emotion	37.47/0.27	37.50/0.22	37.24/0.31	40.92/0.36	42.24/0.36
DEAP + AMIGOS + MAHNOB-HCI Combined Dataset					
Valence	61.24/0.60	58.57/0.59	58.98/0.57	61.59/0.61	62.56/0.63
Arousal	65.15/0.63	61.84/0.61	61.02/0.59	65.94/0.65	67.15/0.66
Emotion	40.21/0.35	36.33/0.31	35.71/0.28	42.51/0.33	43.00/0.32

The DEAP + AMIGOS combined dataset consists of the data from 72 subjects and more than 1,900 trials. The DEAP + AMIGOS + MAHNOB-HCI combined dataset consists of the data from 99 subjects and more than 2,400 trials. Only the deep learning-based methods are used for extracting features for evaluation from various modalities because these can be extracted from all datasets in the same manner.

Table 5.6: Transfer Learning Performance Evaluation

Response	EEG	Cardiac	GSR	Face-1	Face-2
DEAP + AMIGOS (Train Dataset), MAHNOB-HCI (Test Dataset)					
Valence	63.55/0.60	64.77/0.54	64.96/0.55	55.02/0.52	62.01/0.62
Arousal	58.37/0.55	62.50/0.52	62.50/0.52	59.32/0.54	58.60/0.58
Emotion	36.65/0.32	39.58/0.28	38.64/0.28	36.38/0.39	34.05/0.37
DEAP (Train Dataset), MAHNOB-HCI (Test Dataset)					
Valence	62.70/0.54	63.59/0.46	65.19/0.47	56.48/0.49	59.86/0.59
Arousal	61.99/0.55	61.46/0.48	63.23/0.52	59.33/0.56	61.99/0.60
Emotion	35.88/0.23	38.01/0.24	39.08/0.24	33.57/0.33	32.50/0.22

Only the deep learning-based methods are used for extracting features for evaluation from various modalities because these can be extracted from all datasets in the same manner.

approach was also employed to train the ELM classifier with data from some of the four datasets and then test it against the remaining dataset. The combined datasets were randomly divided into an 80:20 ratio for training and testing. This allowed the verification of how scalable are the proposed feature extraction methods across datasets having discrepancies in recording devices (e.g. ECG vs PPG) and parameters (e.g. channel numbers). Table 5.5 shows that despite all these discrepancies across the datasets, the proposed framework works well and always performs considerably better than the chance accuracy and the baseline accuracies for individual datasets [5, 6, 7] reported in Table 5.2. Table 5.6 shows the results of training with two datasets and testing on the third. The above combinations of datasets were chosen because all the sensor modalities were used in the datasets and the DEAP dataset contains more trials (1,280 trials) than the other two datasets combined together (AMIGOS and MAHNOB-HCI containing 640 and 540 trials respectively). Even when ELM was tested on a dataset, the trials from which were not used for training, the results were consistently better (more so for ECG/PPG and GSR modalities) than many previous studies and far above the chance accuracy. The slight decrease in performance for some modalities compared to those trained with the data from the same dataset might be due to two factors, namely, the varying trial length between the datasets and only using the VGG-based features common to the datasets (for consistency among the datasets) as opposed to combining features from other methods like conditional entropy, HRV, face-localization, etc.

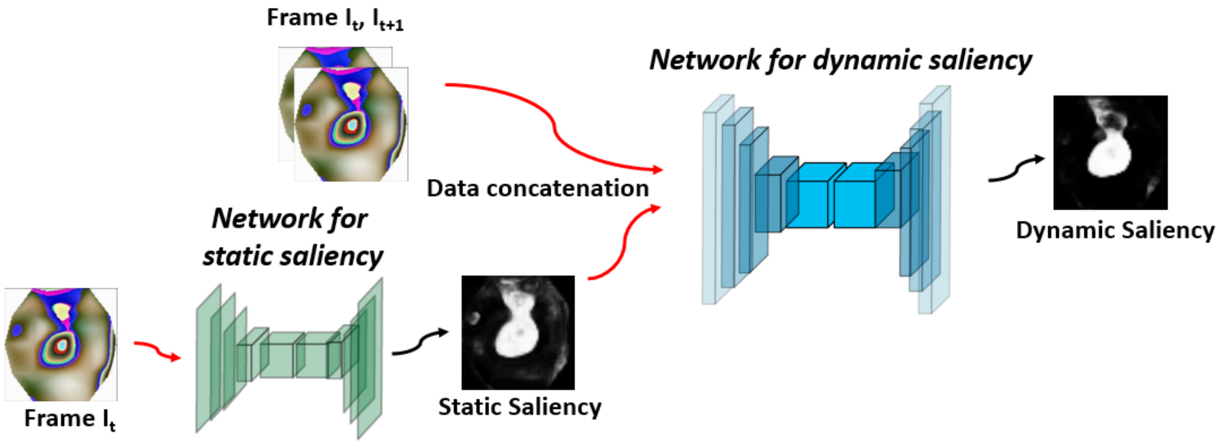


Figure 5.7: Salient brain regions corresponding to low/high valence/arousal in DEAP dataset. The frontal lobe has high activation.

5.3.5 Identifying the salient brain regions that contribute towards processing various emotions

As is clear from the performance evaluation sections above, the deep learning-based methods are able to extract more meaningful features and perform better than traditional features. This section aims to explore what insights the proposed deep learning-based method can provide on the brain regions contributing to emotional responses. To this end, a reverse VGG network (before the final max pooling step) was added to the pre-trained VGG network that extracted the informative features used above. That is, a deconvolving network was added to the convolving network. As shown in [136], the convolution-deconvolution network can be used to identify the most salient areas in the images in both static and dynamic manner. This network (Fig. 5.7) was utilized to detect those regions in the EEG-PSD brain images that contribute most towards processing various emotions. The pairs of EEG-PSD images for consecutive seconds for a trial I_t, I_{t+1} were sent to the dynamic convolution-deconvolution network along with the output of the static saliency for the image at I_t. The static saliency network identified the most salient areas whereas the dynamic saliency network was able to learn the variations between these image areas for every consecutive second. This procedure was done for every second for all the trials. The results are reported only from the DEAP dataset for this method because of its fixed trial length.

The RGB combined images (Figure 5.1) were used for every second for every trial of the low/high valence/arousal instances from the DEAP dataset by first convolving and then deconvolving them in the network described above. Hence, theoretically, the areas with most salient variations across the trials would represent the brain regions that are most receptive to the particular affective state. Fig. 5.8 shows the brain activity for these affective states after averaging the output of the network across all the trials for the affective state (valence/arousal). Most of the activity is over the frontal and central lobes around the FC3, FCz, FC4, and Cz locations according to the EEG 10-20 system. This is consistent with the textbook evidence regarding

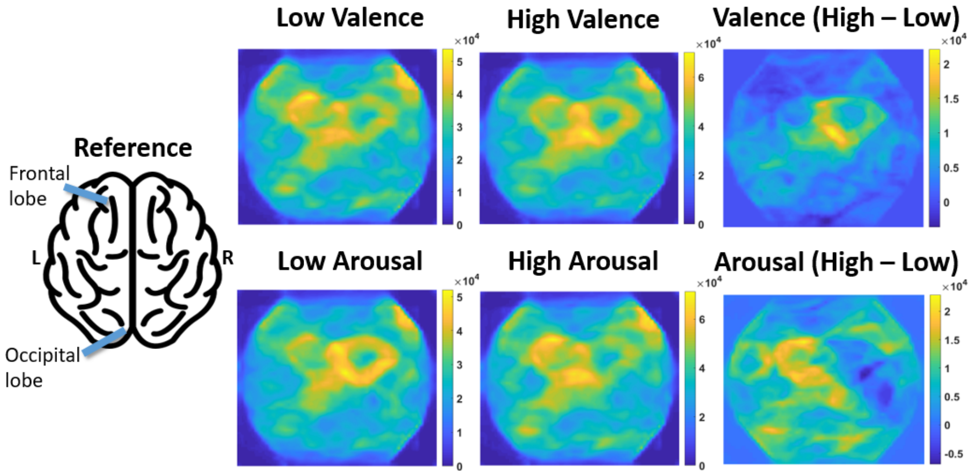


Figure 5.8: Convolution-Deconvolution network on EEG-PSD images to identify salient brain regions corresponding to affective states. Pixels in individual images were scaled from 0 to 1.

the processing of human emotions [137, 138]. More interestingly, it was observed from the difference image between high and low arousal that the processing of arousal affective state is much more widely distributed across the brain than valence. Hence, this method allowed for the usage of a single image to represent such areas across the brain, and across all subjects and trials, that are most activated for a particular affective measure rather than using multiple such EEG images. These results are presented as a starting point to take this work towards using the EEG for investigating the generation and processing of emotions inside the brain.

5.4 Chapter Concluding Remarks

Advances in deep learning have not translated into bio-sensing and multi-modal affective computing research domains mostly due to the absence of very-large-scale datasets. Such datasets are available for vision/audio modalities due to the ease of data collection. Hence, for the time being, it seems that the only viable solution is to use “off-the-shelf” pre-trained deep learning networks to extract features from bio-sensing modalities. The proposed methods present the advantages of being scalable and able to extract features from different datasets. Such “off-the-shelf" features prove to work better than the traditionally used features of various bio-sensing modalities.

This study proposed novel methods to affective computing research by employing deep learning features across various modalities. It showed that these methods perform better than previously reported results on four different datasets containing various recording discrepancies. The methods were also evaluated on the combined datasets. Furthermore, the various modalities were fused to augment the performance of our models. The LSTM was used to learn the temporal dynamics of the features during stimulus presentation and increase the classification accuracy, compared to averaging the features across that trial. This study also showed that features extracted from bio-sensing modalities such as EEG can be combined with those from the video-based modality to increase the accuracy further.

Since affective computing encompasses the study of human affects, awareness, and attention is broadly a component of the same area of research. The next chapter discusses how such multi-modal bio-sensing tools may actually be used to save precious human lives by continuously monitoring driver's awareness and identifying his/her physiological response to hazardous on-road situations.

This chapter is in part a reprint of material that has been accepted for publication in the journal IEEE Transactions on Affective Computing (2019), by Siddharth Siddharth, Tzyy-Ping Jung, and Terrence J. Sejnowski. The dissertation author was the primary author of this paper.

Chapter 6

Utilizing Multi-modal Bio-sensing Tools Toward Driver Awareness Analysis

Affective Computing being the study of human affects, one of the key themes in the real-world is to study human awareness utilizing bio-sensors. This chapter presents the tools developed in the previous chapters for driver awareness analysis in situations approximating real-world driving scenarios. This study focuses on the specific problem of inferring driver awareness in the context of attention analysis and hazardous incident activity. Even after the development of wearable and compact multi-modal bio-sensing systems in recent years, their application in driver awareness context has been scarcely explored. The capability of simultaneously recording different kinds of bio-sensing data in addition to traditionally employed computer vision systems provides exciting opportunities to explore the limitations of these sensor modalities. This study explores the applications of three different bio-sensing modalities namely electroencephalogram (EEG), photoplethysmogram (PPG), and galvanic skin response (GSR) along with a camera-based vision system in driver awareness context. This study further assesses the information from these sensors independently and together using both signal processing- and deep learning-based tools. This chapter shows that the developed methods outperform previously reported studies to classify driver attention and detecting hazardous/non-hazardous situations for short time scales of two seconds. This study utilizes EEG and vision data for high-resolution temporal classification (two seconds) while additionally also employ PPG and GSR over longer time periods. The methods are evaluated by collecting user data on twelve subjects for two real-world driving datasets among which one is publicly available (KITTI dataset) while the other was collected by the authors (LISA dataset) with the vehicle being driven in an autonomous mode. This work presents an exhaustive evaluation of multiple sensor modalities on two different datasets for attention monitoring and hazardous events classification.

6.1 Introduction and Related Research

With the development of increasingly intelligent vehicles, it has now become possible to assess the criticality of a situation much before the event actually happens. This makes it imperative to understand the criticality of a situation from the driver's perspective. For example, one important area where driver's attention monitoring is most crucial is predicting takeover readiness of the automobile [139]. While computer vision continues to be the preferred sensing modality for achieving the goal of assessing driver awareness, the use of bio-sensing systems in this context has received wide attention in recent times [140, 141, 142]. Most of these studies have used electroencephalogram (EEG) as the preferred bio-sensing modality.

The emergence of wearable multi-modal bio-sensing systems [90, 30] has opened a new possibility to overcome the limitations of individual sensing modalities through the fusion of features from multiple modalities. For example, the information related to driver's drowsiness extracted from EEG (which suffers from low spatial resolution especially when not using a very large number of sensors) may be augmented by the use of galvanic skin response (GSR) which does not suffer from electromagnetic noise (but has a low temporal resolution).

Driver awareness depends highly on the driver's physiology since different people react differently to fatigue and to their surroundings. This means that one-fit-for-all type of approach using computer vision based on eye blinks/closure etc. might not scale very well across drivers. It is here that the use of bio-sensing modalities (EEG, GSR, etc.) may play a useful role in assessing driver awareness by continuously monitoring the human physiology. The fusion of data from vision-based systems and bio-sensors might be able to generate more robust models for the same. Furthermore, EEG with its higher temporal resolution than other common bio-sensors may prove to be very useful for detecting hazardous vs. non-hazardous situations on short time scales (such as 1-2 seconds) if such situations do not register in the driver's facial expressions. Additionally, the driver's physiology may provide insights into how they react to various situations during the drive which may have a correlation with the driver's safety. For example, heart-rate variability, which has been shown to model human stress [143] may be used as an indicator of when it is unsafe for a person to drive a vehicle.

Deep Learning has many applications in computer vision-based driver-monitoring systems [144, 145]. But, these advances have not translated towards the data from bio-sensing modalities. This is primarily due to the difficulty in collecting very large scale bio-sensing data which is a prerequisite for training deep neural networks. Collecting bio-sensing data on a large scale is costly, laborious, and time-consuming. It requires sensor preparation and instrumentation on the subject before the data collection can be started whereas for collecting images/videos even a smartphone's camera may suffice without the need to undergo any sensor preparation in most cases.

6.1.1 Contributions

- This study focuses on driver awareness and his/her perception of hazardous/non-hazardous situations from bio-sensing as well as vision-based perspectives. The features are individually used from three bio-sensing modalities namely EEG, PPG, and GSR, and vision data to compare the performance of these modalities. The fusion of features is also used to understand if and in what circumstances can it be advantageous.

- To this end, a novel feature extraction and classification pipeline is presented that has the ability to work with real-time capability. The pipeline utilizes pre-trained deep neural networks even in the absence of very large scale bio-sensing data.

- The developed framework is completely scalable for signal acquisition, feature extraction, and classification that has been designed with the intent to work in real-world driving scenarios. The framework is modular since information is extracted separately from each sensor modality.

- Finally, two hypotheses are tested in this paper. First, it is evaluated if the modalities with low-temporal resolution (but easily wearable) namely PPG and GSR can work as well as EEG and vision modality for assessing driver's attention. Second, the study tests if (and when) the fusion of features from different sensor modalities boost the classification performance over using each modality independently for attention and hazardous/non-hazardous event classification.

6.1.2 Related Studies

Driver monitoring for assessing attention, awareness, behavior prediction, etc. has usually been done using vision as the preferred modality [146, 147, 148]. This is carried out by monitoring the subject's facial expressions and eye-gaze [149] which are used to train machine learning models. But, almost all such studies utilizing "real-world" driving scenarios have been conducted during daylight when ample ambient light is present. Even if infra-red cameras are used to conduct such experiments at night, vision modality suffers from occlusion and widely varying changes in illumination [146], both of which are not uncommon in such scenarios. Furthermore, it has been shown that EEG can classify hazardous vs. non-hazardous situations over short periods which is not possible with images/videos [10].

On the other hand, those driving studies focusing on bio-sensing modalities suffered from impracticality in the "real-world" situations. This is because the bio-sensors were usually bulky, required wet electrodes, and were very prone to noise in the environment. Hence, the studies carried out with such sensors required wet electrode application and monitors in front of participants with minimal motion [150, 151]. In the early years of this decade, such bio-sensing systems gave way to more compact ones capable of transmitting data wirelessly while being more resistant to the noise by better EM (electro-magnetic) shielding and advances in mechanical design. Finally, recent advances have led to the development of multi-modal bio-sensing systems and the ability to design algorithms utilizing the fusion of features from various modalities. This has been utilized for various applications such as in affective computing and virtual reality [63, 152].

The use of deep learning for various applications relating to driver safety and autonomous driving systems has skyrocketed in the past few years. These studies have ranged from understanding driving behavior [153] to autonomous driving systems on highways [154] to detecting obstacles for cars [155] among other applications. All such studies only used vision modality since as pointed out before due to the prevalence of large-scale image datasets. However, the use of "pre-trained" neural networks for various applications [124, 125] may provide a new opportunity. Hence, if bio-sensing data can be represented in the form of an image, it should be possible to use such networks to extract deep-learning-based optimal feature representation of the image (henceforth called most significant features) even in the absence of large-scale bio-sensing datasets.

The developed software pipeline uses multiple bio-sensing modalities in addition to the vision, which was not the case with previous state-of-the-art evaluations done on the KITTI dataset [156]. The authors' previous work on EEG and visual modality data with the KITTI dataset [157] showed the utility of using these two modalities for driver attention monitoring but did not present a holistic view of multiple bio-sensing modalities with short-time-interval analysis on other datasets. The use of another dataset collected by the authors while driving a car in autonomous mode with the driver strapped with bio-sensing modalities and holistic comparison of multiple sensing modalities is a chief feature of this research study. Also, the previous research studies generally [10, 158] used a single modality and traditional features (i.e. not based on deep neural networks) for classification. Through the presented evaluation, it is shown that the developed software pipeline easily outperforms previous best results with higher-order features.

6.2 Research Methods

This section discusses the various research methods that were employed to pre-process the data and extract features from each of the sensor modalities used in this study. It also shows the visualization of sensor data for each sensor modality. This is done by showing the sensor data in the time and/or frequency domains as no threshold exists that can distinguish between different mental or psychological states directly for each sensing modality even for a single subject.

6.2.1 EEG-based Feature Extraction

The cognitive processes pertaining to attention and mental load such as while driving are not associated with only one part of the brain. Hence, the goal was to map the interaction between various regions of the brain to extract relevant features related to attention. The EEG was initially recorded from a 14-channel Emotiv EEG headset at 128 Hz sampling rate [36]. The EEG channel locations as per the International 10-20 system were AF3, AF4, F3, F4, F7, F8, FC5, FC6, T7, T8, P7, P8, O1, and O2. Artifact subspace reconstruction (ASR) pipeline was used in the EEGLAB [120] toolbox to remove artifacts related to eye blinks, muscle movements, line noise, etc. [87] This pipeline is capable of working in real-time and unlike Independent Component Analysis (ICA) [48] has the added advantage of being able to remove noise without much loss of EEG data when a very large number of EEG sensors are not present. Then, the EEG data was band-pass filtered between 4-45 Hz. The band-pass filter was designed to capture the theta, alpha, beta, and low-gamma EEG bands i.e. frequency information between 4-45 Hz. On this processed EEG data, two distinct and novel methods were employed to extract EEG features that capture the interplay between various brain regions to map human cognition.

Mutual Information-based features

Similar to the method presented above in Chapter 5, the conditional entropy using mutual information between all possible pairs of EEG electrodes for a given trial were calculated. Hence, for 14 EEG electrodes, 91 EEG features were calculated based on this measure.

Deep Learning-based features

The most commonly used EEG features are the calculation of power-spectrum density (PSD) of different EEG bands. But, these features in themselves do not take into account the EEG-topography i.e. the location of EEG electrodes for a particular EEG band. Hence, similar to the method in Chapter 5, EEG-topography was exploited for extracting information regarding the interplay between different brain regions.

Since the 2D spectrum image of the brain PSD has the information about amplitude distribution of each frequency band across the brain, the PSD of three EEG bands namely theta (4-7 Hz), alpha (7-13 Hz) and Beta (13-30 Hz) for all the EEG channels were calculated. The choice of these three specific EEG bands was made since they are the most commonly used bands and are thought to carry a lot of information about human cognition. The PSD for each band thus calculated were averaged over the complete trial. These features from different EEG channels were then used to construct a two-dimensional EEG-PSD heatmap for each of the three EEG bands using bicubic interpolation. These heat-maps now contained the information related to EEG topography in addition to spectrum density at each of these locations.

Fig. 6.1 shows these 2-D heatmaps for each of the three EEG bands. As can be seen from

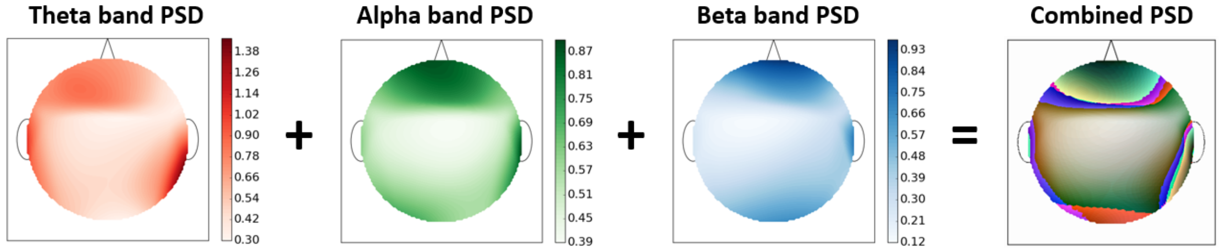


Figure 6.1: PSD heat-maps of the three EEG bands i.e. theta (red), alpha (green), and beta (blue) EEG bands are added according to respective color-bar range to get combined RGB heat-map image.(Circular outline, nose, ears, and color-bars have been added for visualization only. All units are in Watt per Hz.)

the figure, each of the three EEG bands was plotted using a single color channel i.e. red, green, and blue. These three color band images were then added to get a color RGB image containing information from the three EEG bands. The three color band images were added in proportion to the amount of EEG power in the three bands using alpha blending [122] by giving weights to the three individual bands' images by normalizing them using the highest value in the image. Hence, this procedure was able to represent the information in the three EEG bands along with their topography using a single color image. The interaction through the mixture of these three colors (thus forming new colors by adding these primary colors) in various quantities was contained the information regarding the distribution of power spectrum density across the various brain regions.

Notably, the PSD in three EEG power bands (theta, alpha, and beta) were computed separately because these three EEG power bands contribute the most towards human cognition. Subsequently, the power in these EEG bands were averaged separately for each band and not together. This process was done so that the RGB colored image with three EEG bands could be constructed utilizing the averaged power of theta, alpha, and beta band separately. It was to execute this complete procedure of generating the RGB colored image with each color band representing the averaged PSD of a particular EEG band that it was not possible to directly average the PSD values between 4-30 Hz.

Since it is not possible to train a deep neural network from scratch without thousands of trials from the EEG data (and no such dataset currently exists in a driving scenario), the combined colored image representing EEG-PSD with topography information was then fed to a pre-trained deep learning-based VGG-16 convolution neural network [97] to extract features from this image. This network has been trained with more than a million images for 1,000 object categories using the Imagenet Database [123]. Specifically, the VGG-16 deep learning network used has 16 total layers (as shown in Fig. 4), the input to the network being 224×224×3 sized colored image and Rectified Linear Unit (ReLU) optimizer as the activation function for all hidden layers. The convolution layers are 3×3 kernel size while the max-pooling layers are 2×2 kernel size and finally, there are three layers of the VGG-16 network which are fully connected ones.

Previous research studies [124, 125] have shown that the features from such an "off-the-shelf" neural network can be used for various classification problems with good accuracy. Even for the research problems where the neural networks were trained on a different vision-based problem and applied to a totally different application they still worked very well [125, 159, 160]. This is mostly because the low-level features such as texture, contrast, etc. reflected in the initial layers of the Convolution Neural Network (CNN) are ubiquitous in any type of image. The EEG-PSD colored image was resized to 224×224×3 for input to the network. The last layer of the network classifies the image into one of the 1000 classes but since the interest was only in "off-the-shelf" features, 4,096 features were extracted from the last but one layer of the network. The EEG features from this method were then combined with those from the previous one for further analysis.

6.2.2 PPG-based Feature Extraction

PPG measures the changes in blood volume in the microvascular tissue bed. This is done to assess the blood flow as being modulated by the heart beat. Using a simple peak detection algorithm on the PPG signal, it is possible to calculate the peaks of the blood flow and measure the subject's heart rate in a much more wearable manner than a traditional electrocardiogram (ECG) system. The PPG signal was recorded using an armband (Biovotion) that measures PPG at a sampling rate of 51.2 Hz.

HRV and Statistical Time-domain features

Heart-rate variability (HRV) has shown to be a good measure for classifying cognitive states such as emotional valence and stress [126]. HRV is much more robust than heart rate (HR) which changes slowly and generally only correspond to physical exertion. A moving-average filter with a window length of 0.25 seconds for filtering the noise in the PPG data was first used for each trial. The filtered PPG data so obtained was then scaled between 0 and 1 and subsequently a peak-detection algorithm [127] was applied to find the inter-beat intervals (RR) for the calculation of HRV. The minimum distance between successive peaks was taken to be 0.5 seconds to remove any false positives as in Fig. 6.2.

HR is defined as the total number of peaks per minute in the PPG. pNN50 algorithm [58] was then used to calculate HRV from RR intervals. To explore the statistics related to the PPG wave itself in time-domain, six statistical features were calculated on the PPG wave as defined in [131]. These features mapped various trends in the signal by the calculation of mean, standard deviation, etc. at first and subsequent difference signals formed using the original signal.

Spectrogram Deep Learning-based features

Recent research studies have shown promising results in analyzing PPG in the frequency domain for applications such as blood pressure estimation and gender identification [161, 162].

The frequency range of PPG signals is low and hence the focus was only on 0-5 Hz range. Fig. 5.2 shows the generated frequency spectrogram [130] for this frequency range for the PPG signal in a trial. The different color values generated using the “Parula” color-map shows the intensity of the spectrogram at a specific frequency bin. Then, the spectrogram images were resized to feed them to the VGG-16 network, and after which the 4,096 extracted features were extracted from the VGG-CNN network. Time-domain statistical and HRV features from the

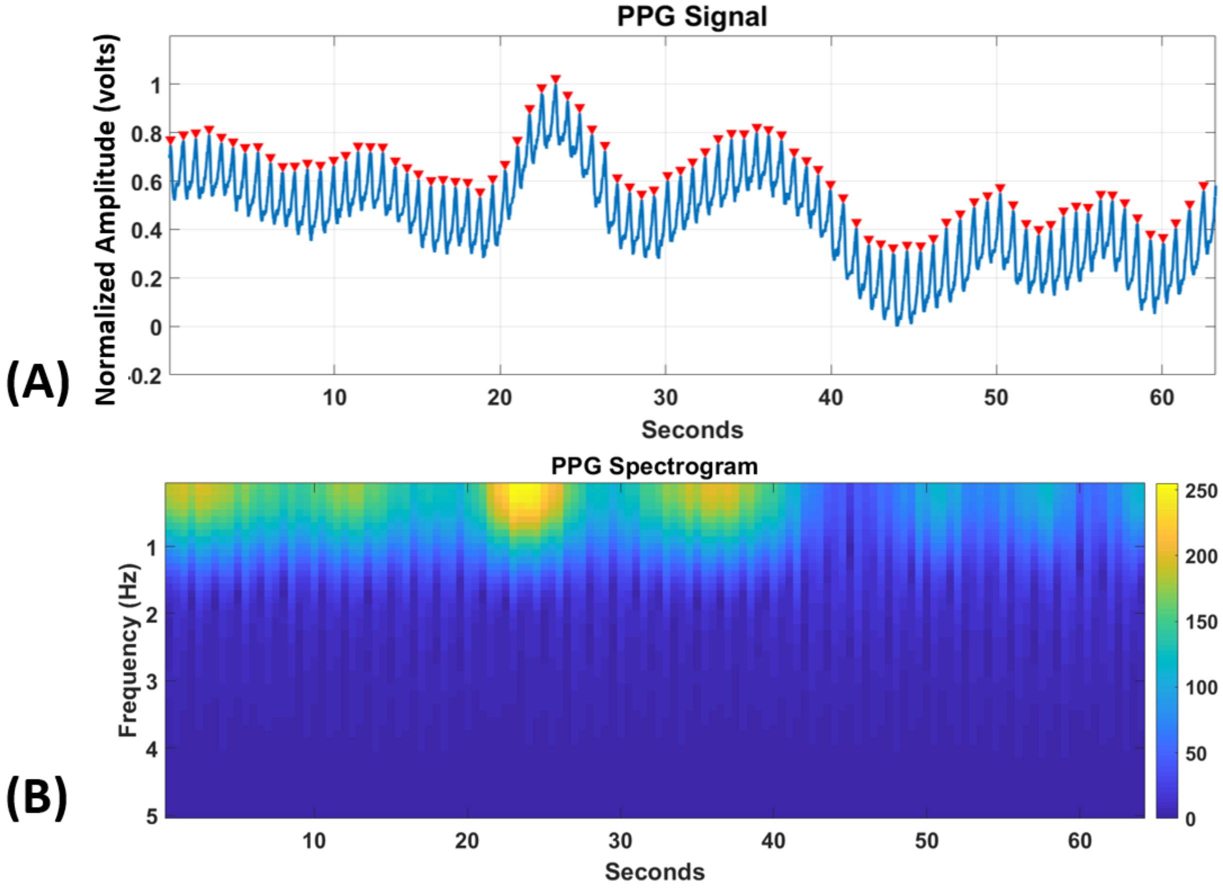


Figure 6.2: For a trial, PPG signal with peaks (in red) being detected for the calculation of RRs and HRV (above), and PPG spectrogram (below).

method above were concatenated with these features for further analysis.

6.2.3 GSR-based Feature Extraction

The feature extraction pipeline on the GSR signal was similar to that on the PPG. The same two methods that were applied to the PPG were used for GSR too. Same as PPG, the signals are sampled at 51.2 Hz by the device.

Statistical Features

The GSR data were first low-pass filtered with a moving average window of 0.25 seconds to remove any bursts in the data. Eight features based on the profile of the signal were then calculated. The first two of these features were the number of peaks and the mean of absolute heights of the peaks in the signal. Such peaks and their time differences may prove to be a good measure of arousal. The remaining six features were calculated as in [131] like the PPG signal above. For this time-series analysis of GSR signal, the features computed based on the peaks of GSR takes into account the Skin Conductance Response (SCR) i.e. the "peaks" of the activity while other features based on the GSR signal profile (by calculating mean and standard deviation of successive differences of the signal) accounts for the Skin Conductance Level (SCL).

Deep Learning-based features of GSR Spectrogram

Since GSR signals change very slowly the focus was only on the 0-2 Hz frequency range. A spectrogram image was generated for GSR in the above frequency range for each trial. This choice of utilizing low-frequency features was done because similar to PPG, GSR signals change very slowly. The spectrogram image was then sent to the VGG-16 deep neural network and extract the most significant 4,096 features from the same. These features were then concatenated with the features from the time-domain analysis.

6.2.4 Facial Expression-based Feature Extraction

As discussed above, the analysis of facial expressions has been the preferred modality for driver attention analysis. Hence, the goal was to use this method to compare it against the bio-sensing modalities. Furthermore, most of the research work in this area has been done by tracking fixed localized points on the face based on face action units (AUs).

First, the face region was extracted from the frontal body image of the person captured by the camera for each frame. This was done by fixing a threshold on the image size to reduce its extreme ends and placing a threshold of minimum face size to be 50×50 pixels. This resizing was done to remove any false positives and decrease the computational space for face detection. The Viola-Jones object detector was then used with Haar-like features [92] to detect the most likely face candidate.

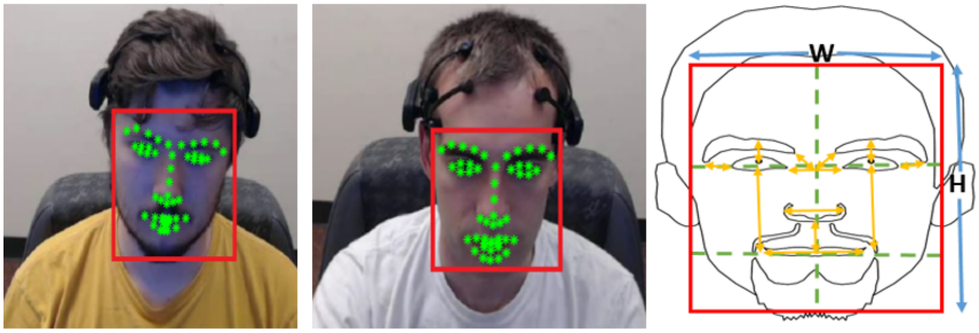


Figure 6.3: Detected face (marked in red) and face localized points (marked in green) for two participants (left and center) in the study, and some of the features (marked in yellow) computed using the coordinates of the face localized points. These features were then normalized using the size of the face in the camera i.e. number of pixels in height (H) and width (W)

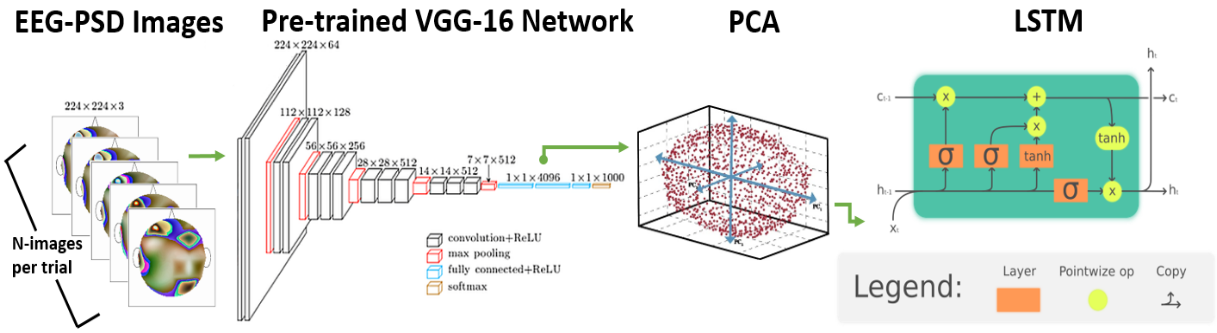


Figure 6.4: Network architecture for EEG-PSD trend-based Deep Learning method.

Facial-points Localization-based features

Face action units have been used for a variety of applications ranging from affective computing to face recognition [68]. Facial Action Coding System (FACS) is the most commonly used method to code facial expressions and map them to different emotional states [67]. The goal was to use face localized points similar to the ones used in FACS without identifying facial expressions such as anger, happiness, etc. since they are not highly relevant in the driving domain and short time intervals. The use of FACS initially involves the identification of multiple facial landmarks that are then tracked to map the changes in facial expressions. The state-of-the-art

Chehra algorithm [93] was applied to the extracted face candidate region from above. This algorithm outputs the coordinates of 49 localized points (landmarks) representing various features of the face as in Fig. 6.3. The choice of this algorithm was done because of its ability to detect these landmarks through its pre-trained models and hence not needing training for any new set of images. These face localized points were then used to calculate 30 different features based on the distances such as between the center of the eyebrow to the midpoint of the eye, between the midpoint of nose and corners of the lower lip, between the midpoints of two eyebrows, etc. and the angles between such line segments. To remove variations by factors such as distance from the camera and face tilt, these features were normalized using the dimensions of the face region. All these features were calculated for individual frames, many of which make a trial. Hence, to map the variation in these features across a trial (which may directly correspond to driver's attention and driving condition) mean, 95^{th} percentile (more robust than maximum), and the standard deviation of these 30 features were calculated across the frames in the trial. In this manner, 90 features were computed based on face-localized points from a particular trial.

Deep Learning-based features

For the extraction of deep learning-based features from face images, the VGG-Faces network was used instead of VGG-16. This was done to extract features more relevant to faces since the VGG-Faces network has been trained on more than 2.6 million face images from more than 2,600 people rather than on various object categories in the VGG-16 network. Each face region part was sent to the network and extracted the most significant 4,096 features. To represent the changes in these features across the trial i.e. across the frames, mean, 95^{th} percentile, and standard deviation of the features across the frames in a trial were computed.

6.2.5 Assessing Trends of EEG/Face Features using Deep Learning

The EEG features discussed in section 6.2.1 above were computed over the whole trial such as by generating a single EEG-PSD image for a particular trial. This is a special case when the data from the whole trial is being averaged. Here, a novel method is proposed to compute the trend of EEG features i.e. their variation in a trial based on deep learning. To compute features with more resolution, multiple EEG-PSD images were generated for successive time durations in a trial. One image per second of the data was generated for driver attention analysis and 30 images were used for every second for a 2-second incident classification analysis as detailed below in the Quantitative Analysis section. Fig. 6.4 shows the network architecture for this method. The EEG-PSD images were generated for multiple successive time durations in a trial, each of which was then sent to the VGG-16 network to obtain 4,096 most significant features. Similarly, this process was done for conditional entropy features by calculating this over multiple periods in a trial rather than once on the whole trial. Principal component analysis (PCA) [60] was then used to reduce the feature size to 60 to save computational time in the next step. These $60 \times N$ (N = number of successive time intervals) features were then sent as input to a Long Short Term Memory (LSTM) network [133].

In particular, the use of LSTM was motivated by extracting information from sensor modalities with higher temporal resolution. Example: For the EEG and modality, the extraction of deep learning features without LSTM was being done by representing the whole trial with a single 2D EEG image. This was a sort of average power-spectrum density image for the whole trial and thus had a bad temporal resolution. Instead, one such power-spectrum density image was calculated for every second to observe for the 2D image patterns change from second to second. This high temporal resolution was modeled using LSTM which further increases the accuracy by utilizing these shifting patterns.

The LSTM treats each of these features as a time-series and was trained to capture the trend in each of them for further analysis. This method could only be applied when the time duration of the trials is fixed since the length of each time series should be the same. Hence, this method was applied only in the trials used for detecting hazardous/non-hazardous situations and on EEG and face i.e. vision sensor modalities.

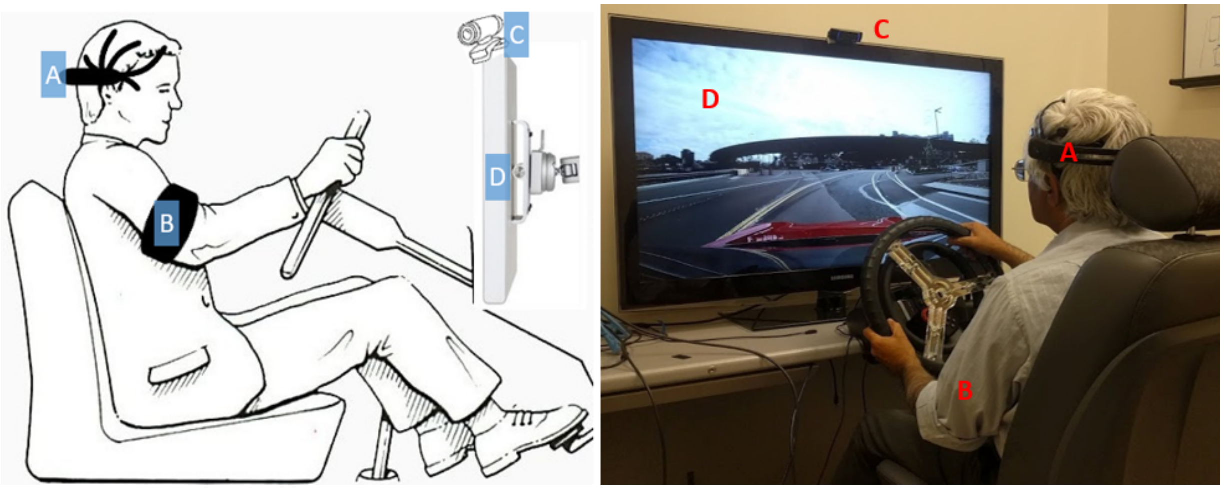


Figure 6.5: The experiment setup for multi-modal data collection. (A) EEG Headset, (B) PPG and GSR armband, (C) External camera, and (D) Driving videos displayed on the screen. The subject sits with her/his arms and feet on a driving simulator with which s/he interacts while watching the driving videos.

6.3 Dataset Description

Fig. 6.5 shows the experimental setup for data collection with driving videos used as the stimulus in the experiment. Twelve participants (most of them in their 20s with two older than 30 years) based in San Diego participated in the study. The participants were comfortably seated and equipped with EEG headset (Emotiv EPOC) containing 14 EEG channels (sampling rate of 128 Hz.) and an armband (Biovotion) for collecting PPG and GSR (sampling rate of 51.2 Hz.). This EEG headset was chosen since it is easily wearable and does not require the application of electrode gel. This made the headset conform more closely to real-world applications such as in the driving context. However, these advantages came at the cost of two limitations, namely, lower sampling rate and fewer EEG channels as compared to bulky EEG headsets used in the

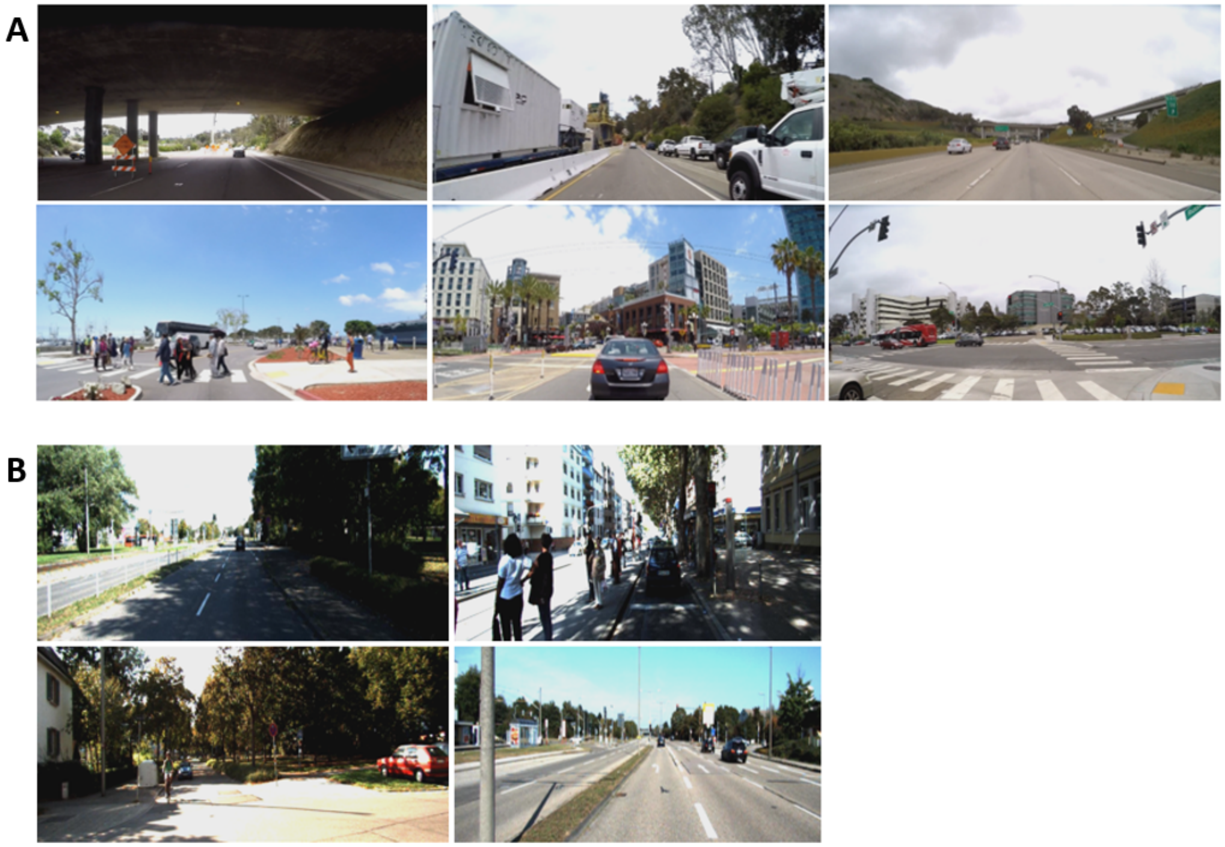


Figure 6.6: Various image instances with varying illumination conditions and types of roads (street, single-lane, highway, etc.) from (A) LISA Dataset and (B) KITTI Dataset.

laboratory. The positioning of the GSR sensor was however sub-optimal since it was not placed at the palm or the feet. This choice was driven by the practicality of data collection in the driving scenario since users interact with multiple vehicle modules from their palms and feet during driving. The facial expressions of the subject were recorded using a camera in front of him/her. The participants were asked to use a driving simulator that they were instructed to control as per the situation in the driving stimulus. For example, if there was a “red light” or “stop sign” at any point in a driving stimulus video, the participants should press and hold the brake.

For consistency between this and previous studies [10, 158], 15 video sequences from the KITTI dataset [156] were used. These previous research studies have used the same dataset but without detailing the exact image sequences used to generate the video sequences. Thus, the video sequences in this study chosen based on external annotation by two subjects to judge them based on potential hazardous events in them. These video sequences ranged from 14 to 105 seconds. These video sequences were recorded at 1242×375 resolution at 10 frames-per-second (fps). The videos were resized to 1920×580 to fit the display screen in a more naturalistic manner. But, video sequences from the KITTI dataset suffer from three limitations namely low resolution, low fps, and few sequences of driving on highways. Additionally, since the images in the KITTI dataset were captured at 10 fps it may elicit steady-state visual-evoked potential (SSVEP) in EEG [163]. This is undesirable because the focus should be exclusively on driver attention and hazardous/non-hazardous events analysis whereas SSVEP might act as a noise in the process.

Table 6.1: Table showing the various parameters pertaining to datasets and features used in evaluation.

Dataset	**KITTI**	**LISA**
Number of video sequences	15	20
Time Duration per video (s)	14-105	30-50
Frames per second	10	30
Video resolution	1920×580	1920×1200
Sensor Modality	**Traditional Features**	**Deep Learning Features**
EEG	96 (Conditional Entropy)	4096 (EEG-PSD 2-D spectrum image)
PPG	7 (HRV and Statistical)	4096 (PPG spectrogram image)
GSR	8 (Statistical)	4096 (GSR spectrogram image)
Face video	30 (Face AUs-based)	4096 (Face image-based)

Hence, the authors also collected a dataset of 20 video sequences containing real-world driving data on freeways and downtown San Diego, California. This dataset was collected using a LISA-T vehicle testbed in which a Tesla Model 3 is equipped with 6 external facing GoPro cameras. It is also to be noted that while recording these videos the vehicle was in the autonomous driving mode making LISA dataset the first of its kind. The cameras were operating at 122 degrees field-of-view which is very representative of the human vision. Furthermore, these video sequences were presented on a large screen (45.9 inches diagonally) at a distance of a meter from the participants to model a real-world driving scenario. These video sequences ranged from 30 to 50 seconds in length and were shown to the participants with 1920×1200 resolution at 30 fps.

External annotation was done to classify parts of the video sequences from both datasets into hazardous/non-hazardous events. For example, an event where a pedestrian suddenly appears to cross the road illegally was termed hazardous whereas an event where a stop sign can be seen from a distance and the vehicle's speed is decreasing was termed non-hazardous. External annotation was performed to classify every video sequence into how attentive the driver ought to be in that particular sequence. Table 6.1 catalogs the different datasets and features that were used in the evaluation of the proposed pipeline.

6.4 Quantitative analysis of multi-modal bio-sensing and vision sensor modalities

This section presents the evaluation results using various singular and multiple modalities for driver attention analysis and hazardous/non-hazardous instances classification. First, the videos in both datasets were externally annotated by two annotators for low/high driver attention required. For example, the video instances where the car is not moving at all were characterized as low attention instances whereas driving through narrow streets with pedestrians on the road were labeled as instances with high driver attention required. Hence, among the 35 videos (15 from KITTI dataset and 20 from LISA dataset), 20 were characterized as requiring low-attention and 15 as high-attention ones.

Second, 70 instances, each two-second long video clips were found in the videos and were characterized as hazardous/non-hazardous. Fig. 6.7 presents some examples of instances from both categories. As an example, a pedestrian suddenly crossing the road "unlawfully" or a vehicle overtaking suddenly represents hazardous events whereas "red" traffic sign at a distance and a pedestrian at a crossing with ego vehicle not in motion are examples of non-hazardous events. Among the 70 instances, 30 instances were labeled as hazardous whereas rest were labeled as non-hazardous. Hence, the goal was to classify such instances in a short period of two seconds using the above modalities. Since PPG and GSR have low temporal resolution and do not reflect changes in such short time intervals, only facial features and EEG were used for hazardous/non-hazardous event classification.

For each modality, PCA [60] was first used to reduce the number of features from the above algorithms to 30. Extreme learning machines (ELM) [62] were then employed for classification. The choice of using ELM over other feature classification methods was driven by previous studies that have shown how ELM performs better for features derived from bio-signals [134, 164]. These features were normalized between -1 and 1 across the subjects before training. A single hidden layer ELM was used with a triangular basis function for activation. For the method with trend-based temporal EEG and face feature data, two-layer LSTM was used with 200 and 100 neurons in respective layers instead of ELM for classification. The LSTM network's training was done using stochastic gradient descent with a momentum (SGDM) optimizer. Leave-one-subject-out cross-validation was used for each case. This meant that the data from 11 subjects (385 trials) were used for training at a time and the classification was done on the 35 trials from the remaining 12^{th} subject. This choice of cross-validation was driven by two factors. First, this method of cross-validation is much more robust and less prone to bias than models such as leave-one-sample-out cross-validation that constitutes training data from all the subjects at any given time. Second, since the data contained 420 trials only as opposed to thousands of trials for any decent image-based deep learning dataset, it does not make sense to randomly divide such a small number of trials to training, validation and test sets since it might introduce bias by uneven division across trials from individual subjects.

Both of the feature classification methods i.e. LSTM-based and ELM-based were used independently for feature classification with labels. When a higher temporal resolution was taken into consideration i.e. trends in a series of EEG-PSD images, then the LSTM-based method was used for feature classification. This is because now the features vary as a time series for each trial and ELM cannot be used for such a time-series based classification. The ELM-based method was performed for the other case i.e. the case when high temporal resolution data (multiple data point features for each trial) were not present. The data from the complete trial was represented by a single (non-varying in time) value for each feature.

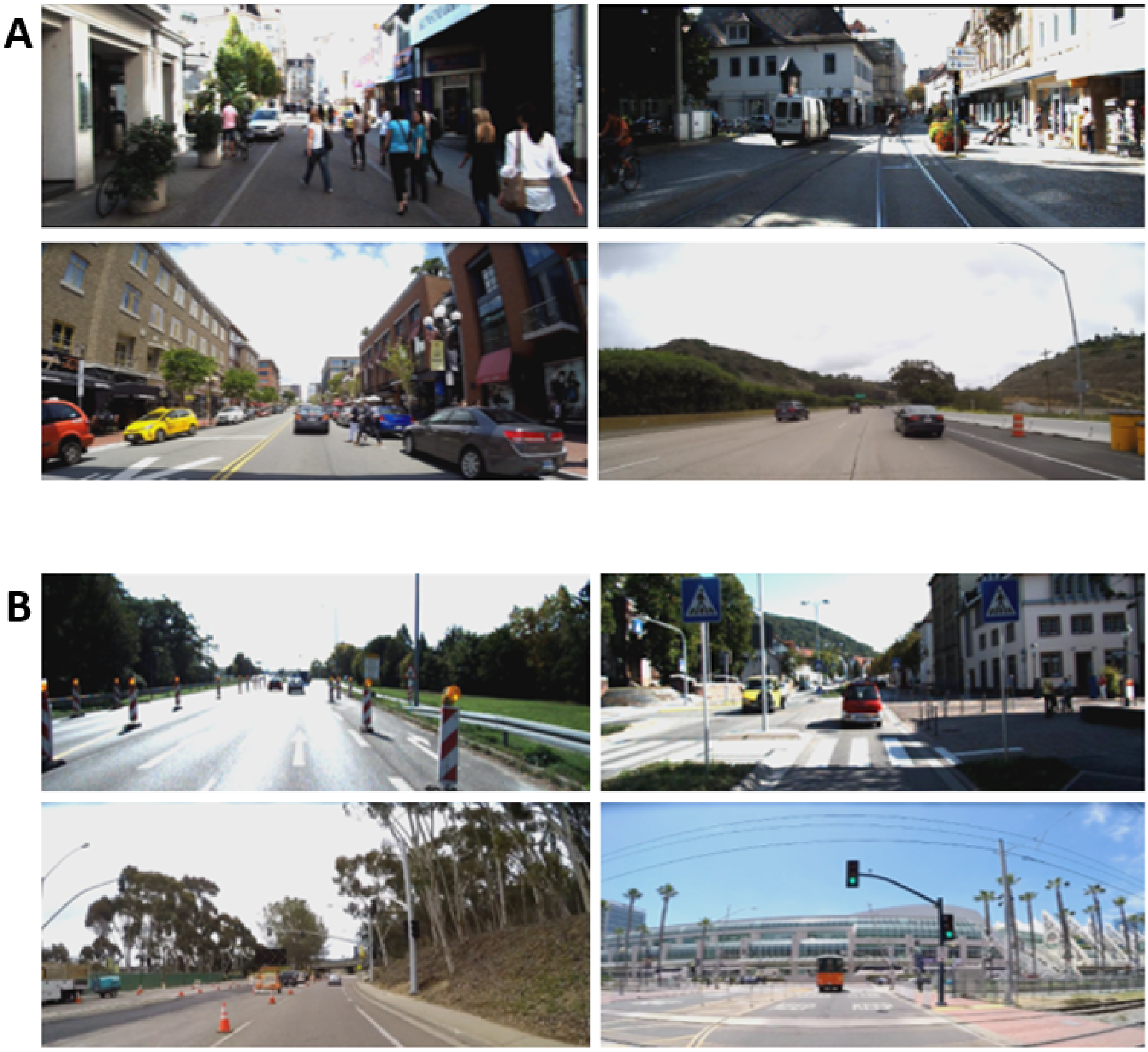


Figure 6.7: (A) Examples of 2-second incidents classified as hazardous. Examples include pedestrians crossing the street without a crosswalk while the ego vehicle is being driven and another vehicle overtaking suddenly. (B) Examples of 2-seconds incidents classified as non-hazardous. Examples include stop signs and railway crossing signs. For each category, the top images are from KITTI dataset whereas the bottom images are from LISA dataset.

6.4.1 Evaluating Attention Analysis Performance

This section evaluates the performance of assessing the driver's attention using single and multiple modalities across the video trials. For all the four modalities, the features as defined above were calculated for data from each video trial. The ELM-based classifier was then trained based on each video trial divided into one of the two classes representing low-attention and high-attention required by the driver.

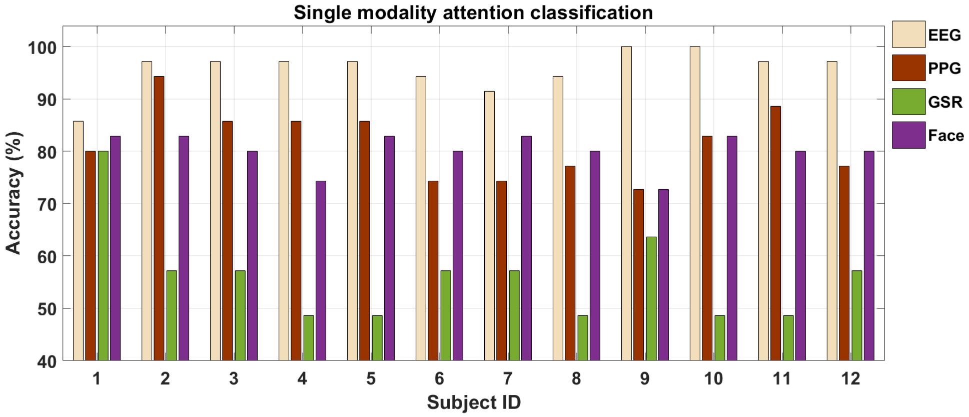


Figure 6.8: Single modality classification performance for driver attention analysis.

Single Modality Analysis

To compare the performance among the different modalities, the number of neurons in the hidden layer was set to 170 for each of the modality. Fig. 6.8 shows these results. Clearly, EEG performs the best among the four modalities for driver attention classification. The average classification accuracy for EEG, PPG, GSR, and face-videos was $95.71 \pm 3.95\%, 81.54 \pm 6.67\%, 56.02 \pm 3.04\%$, and $80.11 \pm 3.39\%$ respectively. The AUC (area under the curve) for the above four cases were $0.84 \pm 0.01, 0.83 \pm 0.02, 0.71 \pm 0.19$, and 0.79 ± 0.18 respectively. The statistical t-test was also performed on the above observations. The p-values for the above four classification cases were 10^{-5}, 10^{-6}, 10^{-3}, and 10^{-6} respectively. Hence, GSR performs only at about chance level whereas on average PPG and face videos perform equally well. Thus, it was observed that the sensor modalities with good temporal resolution i.e. EEG and vision perform better or at least as good as the ones with low temporal resolution (PPG and GSR) thus evaluating this study's first hypotheses. Pairwise t-test were performed for all six pair combinations of the above four signal modalities and the p-values were less than 0.05 for all cases. The statistical tests showed the results to be statistically significant. It was also observed that for all the subjects except one, EEG's classification accuracy is above 90% while for three modalities (EEG, GSR, and vision) the standard deviation in performance across the subjects is not too high.

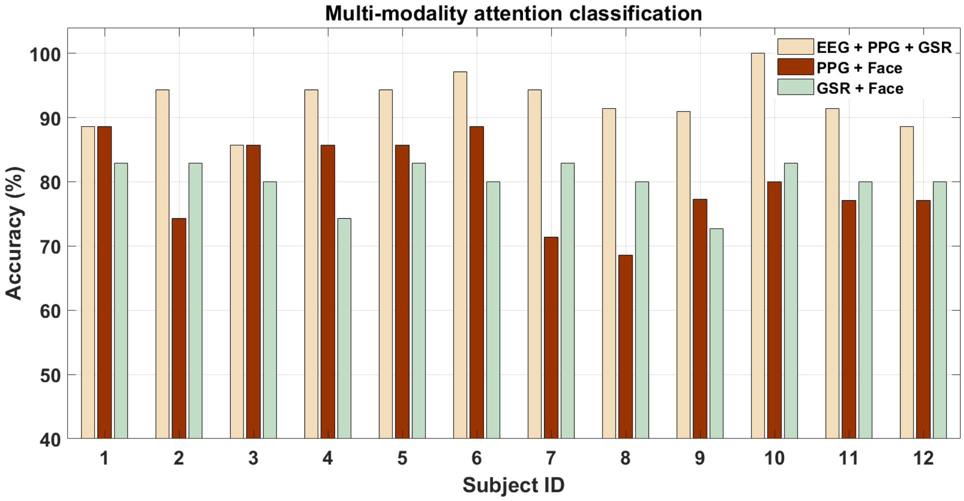


Figure 6.9: Multi-modality classification performance for driver attention analysis.

Multi-modality Analysis

Fig. 6.9 shows that on combining EEG with PPG and GSR there is no increase in the performance across the subjects (it might be that for a few subjects this is not the case). When the features from the low-performing (and poor temporal resolution) modalities i.e. PPG and GSR are combined with EEG, the performance is not as good as EEG alone for most of them. The mean accuracy across all the subjects was $92.58 \pm 3.96\%, 80.11 \pm 3.39\%$, and $80.01 \pm 6.78\%$ for the three cases respectively, all of which were significantly above the chance accuracy. The AUC for the above three cases were $0.85 \pm 0.01, 0.80 \pm 0.03$, and 0.80 ± 0.01 respectively. The p-values for the above four classification cases were 10^{-6}, 10^{-6}, and 10^{-3} respectively. To compare the different signal combinations, pairwise t-test was performed for the above cases. The p-values of pairwise t-test for multi-modal attention classification were 10^{-6} between (EEG + PPG + GSR) and (GSR +Face) cases and 0.9 between (GSR + Face) and (PPG + Face). Finally, pairwise t-test analysis was also performed between multi-modality and single-modality cases and found that the p-values between all four singular modalities (EEG, PPG, GSR, and Face) and the three multi-modality cases mentioned above were less than 0.05. These p-values thus denote that not all signal combinations between multi-modality cases are statistically significant in a pairwise manner while those between singular and multi-modality cases were statistically significant. Hence, it was observed it is not always beneficial to use features from multiple sensor modalities. For most of the subjects and modalities, the fusion of features does not perform better at all and hence may not be advantageous in this case. This might be due to the vast difference in the performance of each modality when used independently, based on the subject's physiology. This leads to an increase in performance for some of the subjects but not for all. But, for all combinations of sensor modalities, it was observed that the accuracy values were as good or better than using individual sensor modalities. This proves this study's first hypotheses about the performance improvement that could be gained by the use of multiple sensor modalities.

6.4.2 Evaluating Hazardous/Non-hazardous Incidents Classification

This section presents the results of the evaluation of the modalities over very short time intervals (2 seconds) pertaining to hazardous/non-hazardous driving incidents as shown in Fig. 6.7. Since GSR and PPG do not provide such a fine temporal resolution, these modalities were not used for this evaluation. This is because GSR changes very slowly i.e. take more than a few seconds to vary and PPG for a very short period such as 2 seconds would mean only 2-4 heartbeats which are not enough for computing heart-rate or heart-rate variability. Previous studies to assess human emotions using GSR and PPG on the order of multiple seconds (significantly greater than two seconds hazardous incident evaluation for driving context) [165]. Also, the subjects cannot tag the incidents while they are participating in the driving simulator experiment and hence these incidents were marked by the external annotators.

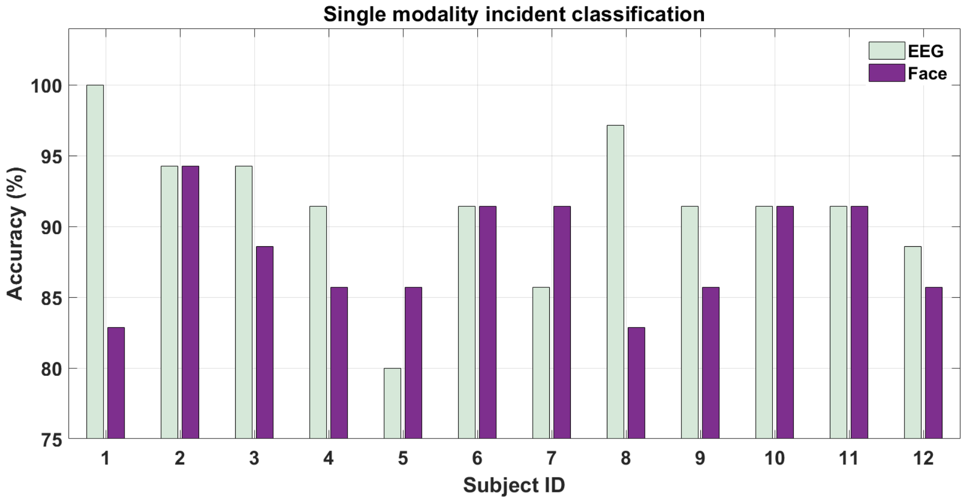


Figure 6.10: Single modality classification performance for driver attention during hazardous/non-hazardous incidents.

Single-modality Analysis

Fig. 6.10 shows the results for classifying hazardous/non-hazardous incidents using EEG and face-expression features. As the figure shows, the accuracy for both modalities for all the subjects is well above chance level (50%). The inter-subject variability for different sensor modalities can also be visualized from the above figure. For example, EEG outperforms face-based features for half of the participants but not for the other half. This variation in results is natural since some people tend to be more expressive with their facial expressions while on the other hand, the “perceived hazardousness” of a situation varies across subjects. The mean accuracy among subjects were $91.43 \pm 5.17\%$ and $88.10 \pm 3.82\%$ for EEG- and face-based features respectively. The AUC for these two cases were 0.85 ± 0.02 and 0.84 ± 0.02 respectively. The p-values for the above two classification cases were 10^{-5} and 10^{-5} respectively. Finally, a statistical pairwise t-test was also performed on the above two sensor modalities and the p-value was found to be 0.06. Thus, statistical significance was found within EEG and Face sensor modalitiy pair. Since the evaluation was done on 2-second time intervals i.e. without a lot of data it is notable that such a high mean accuracy for both modalities was only possible due to using deep learning-based features in addition to the traditional features for both modalities. This is further substantiated by the fact that an EEG system was used with a much lesser number of channels than such previous studies using EEG [10]. This study also shows that using such deep learning features the developed method outperforms the previous results for EEG on the KITTI dataset [10, 158] in a similar experimental setup with hazardous/non-hazardous event classification. Specifically, the single-modality approach for both EEG (AUC 0.85) and Face-videos (AUC 0.84) outperform on both datasets the previous best result (AUC 0.79) shown only on the KITTI dataset using EEG alone in [10].

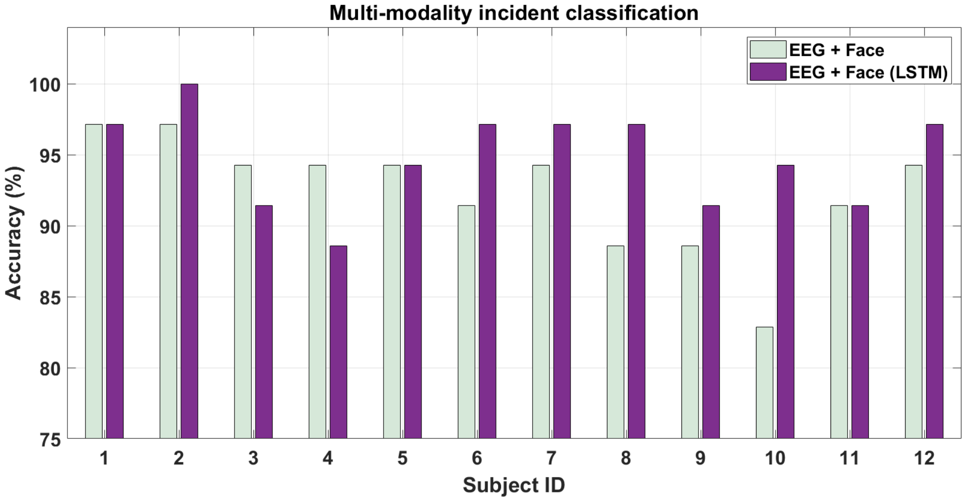


Figure 6.11: Multi-modality classification performance for driver attention during hazardous/non-hazardous incidents.

Multi-modality Analysis

This section presents the results for classifying hazardous/non-hazardous incidents using features from both EEG and face modalities. This was done in two ways. First, directly combining the features from single modality analysis by concatenating them. Second, by using an LSTM classifier over the features from both modalities calculated for every frame in the 2-second long sequence. The trend of these features is then fed to LSTM for training. Fig. 6.11 shows the results of the two approaches. As is clear from the figure, combining the features from the two modalities may or may not improve the performance, compared to using individual modalities shown earlier in Fig. 6.10. But, it is to be noted that on taking the trend of features i.e. increased temporal resolution into account, the performance of combining the modalities increases for most of the subjects. This can be seen from Fig. 6.10 where using the LSTM-based method, the accuracy increases across the subjects. The average accuracy across subjects being $92.38 \pm 4.10\%$ and $94.76 \pm 3.41\%$ respectively are also more than for singular modality analysis. The AUC for these two cases were 0.84 ± 0.01 and 0.88 ± 0.01 respectively. The p-values for the above two classification cases were 10^{-6} and 10^{-7} respectively. The pairwise t-test was also performed for the above multi-modal cases. The pairwise p-value was 0.10 for driving hazardous/non-hazardous incident classification between (EEG + Face) and (EEG + Face (LSTM)) cases and thus show that unlike individual modality cases, the pairwise statistical analysis was not significant for the above pair of signal modality combinations. Finally, a statistical pairwise t-test evaluation was performed between singular sensor modalities from Fig. 6.10 and multiple modalities from Fig. 6.11. The p-values between the two singular sensor modality cases (EEG and Face) and two multi-modality cases were less than 0.05. Thus, statistical significance was observed between singular and multi-modality sensor combinations as well. In conclusion, using multiple modalities with high temporal resolution (EEG and vision) may prove to be best when computing features over a short time duration with their trend (though it will involve more computational power).

Since EEG and Face modalities can be used in short-time intervals, Table 6.2 shows the mean accuracy across subjects for using EEG and faces separately and combining them for the two types of analysis done above. It is observable that the performance of EEG combined with faces can be better than when either modality is used independently for hazardous incident analysis when using features from the LSTM i.e. trend over the changes in features. However, adding

Table 6.2: The table shows the average classification accuracy across subjects using EEG and face-based features for driver attention analysis over the whole video and 2-second hazardous/non-hazardous incident classification. EEG features generally outperform Face features for both cases. Using LSTM i.e. better temporal resolution also increases the accuracy. LSTM could not be used in attention analysis since the duration of the videos varies widely among the datasets.

Modality	Attention Analysis	Incident Analysis
EEG	$95.71 \pm 3.95\%$	$91.43 \pm 5.17\%$
Faces	$80.11 \pm 3.39\%$	$88.10 \pm 3.82\%$
EEG + Faces	$95.10 \pm 3.62\%$	$92.38 \pm 4.10\%$
EEG + Faces (LSTM)	—	$94.76 \pm 3.41\%$

multiple modalities together without using trend-based LSTM analysis may not prove much beneficial. This answers the study's second hypotheses by showing that it is beneficial to use a fusion of the modalities if both modalities have a good temporal resolution to extract short-duration features over them to map the trend. Thus, connecting dots with the single-modality analysis, it can be surmised that multi-modality boosts performance over using individual modalities for hazardous/non-hazardous incident classification (like it did for driver attention analysis) while further improvement in performance is observed by utilizing higher temporal resolution using LSTMs.

6.5 Chapter Concluding Remarks

The use of multiple bio-sensing modalities combined with audio-visual ones is rapidly expanding. With the advent of compact bio-sensing systems capable of collecting data during real-world tasks such as driving, it is natural that this research area will gather more interest in the coming years. This work evaluated multiple bio-sensing modalities with the vision modality for driver attention and hazardous event analysis. A pipeline was also presented to process data from individual modalities by being able to use pre-trained convolution neural networks to extract deep learning-based features from these modalities in addition to traditionally used ones. In this process, this study was able to compare the performance of the modalities against each other while also combining them.

This chapter is in part a reprint of material that has been accepted for publication in the journal MDPI Brain Sciences (2020), by Siddharth Siddharth and Mohan M. Trivedi. The dissertation author was the primary author of this paper.

Chapter 7

Conclusion

Affective computing has predominantly focused on computer vision, which suffers from limitations such as dependence on illumination and privacy-related issues. On the other hand, bio-sensing modalities such as EEG suffer from low spatial resolution and motion artifacts. Thus, there is a need to implement a multi-modal sensory approach to study human affects.

In this dissertation, we presented a multi-modal approach to study human affects. The chapters of this dissertation were arranged in a systemic manner with each building on the content from the previous one. In the process, we presented a compact headset able to record multi-modal bio-sensing data in real-time and a real-world manner. Subsequently, such multi-modal sensor modalities were used with tools from signal processing and deep learning for designing emotion classification algorithms. These tools were developed in a scalable and robust manner. Such tools were utilized in applications ranging from assessing emotion states induced by emotional multimedia content to driver awareness.

As discussed in Chapter 2, there are five objectives that any multi-modal bio-sensing system for affective computing should achieve. Among those five, this dissertation focused on detecting and classifying emotional states, inferring emotional state from a minimal number of sensor modalities, and inferring the context in real-time to process the data. To take this research forward by utilizing such processed data for executing different tasks by the system and being able to achieve the above in real-world settings throughout the day are future research avenues. An insight from this dissertation is that since one of the main limitations of affective computing research being subjective feedback by users, it may be wise to rely rather on games (as we did in Chapter 3) to assess emotional states. Another insight from this research (as shown in Chapter 5 and 6) is that facial expressions and EEG that has high temporal resolution perform better for human affect recognition than those with low temporal resolution such as PPG and GSR.

The modular nature of the platform makes it apt for being used in other applications too such as studying cognitive attention and mental health. Specifically, the hardware platform developed in Chapter 3 can be used for collecting bio-sensing data for any application while the algorithms developed in Chapter 5 can be tuned for applications in medical diagnosis. Chapter 6 showed how such algorithms developed for classifying emotions worked well for studying driver awareness as well.

Bibliography

[1] T. C. Brickhouse and N. D. Smith, “Socrates on the emotions,” *PLATO JOURNAL: The Journal of the International Plato Society*, vol. 15, pp. 9–28, 2015.

[2] J. Tuske, *The Concept of Emotion in Classical Indian Philosophy*. Metaphysics Research Lab, Stanford University, fall 2016 ed., 2016.

[3] J. Tao and T. Tan, “Affective computing: A review,” *International Conference on Affective computing and intelligent interaction*, pp. 981–995, 2005.

[4] W. Jänig, “Neurocardiology: A neurobiologist’s perspective,” *The Journal of physiology*, vol. 594(14), 2016.

[5] S. Koelstra, C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, N. A., and I. Patras, “DEAP: A database for emotion analysis; using physiological signals,” *IEEE Transactions on Affective Computing*, vol. 3(1), pp. 18–31, 2011.

[6] J. A. M. Correa, M. K. Abadi, N. Sebe, and I. Patras, “AMIGOS: A dataset for affect, personality and mood research on individuals and groups,” *IEEE Transactions on Affective Computing*, 2018.

[7] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE Transactions on Affective Computing*, vol. 3(1), pp. 42–55, 2012.

[8] S. Katsigiannis and N. Ramzan, “DREAMER: A database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices,” *IEEE journal of biomedical and health informatics*, vol. 22(1), pp. 98–107, 2017.

[9] J. Morales, C. Díaz-Piedra, H. Rieiro, J. Roca-González, S. Romero, A. Catena, L. Fuentes, and L. Di Stasi, “Monitoring driver fatigue using a single-channel electroencephalographic device: A validation study by gaze-based, driving performance, and subjective data,” *Accident Analysis & Prevention*, vol. 109, pp. 62–69, 2017.

[10] H. Kolkhorst, W. Burgard, and M. Tangermann, “Decoding hazardous events in driving videos,” *Proceedings of the 7th Graz Brain-Computer Interface Conference*, 2017.

[11] P. Bizopoulos, G. I. Lambrou, and D. Koutsouris, "Signal2Image modules in deep neural networks for EEG classification," *IEEE 41st International Engineering in Medicine and Biology Conference (EMBC)*, 2019.

[12] H. Yoon and S. Chung, "EEG-based emotion estimation using bayesian weighted-log-posterior function and perceptron convergence algorithm," *Computers in biology and medicine*, vol. 43(12), pp. 2230–2237, 2013.

[13] J. Atkinson and D. Campos, "Improving BCI-based emotion recognition by combining EEG feature selection and kernel classifiers," *Expert Systems with Applications*, vol. 47, pp. 35–41, 2016.

[14] Y. Liu and O. Sourina, "EEG-based valence level recognition for real-time applications," *In Cyberworlds (CW), 2012 International Conference on*, pp. 53–60, 2012.

[15] S. Koelstra and I. Patras, "Fusion of facial expressions and EEG for implicit affective tagging," *Image and Vision Computing*, vol. 31(2), pp. 164–174, 2013.

[16] Z. Yin, M. Zhao, Y. Wang, J. Yang, and J. Zhang, "Recognition of emotions using multimodal physiological signals and an ensemble deep learning model," *Computer methods and programs in biomedicine*, vol. 140, pp. 93–110, 2017.

[17] A. Kołakowska, A. Landowska, M. Szwoch, W. Szwoch, and M. Wrobel, "Emotion recognition and its applications," *In Human-Computer Systems Interaction: Backgrounds and Applications, Springer, Cham.*, vol. 3, pp. 51–62, 2014.

[18] T. Eerola and J. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39(1), pp. 18–49, 2011.

[19] S. Katsigiannis and N. Ramzan, "Emotion and sociable humanoid robots," *International Journal of Human-Computer Studies*, vol. 59(1-2), pp. 119–155, 2003.

[20] C. Bell, P. Shenoy, R. Chalodhorn, and R. Rao, "Control of a humanoid robot by a noninvasive brain–computer interface in humans," *Journal of neural engineering*, vol. 5(2), p. 214, 2008.

[21] P. Bamidis, C. Papadelis, C. Kourtidou-Papadeli, C. Pappas, and A. B. Vivas, "Affective computing in the era of contemporary neurophysiology and health informatics," *Interacting with Computers*, vol. 16(4), pp. 715–721, 2004.

[22] W. Liu, W. Zheng, and B. Lu, "Multimodal emotion recognition using multimodal deep learning," *arXiv preprint arXiv:1602.08225*, 2016.

[23] K. LaFleur, K. Cassady, A. Doud, K. Shades, E. Rogin, and B. He, "Quadcopter control in three-dimensional space using a noninvasive motor imagery-based brain–computer interface," *Journal of neural engineering*, vol. 10(4), 2013.

[24] T. Carlson and J. del R. Millan, “Brain-controlled wheelchairs: a robotic architecture,” *IEEE Robotics & Automation Magazine*, vol. 20.1, pp. 65–73, 2013.

[25] S. Siddharth, T. Jung, and T. J. Sejnowski, “Multi-modal approach for affective computing,” *IEEE 40th International Engineering in Medicine and Biology Conference (EMBC)*, 2018.

[26] J. Lei, J. Sala, and S. Jasra, “Identifying correlation between facial expression and heart rate and skin conductance with imotions biometric platform,” *Journal of Emerging Forensic Sciences Research*, vol. 2(2), pp. 53–83, 2017.

[27] J. Kamienkowski, M. Ison, R. Quiroga, and M. Sigman, “Fixation-related potentials in visual search: A combined EEG and eye tracking study fixation-related potentials in visual search,” *Journal of vision*, vol. 12(7), pp. 4–4, 2012.

[28] L. Ackualagna and B. Blankertz, “Gaze-independent BCI-spelling using rapid serial visual presentation (RSVP),” *Clinical Neurophysiology*, vol. 124(5), pp. 901–908, 2013.

[29] R. Rawassizadeh, B. Price, and M. Petre, “Wearables: Has the age of smartwatches finally arrived?,” *Communications of the ACM*, vol. 58(1), pp. 45–47, 2015.

[30] T. Wyss, L. Roos, N. Beeler, B. Veenstra, S. Delves, M. Buller, and K. Friedl, “The comfort, acceptability and accuracy of energy expenditure estimation from wearable ambulatory physical activity monitoring systems in soldiers,” *Journal of Science and Medicine in Sport*, vol. 20, pp. S133–S134, 2017.

[31] G. Bertolaccini, I. Carvalho Filho, G. Christofoletti, L. Paschoarelli, and F. Medola, “The influence of axle position and the use of accessories on the activity of upper limb muscles during manual wheelchair propulsion,” *International Journal of Occupational Safety and Ergonomics*, vol. 24(2), pp. 311–315, 2018.

[32] H. Suryotrisongko and F. Samopa, “Evaluating OpenBCI Spiderclaw V1 headwear’s electrodes placements for brain-computer interface (BCI) motor imagery application,” *Procedia Computer Science*, vol. 72, pp. 398–405, 2015.

[33] W. Von Rosenberg, T. Chanwimalueang, V. Goverdovsky, D. Looney, D. Sharp, and D. Mandic, “Smart helmet: Wearable multichannel ECG and EEG,” *IEEE journal of translational engineering in health and medicine*, vol. 4, 2016.

[34] J. Patterson, D. McIlwraith, and G. Yang, “A flexible, low noise reflective PPG sensor platform for ear-worn heart rate monitoring,” *In 2009 sixth international workshop on wearable and implantable body sensor networks*, pp. 286–291, 2009.

[35] M. Poh, N. Swenson, and R. Picard, “Motion-tolerant magnetic earring sensor and wireless earpiece for wearable photoplethysmography,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 14(3), pp. 786–794, 2010.

[36] M. Duvinage, T. Castermans, M. Petieau, T. Hoellinger, G. Cheron, and T. Dutoit, "Performance of the Emotiv Epoc headset for P300-based applications," *Biomedical engineering online*, vol. 12(1), p. 56, 2013.

[37] Y. Chi, Y. Wang, Y. Wang, T. Jung, T. Kerth, and Y. Cao, "A practical mobile dry EEG system for human computer interfaces," *In International Conference on Augmented Cognition*, pp. 649–655, 2013.

[38] Y. Chi, Y. Wang, Y. Wang, C. Maier, T. Jung, and G. Cauwenberghs, "Dry and noncontact EEG sensors for mobile brain–computer interfaces," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 20(2), pp. 228–235, 2012.

[39] D. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32(3), pp. 478–500, 2010.

[40] F. Cornelissen, E. Peters, and J. Palmer, "The Eyelink Toolbox: eye tracking with MATLAB and the Psychophysics Toolbox," *Behavior Research Methods, Instruments, & Computers*, vol. 34(4), pp. 613–617, 2002.

[41] J. Morgante, R. Zolfaghari, and S. Johnson, "A critical test of temporal and spatial accuracy of the Tobii T60XL eye tracker," *Infancy*, vol. 17(1), pp. 9–32, 2012.

[42] P. Kassner, M., W., and A. Bulling, "Pupil: an open source platform for pervasive eye tracking and mobile gaze-based interaction," *In Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: Adjunct publication*, pp. 1151–1160, 2014.

[43] S. Siddharth, A. Patel, T. Jung, and T. Sejnowski, "An affordable bio-sensing and activity tagging platform for HCI research," *In International Conference on Augmented Cognition*, pp. 399–409, 2017.

[44] B. Widrow, J. Glover, J. McCool, J. Kaunitz, C. Williams, R. Hearn, J. Zeidler, J. Dong, and R. Goodlin, "Adaptive noise cancelling: Principles and applications," *Proceedings of the IEEE*, vol. 63(12), pp. 1692–1716, 1975.

[45] J. Lovelace, T. Witt, and F. Beyette, "Modular, bluetooth enabled, wireless electroencephalograph (EEG) platform," *In Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE*, pp. 6361–6364, 2013.

[46] O.-C. M. Mastinu, E. and B. Håkansson, "Analog front-ends comparison in the way of a portable, low-power and low-cost EMG controller based on pattern recognition," *In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 2111–2114, 2015.

[47] S. Hsu, T. Mullen, T. Jung, and G. Cauwenberghs, "Online recursive independent component analysis for real-time source separation of high-density EEG," *In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pp. 3845–3848, 2014.

[48] S. Makeig, A. Bell, T. Jung, and T. Sejnowski, "Independent component analysis of electroencephalographic data," *Advances in neural information processing systems*, pp. 145–151, 1996.

[49] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[50] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299, 2017.

[51] C. Kothe, "Lab streaming layer (LSL)," *https://github.com/sccn/labstreaminglayer*, vol. Last Accessed 2020, 2014.

[52] D. Altman and J. Bland, "Measurement in medicine: the analysis of method comparison studies," *The statistician*, vol. 32(3), pp. 307–317, 1983.

[53] X. Chen, Y. Wang, S. Gao, T. Jung, and X. Gao, "Filter bank canonical correlation analysis for implementing a high-speed SSVEP-based brain–computer interface," *Journal of neural engineering*, vol. 12(4), 2015.

[54] X. Chen, Y. Wang, M. Nakanishi, X. Gao, T. Jung, and S. Gao, "High-speed spelling with a noninvasive brain–computer interface," *Proceedings of the national academy of sciences*, vol. 112(44), pp. E6058–E6067, 2015.

[55] H. Wang and Y. Wang, "Convolutional neural network for target face detection using single-trial EEG signal," *In Engineering in Medicine and Biology Society (EMBC), 2018 40th Annual International Conference of the IEEE*, pp. 2008–2011, 2018.

[56] J. Kaufmann, S. Schweinberger, and A. Burton, "N250 ERP correlates of the acquisition of face representations across different images," *Journal of Cognitive Neuroscience*, vol. 21(4), pp. 625–641, 2009.

[57] J. Polich, "Updating P300: an integrative theory of P3a and P3b," *Clinical neurophysiology*, vol. 118(10), pp. 2128–2148, 2007.

[58] D. Ewing, J. Neilson, and P. Travis, "New method for assessing cardiac parasympathetic activity using 24 hour electrocardiograms," *Heart*, vol. 52(4), pp. 396–402, 1984.

[59] 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(8), pp. 1226–1238, 2005.

[60] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2(1-3), pp. 37–52, 1987.

[61] K. Mera and T. Ichimura, “Emotion analyzing method using physiological state,” *In International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 195–201, 2004.

[62] G. Huang, Q. Zhu, and C. Siew, “Extreme learning machine: theory and applications,” *Neurocomputing*, vol. 70(1-3), pp. 489–501, 2006.

[63] S. Siddharth, T. Jung, and T. J. Sejnowski, “Utilizing deep learning towards multi-modal bio-sensing and vision-based affective computing,” *IEEE Transactions on Affective Computing*, 2019.

[64] Y. P. Lin, Y. H. Yang, and T. P. Jung, “Fusion of electroencephalographic dynamics and musical contents for estimating emotional responses in music listening,” *Frontiers in Neuroscience*, vol. 8, p. 94, 2014.

[65] S. M. Alarcao and M. J. Fonseca, “Emotions recognition using EEG signals: A survey,” *IEEE Transactions on Affective Computing*, 2017.

[66] P. C. Petrantonakis and L. J. Hadjileontiadis, “A novel emotion elicitation index using frontal brain asymmetry for enhanced EEG-based emotion recognition,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 15(5), pp. 737–746, 2011.

[67] P. Ekman, *Facial action coding system.* Consulting Psychologists Press, 1978.

[68] Y. I. Tian, T. Kanade, and J. F. Cohn, “Recognizing action units for facial expression analysis,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23(2), pp. 97–115, 2001.

[69] J. J. Gross and R. W. Levenson, “Emotion elicitation using films,” *Cognition & Emotion*, vol. 9(1), pp. 87–108, 1995.

[70] W. Sato, M. Noguchi, and S. Yoshikawa, “Emotion elicitation effect of films in a japanese sample,” *Social Behavior and Personality: an international journal*, vol. 35(7), pp. 863–874, 2007.

[71] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural computation*, vol. 16(12), pp. 2639–2664, 2004.

[72] J. A. Russell, “A circumplex model of affect,” *Journal of Personality and Social Psychology*, vol. 39(6), pp. 1161–1178, 1980.

[73] A. Gabrielsson and E. Lindström, *Handbook of Music and Emotion: Theory, Research, Applications.* New York, NY: Oxford University Press, 2010.

[74] O. Lartillot, P. Toiviainen, and E. T, *A MATLAB Toolbox for Music Information Retrieval*. Springer, Berlin, Heidelberg, 2008.

[75] E. Asutay and D. Västfjäll, “Perception of loudness is influenced by emotion,” *PLoS ONE*, vol. 7(6), 2012.

[76] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Processing Letters*, vol. 6(1), pp. 1–3, 1999.

[77] L. Jaquet, B. Danuser, and P. Gomez, “Music and felt emotions: How systematic pitch level variations affect the experience of pleasantness and arousal,” *Psychology of Music*, vol. 42(1), pp. 51–70, 2014.

[78] H. L. Wang and L. F. Cheong, “Affective understanding in film,” *IEEE Transactions on circuits and systems for video technology*, vol. 16(6), pp. 689–704, 2006.

[79] R. F. Simons, B. H. Detenber, B. N. Cuthbert, D. D. Schwartz, and J. E. Reiss, “Attention to television: Alpha power and its relationship to image motion and emotional content,” *Media psychology*, vol. 5(3), pp. 283–301, 2003.

[80] B. Adams, C. Dorai, and S. Venkatesh, “Toward automatic extraction of expressive elements from motion pictures: Tempo,” *IEEE Transactions on Multimedia*, vol. 4(4), pp. 472–481, 2002.

[81] B. Castellano, “PySceneDetect,” Last accessed 2020.

[82] H. Zettl, *Sight, Sound, Motion: Applied Media Aesthetics*. Belmont, CA: Wadsworth, 3rd ed., 1998.

[83] Z. Rasheed, Y. Sheikh, and M. Shah, “On the use of computable features for film classification,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15(1), pp. 52–64, 2005.

[84] P. Valdez and A. Mehrabian, “Effects of color on emotions,” *Journal of experimental psychology*, vol. 123(4), pp. 394–409, 1994.

[85] D. Bordwell, K. Thompson, and J. Smith, *Film Art: An Introduction*. New York: McGraw-Hill, 7th ed., 2004.

[86] R. M. Haralick, K. Shanmugam, and I. Dinstein, “Textural features for image classification,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6, pp. 610–621, 1973.

[87] T. Mullen, C. Kothe, Y. Chi, A. Ojeda, T. Kerth, S. Makeig, G. Cauwenberghs, and T. Jung, “Real-time modeling and 3D visualization of source dynamics and connectivity using wearable,” *IEEE 35th International Engineering in Medicine and Biology Conference (EMBC)*, pp. 2184–2187, 2013.

[88] W. O. Tatum, “Ellen R. Grass Lecture: Extraordinary EEG,” *The Neurodiagnostic Journal*, vol. 54, pp. 3–21, 2014.

[89] C. H. Vanderwolf, “Are neocortical gamma waves related to consciousness?,” *Brain Research*, vol. 855(2), pp. 217–24, 2000.

[90] S. Siddharth, A. Patel, T. Jung, and T. J. Sejnowski, “A wearable multi-modal bio-sensing system towards real-world applications,” *IEEE Transactions on Biomedical Engineering*, vol. 66(4), pp. 1137–1147, 2018.

[91] A. Ojeda, “headModel,” Last accessed 2020.

[92] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2001.

[93] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, “Incremental face alignment in the wild,” *Proceedings of the 2014 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1859–1866, 2014.

[94] G. Pajares and J. M. De La Cruz, “A wavelet-based image fusion tutorial,” *Pattern recognition*, vol. 37(9), pp. 1855–1872, 2004.

[95] P. McCullagh and J. A. Nelder, *Generalized Linear Models.* New York: Chapman & Hall, 1990.

[96] X. Zhang, L. Yao, X. Wang, J. Monaghan, D. Mcalpine, and Y. Zhang, “A survey on deep learning based brain computer interface: Recent advances and new frontiers,” 2019.

[97] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014.

[98] N. Kriegeskorte, M. Mur, and P. A. Bandettini, “Representational similarity analysis-connecting the branches of systems neuroscience,” *Frontiers in systems neuroscience*, vol. 2, p. 4, 2008.

[99] J. Bassili, “Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face,” *Journal of Personality and Social Psychology*, vol. 37(11), pp. 2049–2058, 1979.

[100] W. J. Ray and H. W. Cole, “EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes,” *Science*, vol. 228(4700), pp. 750–752, 1985.

[101] S. Coyle, Y. Wu, K. Lau, S. Brady, G. Wallace, and D. Diamond, “Bio-sensing textiles-wearable chemical biosensors for health monitoring,” *In 4th International Workshop on Wearable and Implantable Body Sensor Networks (BSN 2007)*, pp. 35–39, 2007.

[102] G. Riva, F. Mantovani, C. Capideville, A. Preziosa, F. Morganti, D. Villani, A. Gaggioli, C. Botella, and M. Alcañiz, “Affective interactions using virtual reality: the link between presence and emotions,” *CyberPsychology & Behavior*, vol. 10(1), pp. 45–46, 2007.

[103] J. Millán, F. Renkens, J. Mourino, and Gerstner, “Noninvasive brain-actuated control of a mobile robot by human EEG,” *IEEE Transactions on Biomedical Engineering*, vol. 51(6), pp. 1026–1033, 2004.

[104] S. Samant, M. Chapko, and H. Seo, “Predicting consumer liking and preference based on emotional responses and sensory perception: A study with basic taste solutions,” *Food Research International*, vol. 100, pp. 325–334, 2017.

[105] H. Ehmen, M. Haesner, I. Steinke, M. Dorn, M. Gövercin, and E. Steinhagen-Thiessen, “Comparison of four different mobile devices for measuring heart rate and ECG with respect to aspects of usability and acceptance by older people,” *Applied ergonomics*, vol. 43(3), pp. 582–587, 2012.

[106] J. Bailenson, E. Pontikakis, I. Mauss, J. Gross, M. Jabon, C. Hutcherson, C. Nass, and O. John, “Real-time classification of evoked emotions using facial feature tracking and physiological responses,” *International journal of human-computer studies*, vol. 66(5), pp. 303–317, 2008.

[107] M. Yeasin, B. Bullot, and R. Sharma, “Recognition of facial expressions and measurement of levels of interest from video,” *IEEE Transactions on Multimedia*, vol. 8(3), pp. 500–508, 2006.

[108] S. Poria, E. Cambria, A. Hussain, and G. Huang, “Towards an intelligent framework for multimodal affective data analysis,” *Neural Networks*, vol. 63, pp. 104–116, 2015.

[109] S. Kahou, C. Pal, X. Bouthillier, P. Froumenty, C. Gulcehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. Ferrari, and M. Mirza, “Combining modality specific deep neural networks for emotion recognition in video,” *In Proceedings of the 15th ACM on International conference on multimodal interaction*, pp. 543–550, 2013.

[110] Z. Yu and C. Zhang, “Image based static facial expression recognition with multiple deep network learning,” *In Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 435–442, 2015.

[111] A. Gudi, H. Tasli, T. Den Uyl, and A. Maroulis, “Deep learning based FACS action unit occurrence and intensity estimation. in automatic face and gesture recognition (FG),” *In Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, vol. 6, pp. 1–5, 2015.

[112] A. Krizhevsky, V. Nair, and G. Hinton, “The CIFAR-10 dataset,” *http://www. cs. toronto. edu/kriz/cifar. html*, vol. 55, 2014.

[113] L. Deng, “The MNIST database of handwritten digit images for machine learning research [best of the web],” *IEEE Signal Processing Magazine*, vol. 29(6), pp. 141–142, 2012.

[114] M. Soleymani, M. Pantic, and T. Pun, “Multimodal emotion recognition in response to videos,” *IEEE Transactions on Affective Computing*, vol. 3(2), pp. 211–223, 2011.

[115] V. Gonuguntla, R. Mallipeddi, and K. Veluvolu, “Identification of emotion associated brain functional network with phase locking value,” *In Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference*, pp. 4515–4518, 2016.

[116] D. Wang and Y. Shang, “Modeling physiological data with deep belief networks,” *International journal of information and education technology (IJIET)*, vol. 3(5), p. 505, 2013.

[117] H. Ferdinando, T. Seppänen, and E. Alasaarela, “Comparing features from ECG pattern and HRV analysis for emotion recognition system,” *In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE*, vol. 31(2), pp. 1–6, 2016.

[118] Y. Zhu, S. Wang, and Q. Ji, “Emotion recognition from users’ EEG signals with the help of stimulus videos,” *In Multimedia and Expo (ICME), 2014 IEEE International Conference on, IEEE*, pp. 1–6, 2014.

[119] V. Gonuguntla, G. Shafiq, Y. Wang, and K. Veluvolu, “EEG classification of emotions using emotion-specific brain functional network,” *In Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*, pp. 2896–2899, 2015.

[120] A. Delorme and S. Makeig, “EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis,” *Journal of Neuroscience Methods*, vol. 134(1), pp. 9–21, 2004.

[121] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, “Learning representations from EEG with deep recurrent-convolutional neural networks,” *arXiv preprint arXiv:1511.06448*, 2015.

[122] J. Hunter, “Matplotlib: A 2D graphics environment,” *Computing in science & engineering*, vol. 9(3), pp. 90–95, 2007.

[123] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. Berg, “ImageNet large scale visual recognition challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115(3), p. 211–252, 2015.

[124] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: an astounding baseline for recognition,” *In Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

[125] S. Siddharth, A. Rangesh, E. Ohn-Bar, and M. Trivedi, “Driver hand localization and grasp analysis: A vision-based real-time approach,” *In Intelligent Transportation Systems (ITSC), 2016 IEEE 19th International Conference on*, pp. 2545–2550, 2016.

[126] M. Orini, R. Bailón, R. Enk, S. Koelsch, L. Mainardi, and P. Laguna, "A method for continuously assessing the autonomic response to music-induced emotions through HRV analysis," *Medical & biological engineering & computing*, vol. 48(5), pp. 423–433, 2010.

[127] E. Billauer, "peakdet: Peak detection using MATLAB," *Detect Peaks in a Vector, Billauer, E., Haifa, Israel, accessed July, 20,*, 2012.

[128] K. Umetani, D. Singer, R. McCraty, and M. Atkinson, "Twenty-four hour time domain heart rate variability and heart rate: relations to age and gender over nine decades," *Journal of the American College of Cardiology*, vol. 31(3), pp. 593–601, 1998.

[129] C. Lin, "Frequency-domain features for ECG beat discrimination using grey relational analysis-based classifier," *Computers & Mathematics with Applications*, vol. 55(4), pp. 680–690, 2008.

[130] S. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," *The Journal of the Acoustical Society of America*, vol. 119(1), pp. 360–371, 2006.

[131] K. Mera and T. Ichimura, "Emotion analyzing method using physiological state," *In Knowledge-Based Intelligent Information and Engineering Systems*, pp. 195–201, 2004.

[132] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," *British Machine Vision Conference*, 2015.

[133] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9(8), pp. 1735–1780, 1997.

[134] W. Zheng, J. Zhu, and B. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE Transactions on Affective Computing*, 2017.

[135] L. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of machine learning research*, vol. 9(Nov), pp. 2579–2605, 2008.

[136] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27(1), pp. 38–49, 2017.

[137] E. Rolls, J. Hornak, D. Wade, and J. McGrath, "Emotion-related learning in patients with social and emotional changes associated with frontal lobe damage," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 57(12), pp. 1518–1524, 1994.

[138] K. Phan, T. Wager, S. Taylor, and I. Liberzon, "Functional neuroanatomy of emotion: A meta-analysis of emotion activation studies in PET and fMRI," *Neuroimage*, vol. 16(2), pp. 331–348, 2002.

[139] N. Deo and M. M. Trivedi, "Looking at the driver/rider in autonomous vehicles to predict take-over readiness," *IEEE Transactions on Intelligent Vehicles*, 2019.

[140] Z. Guo, Y. Pan, G. Zhao, S. Cao, and J. Zhang, "Detection of driver vigilance level using EEG signals and driving contexts," *IEEE Transactions on Reliability*, vol. 67(1), pp. 370–380, 2017.

[141] R. Chai, G. Naik, T. Nguyen, S. Ling, Y. Tran, A. Craig, and H. Nguyen, "Driver fatigue classification with independent component by entropy rate bound minimization analysis in an EEG-based system," *IEEE journal of biomedical and health informatics*, vol. 21(3), pp. 715–724, 2016.

[142] R. Chai, S. Ling, P. San, G. Naik, T. Nguyen, Y. Tran, A. Craig, and H. Nguyen, "Improving EEG-based driver fatigue classification using sparse-deep belief networks," *Frontiers in neuroscience*, vol. 11, p. 103, 2017.

[143] R. Dishman, Y. Nakamura, M. Garcia, R. Thompson, A. Dunn, and S. Blair, "Heart rate variability, trait anxiety, and perceived stress among physically fit men and women," *International Journal of Psychophysiology*, vol. 37(2), pp. 121–133, 2000.

[144] A. Kuefler, J. Morton, T. Wheeler, and M. Kochenderfer, "Imitating driver behavior with generative adversarial networks," *In Intelligent Vehicles Symposium (IV)*, pp. 204–211, 2017.

[145] K. Dwivedi, K. Biswaranjan, and A. Sethi, "Drowsy driver detection using representation learning," *In IEEE International Advance Computing Conference (IACC)*, pp. 995–999, 2014.

[146] Y. Dong, Z. Hu, K. Uchimura, and N. Murayama, "Driver inattention monitoring system for intelligent vehicles: A review," *IEEE transactions on intelligent transportation systems*, vol. 12(2), pp. 596–614, 2010.

[147] A. Doshi and M. Trivedi, "Tactical driver behavior prediction and intent inference: A review," *In Intelligent Transportation Systems (ITSC), 14th International IEEE Conference on*, pp. 1892–1897, 2011.

[148] E. Ohn-Bar and M. Trivedi, "Looking at humans in the age of self-driving and highly automated vehicles," *IEEE Transactions on Intelligent Vehicles*, vol. 1(1), pp. 90–104, 2016.

[149] S. Martin, S. Vora, K. Yuen, and M. Trivedi, "Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 3(2), pp. 141–150, 2018.

[150] C. Papadelis, C. Kourtidou-Papadeli, P. Bamidis, I. Chouvarda, D. Koufogiannis, E. Bekiaris, and N. Maglaveras, "Indicators of sleepiness in an ambulatory EEG study of night driving," *In Engineering in Medicine and Biology Society, 28th Annual International Conference of the IEEE*, pp. 6201–6204, 2006.

[151] S. Lal, A. Craig, P. Boord, L. Kirkup, and H. Nguyen, “Development of an algorithm for an EEG-based driver fatigue countermeasure,” *Journal of safety Research*, vol. 34(3), pp. 321–328, 2003.

[152] X. Ma, Z. Yao, Y. Wang, W. Pei, and H. Chen, “Combining brain-computer interface and eye tracking for high-speed text entry in virtual reality,” *In 23rd International Conference on Intelligent User Interfaces*, pp. 263–267, 2018.

[153] H. Liu, T. Taniguchi, Y. Tanaka, K. Takenaka, and T. Bando, “Visualization of driving behavior based on hidden feature extraction by using deep learning,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18(9), pp. 2477–2489, 2017.

[154] B. Huval, T. Wang, S. Tandon, J. Kiske, W. Song, J. Pazhayampallil, M. Andriluka, P. Rajpurkar, T. Migimatsu, R. Cheng-Yue, and F. Mujica, “An empirical evaluation of deep learning on highway driving,” *arXiv preprint arXiv:1504.01716*, 2015.

[155] S. Ramos, S. Gehrig, P. Pinggera, U. Franke, and C. Rother, “Detecting unexpected obstacles for self-driving cars: Fusing deep learning and geometric modeling,” *In Intelligent Vehicles Symposium (IV)*, pp. 1025–1032, 2017.

[156] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The KITTI dataset,” *The International Journal of Robotics Research*, vol. 32(11), pp. 1231–1237, 2013.

[157] S. Siddharth and M. Trivedi, “Attention monitoring and hazard assessment with biosensing and vision: Empirical analysis utilizing CNNs on the KITTI dataset,” *In 2019 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1673–1678, 2019.

[158] H. Kolkhorst, M. Tangermann, and W. Burgard, “Decoding perceived hazardousness from user’s brain states to shape human-robot interaction,” *In Proceedings of the Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, pp. 349–350, 2017.

[159] G. Carneiro, J. Nascimento, and A. Bradley, “Unregistered multiview mammogram analysis with pre-trained deep learning models,” *In International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 652–660, 2015.

[160] M. Schwarz, H. Schulz, and S. Behnke, “RGB-D object recognition and pose estimation based on pre-trained convolutional neural network features,” *In Robotics and Automation (ICRA), IEEE International Conference on*, pp. 1329–1335, 2015.

[161] X. Ding, Y. Zhang, J. Liu, W. Dai, and H. Tsang, “Continuous cuffless blood pressure estimation using pulse transit time and photoplethysmogram intensity ratio,” *IEEE Transactions on Biomedical Engineering*, vol. 63(5), pp. 964–972, 2016.

[162] Y. Djawad, A. Mu’nisa, P. Rusung, A. Kurniawan, I. Idris, and M. Taiyeb, “Essential feature extraction of photoplethysmography signal of men and women in their 20s,” *Engineering Journal*, vol. 21(4), pp. 259–272, 2017.

[163] D. Zhu, J. Bieger, G. Molina, and R. Aarts, “A survey of stimulation methods used in SSVEP-based BCIs,” *Computational intelligence and neuroscience*, 2010.

[164] Q. Yuan, W. Zhou, S. Li, and D. Cai, “Epileptic EEG classification based on extreme learning machine and nonlinear features,” *Epilepsy research*, vol. 96(1-2), pp. 29–38, 2011.

[165] K. Cheng, Y. Chen, and T. Wang, “Physiological parameters assessment for emotion recognition,” *In Biomedical engineering and sciences (IECBES), IEEE EMBS conference on*, vol. 250, pp. 995–998, 2012.