Gender Differences in Facial Emotion Perception for User Profiling via Implicit Behavioral Signals

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Electronics and Communication Engineering
by Research

by

Maneesh Bilalpur
201532589
maneesh.bilalpur@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
JULY 2018

Copyright © MANEESH BILALPUR, 2018
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled " Gender Differences in Facial Emotion Perception for User Profiling via Implicit Behavioral Signals" by MANEESH BILALPUR, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Prof. RAMANATHAN SUBRAMANIAN

To

My Family

Acknowledgments

The research time at IIIT-H was both exciting and occasionally uncertain. I would like to take this opportunity to thank several people who have been there in times of need.

I would like to begin by thanking my advisor, Prof. Ramanathan Subramanian, who has been there in understanding and guiding me not just in terms of the thesis but much more. This work would not be complete without thanking my co-authors, Dr. Seyed Mostafa Kia and Manisha Chawla for their expertise in Neuroscience and experience with EEG have helped me tremendously. I would also take this opportunity to thank Prof. C.V. Jawahar for being understanding and encouraging me in joining the CVIT. I extend my heartfelt thanks to Prof. Mohan Kankanhalli without whom this work might not have assumed this shape, and for his wise words regarding a career in research and the lab environment at SeSaME.

Not leaving behind, I would like to thank the folks I have spent my time with at both IIIT-H and National University of Singapore. I would like to thank Viral Parekh for imparting his knowledge gained through experience at innumerable times and Jithendra Yasaswi for allowing me to share his desk for a short while. I would also like to thank my friends Rhishi Pratap Singh, Nachiket Ayir, Upender Reddy, Prakash and Mahesh Murthy for making my stay at IIIT memorable. I thank Abhinav Ramesh Kashyap, Rahul Soni and Mannu Malhotra for a great time in Singapore.

I would like to conclude by thanking my parents who have always supported my decisions, and the mystic forces in the universe that led me to what I am today.

Abstract

Understanding human emotions has been of research interests to multiple domains of modern day science namely Neuroscience, Psychology and Computer Science. The ultimate goals of each of these domains in studying them might be different such as neuroscientists interest in emotions is primarily to understand the structural and functional abilities of brain, psychologists study them to understand human interactions and computer scientists to design interfaces and automation of certain human-centric tasks.

Several earlier works have suggested the existence of two facets to emotions namely perception and expression. It has been advised to study emotions in the aspects of perception and expression as separate entities. This work attempts to study the existence of gender differences in emotion perception(in specfic the Ekman emotions). Our work aims at utilizing such differences for user profiling, particularly in terms of gender and emotion Recognition. We employed implicit signals–the non-invasive electrical scalp activity of brain through Electroencephalography(EEG) and gaze patterns acquired through low-cost commercial devices to achieve these. We studied the impact of facial emotion intensity and facial regions in invoking the differences through stimuli involving of different intensities and masking face regions which were deemed important in previous studies.

We expressly examined the implicit signals for their ecological validity. Existence of correlations between our study and previous studies from the above said domains in terms of Event Related Potentials(ERPs) and fixation distributions have added uniqueness and strength to our work. We achieved a reliable gender and emotion recognition with Support Vector Machine based classifiers and further designed a deep learning model to significantly outperform them. We also analyzed for emotion specific time windows and key electrodes for maximum gender recognition to arrive at some interesting conclusions.

The appendix chapter on cross-visualization based cognitive workload classification using EEG attempts to quantify workload in order to evaluate user-interfaces. We employ four common yet unique data visualization methods to induce varying levels of workload through a standard n-back task and attempt to classify it across visualizations with deep learning through transfer learning. We compare its performance against the Proximal Support Vector Machines adopted in earlier works for within visualization workload classification.

Contents

Chapter Page

1 Introduction . . . 1
1.1 Key contributions of the work . . . 3
1.2 Organization of the work . . . 3

2 Related works . . . 4
2.1 User-centered ER . . . 4
2.2 Gender Differences in Emotion Recognition . . . 5
2.3 Analysis of Related Work . . . 6

3 Experimental Design and User Responses . . . 7
3.1 Materials and Methods . . . 7
3.2 User Responses . . . 8
3.2.1 Response time (RT) . . . 9
3.2.2 Recognition Rates . . . 10

4 Analyzing Implicit Signals . . . 11
4.1 EEG Analysis . . . 11
4.1.1 EEG preprocessing . . . 11
4.1.2 ERP analysis . . . 11
4.2 Eye-tracking analysis . . . 13
4.2.1 Fixation analysis . . . 14

5 User Profiling with Implicit Signals . . . 15
5.1 Experimental settings . . . 15
5.2 Impact of Emotional Intensity . . . 16
5.2.1 Valence Recognition . . . 16
5.2.2 Gender Recognition . . . 17
5.2.3 Spatio-temporal EEG analysis . . . 17
5.3 Impact of facial regions . . . 19
5.3.1 Valence Recognition . . . 19
5.3.2 Gender Recognition . . . 20
5.3.3 Spatio-temporal EEG analysis . . . 21
5.4 Deep learning and AdaBoost . . . 22
5.4.1 Deep learning Architecture . . . 22
5.4.1.1 Training Strategy . . . 24

5.4.2 AdaBoost . . . 24
5.4.3 Emotion Recognition . . . 25
5.4.4 Gender Recognition . . . 26
5.4.5 Spatio-temporal EEG analysis . . . 27

6 Conclusions . . . 29

Appendix A: EEG-based Cognitive Load Estimation Across Visualizations . . . 33
A.1 Related Work . . . 34
A.1.1 Implicit signals for workload evaluation . . . 34
A.1.2 Analysis of related work . . . 35
A.2 Materials and Methods . . . 35
A.2.1 Experimental Design . . . 36
A.2.1.1 Stimuli . . . 36
A.2.1.2 Protocol . . . 37
A.2.1.3 Hypotheses . . . 38
A.3 User Response Analysis . . . 38
A.3.1 Response times . . . 38
A.3.2 Response accuracies . . . 39
A.3.3 Explicit responses . . . 39
A.4 EEG analysis for Cognitive load . . . 39
A.4.1 Preprocessing . . . 39
A.4.2 Deep learning for Cognitive Load Estimation . . . 40
A.4.3 pSVM based Cognitive Load Estimation . . . 40
A.4.4 Experiments and Results . . . 41
A.4.4.1 Classification settings . . . 41
A.4.4.2 Comparison of results . . . 44
A.5 Conclusions and Future works . . . 44

Bibliography . . . 47

List of Figures

Figure Page

1.1 Emotiv electrode configuration. 2

3.1 Experimental Protocol: Viewers were required to recognize the facial emotion from either an *unmasked* face, or with the *eye/mouth* masked. Trial timelines for the two conditions are shown at the bottom. 8
3.2 Emotion-wise RTs of females and males considering various morphs. y-axis denotes response time in seconds. Error bars denote unit standard error (best-viewed in color). . 9
3.3 Emotion wise RRs for various morph conditions in females and males. 10

4.1 ERPs for HI morphs (a–d) and LI morphs (e–h): (from left to right) O1 and O2 ERPs for females and males.y-axis shows ERP in microvolts, refer (h) for legend. ERP data is plotted upside down (view in color and under zoom). 12
4.2 Female ERPs in the *HI* (a) and *eye mask* (b) conditions. Best-viewed in color and under zoom. 13
4.3 (a–d) Fixation duration distributions for males and females in the *HI*, *LI*, *eye-masked* and *mouth-masked* conditions. Best-viewed in color and under zoom. 13
4.4 Fixation duration distributions: Region-wise fixation distribution across different emotions for (a) females and (b) males for both HI and LI conditions combined. Error bars denote unit standard error. 14

5.1 ER Results: Valence recognition results with HI and LI morphs. 16
5.2 (a,b) Spatio-temporal EEG analyses: GR results for HI and LI morphs over temporal windows (W1–W4). (c) Top 3 electrodes based on AUC metric for HI and LI morphs. x-axis denotes the rank of the electrodes in decreasing order. Error bars denote unit standard error. 19
5.3 ER performance for *HI*, *eyemask* and *mouthmask* epochs for males and females. 19
5.4 (a–c): Temporal variance in GR performance for the *no mask*, *eye mask* and *mouth mask* conditions. (d) Top three electrodes showing best GR performance for the three conditions. 22
5.5 Deep learning architecture showing various layers in the model and parameters. 23
5.6 ER Results: valence recognition results with HI and LI morphs. 26
5.7 Temporal EEG analyses: GR results for HI, LI, Eyemask and Mouthmask morphs respectively over temporal windows. 28

A.1 Various visualizations used in the problem *character*, *position*, *bar* and *pie*. 33
A.2 Sample protocol settings for 1-back task. 36

A.3 (left) Response times(RTs in seconds) and (right) response accuracy(RAs) for various stimuli. 38
A.4 Deep learning architecture. 39

List of Tables

Table Page

3.1 ANOVA summary for behavioral measures. df and Sig respectively denote degrees of freedom and significance level. 9

5.1 Table showing Gender Recognition results over different modalities and their fusion for HI and LI morphs using NB-NaiveBayes,RSVM-SVM with RBF kernel and LSVM-SVM with linear kernel. 18
5.2 GR with different conditions/modalities. The best classifier is denoted within parenthesis. Highest AUC for a given condition is shown in bold. 21
5.3 Parameters for the deep learning model 23
5.4 Gender Recognition results over different modalities and their fusion for HI, LI, Eyemask and Mouthmask morphs using CNN(EEG) and AdaBoost(Gaze). 27
5.5 Spatial performance to evaluate the importance of channels in decreasing order. The parentheses contain the channel for which the score was achieved. 27

A.1 ANOVA summary for RTs and RAs. df and Sig respectively denote degrees of freedom and significance level. 36
A.2 Summary of the features and outputs for CNN and pSVM methods. 40
A.3 Data distribution split with respect to the back and the visualization type. 41
A.4 Deep CNN based cognitive load estimation. Rows represent the training visualization while columns represent the test. 42
A.5 pSVM based cognitive load estimation. Rows represent the training visualization while columns represent the test 43

Chapter 1

Introduction

The birth of Brain-Computer Interfaces was initially aimed at providing accessibility for people with disabilities. It has achieved an appreciable success along those lines. However, its applications are not restricted to medical domain, the current trend of low-cost devices with advancements in amplifiers and reliable sensor technology along with the increased understanding of the human brain has promoted the development of BCIs for commercial applications like cognitive workload estimation, robotic control and user profiling.

BCIs today can be broadly divided into three major types [58]:

- **Invasive BCIs**: These involve surgical implants or probes for recording from or stimulating neurons inside the brain.
- **Semi-invasive BCIs**: These involve recording from or stimulating the brain surface or nerves.
- **Non-invasive BCIs**: As the name suggests, these are non-invasive and user friendly. They record the brain activity without organ penetration in human body.

While several brain imaging methods like functional Magnetic Resonance Imaging(fMRI), Magnetoencephelography and Near IR spectroscopy(NIRS) exist, the importance and advantages of Electroencephalography(EEG) outweigh those of the rest. EEG provides a superior temporal resolution and does not require heavy machinery(superior portability) in order to image the brain [22]. Imaging method through EEG could be both invasive or non-invasive catering to wide variety of applications like functional and structural understanding of the brain and commercial interfaces.

The current work lies at the intersection of Neuroscience, Multimedia and Machine learning. We utilize the backgrounds and established ideas from these varied domains to achieve a reliable and state-of-the-art user profiling in terms of Gender and Emotion Recognition. We summarize the need for such efforts in sensitive applications like user profiling and explain how our work is unique below:

Gender human-computer interaction (HCI) [82] and Affective HCI [55] have evolved as critical HCI sub-fields due to widespread acknowledgment of the fact that computers need to *appreciate* and *adapt to* the user's gender and emotional state. The ability to identify user demographics including gender and

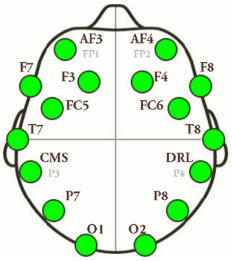

Figure 1.1 Emotiv electrode configuration.

emotion can benefit interactive and gaming systems in terms of (a) visual and interface design [53, 16], (b) game and product recommendation (via ads) [85, 29], and (c) provision of appropriate motivation and feedback for optimizing user experience [62]. Contemporary gender recognition (GR) and emotion recognition (ER) systems primarily work with facial [51, 32] or speech [43, 42] cues; however, face and speech are *biometrics* that encode an individual's identity, and pose grave privacy concerns as they can be recorded without the user's knowledge [2].

This work examines GR and ER from ***implicit user behavioral signals***, in the form of *EEG brain signals* and *eye movements*. The conveyance of implicit behavioral signals is hidden from the outside world, and they cannot be recorded without express user cooperation making them privacy compliant [12]. Also, behavioral signals such as EEG and eye movements are primarily *anonymous* as little is known regarding their uniqueness to a person's identity [84].

Specifically, we attempt GR and ER using signals captured by commercial, off-the-shelf devices which are minimally intrusive, affordable, and popularly used in gaming as input or feedback modalities [72, 45, 73]. The *Emotiv* EEG wireless headset consists of 14 dry (plus two reference) electrodes having a configuration as shown in Fig. 1.1. While being lightweight, wearable and easy-to-use, neuroanalysis with Emotiv can be challenging due to relatively poor signal quality. Likewise, the *EyeTribe* is a low-cost eye-tracker which its suitability for research purposes has been evaluated and endorsed [17]. We show how relevant gender and emotion-specific information is captured by these low-cost devices via examination of event-related potential (ERP) and fixation distribution patterns, and also through recognition experiments.

To capture gender-based differences, we designed a facial emotion recognition experiment as males and females have been known to respond differently to affective information [62, 49, 23, 10]. Our study performed with 28 viewers (14 males and 14 females) confirms that women are superior at facial ER, mirroring prior findings. Hypothesizing that enhanced female emotional sensitivity should reflect via their implicit responses, we examined EEG signals and found that (1) Stronger ERPs are observed for females while processing negative facial emotions, (2) Better ER is achieved with female EEG responses and eye movement features, and (3) Emotion-specific gender differences manifest better than emotion-agnostic differences.

Overall, our work makes the following research contributions: (a) While prior works have identified gender differences in emotional behavior, this is one of the first works to expressly perform GR

and ER from implicit behavioral signals; (b) Apart from recognition experiments, we show that the employed devices capture meaningful information, as typified by gender and emotion-specific event related potentials (ERPs); (c) The use of minimally intrusive, off-the-shelf and low-cost devices extends the ecological validity of our findings, and the feasibility of employing a similar framework for large-scale interactive user profiling.

1.1 Key contributions of the work

We summarize the key contributions of the work as follows:

- Exploring the utility of gender differences in emotion perception for user profiling acquired using low-cost devices.
- Proving the qualitative ecological validity of the acquired signals through established cognitive markers.
- Corroborating the existence of cognitive markers with machine learning methods.
- Use of deep learning methods to overcome the manual interference in preprocessing the EEG signals and boost the performance.

1.2 Organization of the work

The outline of the thesis is summarized below:

- *Chapter 2* provides a succinct understanding of the related works on similar lines in the fields of Neuroscience and imaging methods for emotion understanding and interpretation.
- *Chapter 3* elaborates on the experimental design and setup utilized and studies the explicit user responses.
- *Chapter 4* examines the implicit signals for the presence of differences in emotional perception and validates our hypothesis.
- *Chapter 5* analyzes the implicit signals for their utility in user profiling using machine learning methods.
- *Chapter 6* concludes highlighting the takeaways supporting our hypotheses.
- *Appendix A* elaborates on the contributions to the work on cognitive workload estimation using EEG.

Chapter 2

Related works

Many works have focused on ER with implicit behavioral signals [86, 47, 3, 37, 69], but very few works have examined on GR with such signals [83]. Also, some works have attempted to capture emotion and gender-specific differences by examining eye movement and EEG responses to emotional faces [61, 88, 33, 68]. However, none of these studies have attempted to isolate behavioral differences as well as exploit them in a computational setting. We explicitly show gender-specific EEG and eye movement patterns characteristic of emotional face processing under occluded and unoccluded conditions, and demonstrate recognition results consistent with these findings. Spatio-temporal analysis of EEG data also reveals cognitive differences while processing facial emotions in the *no mask* and *mask* conditions.

Tavakoli *et al* [56] perform valence (+ve vs -ve emotion) recognition using eye movements with an emphasis on evaluating eye-based features and their fusion, and achieved 52.5% accuracy with discriminative feature selection and a linear SVM.

To position our work with respect to the literature, this section reviews related work on (a) user-centered ER, and (b) gender differences in emotional face processing.

2.1 User-centered ER

Emotions evoked by multimedia content have been modeled via *content-centered* as well as *user-centered* approaches. Content-centered approaches typically attempt to find emotional correlates from content-based features [25, 79, 77, 64], while user-centered methods monitor user behavioral cues (eye movements, EEG signals, *etc.*) to recognize the evoked emotion. As emotion represents a subjective human feeling, many user-centered approaches have attempted to model emotions by examining *explicit* and *implicit* user behavioral cues. Conspicuous facial cues are studied in [32] to detect multimedia highlights, while implicit physiological measurements are employed to model emotions induced by music and movie scenes respectively in [37] and [3]. EEG and eye movements are two popular implicit modalities employed for ER, and many works have predicted affect with a combination of both [86, 47, 10], or exclusively using either signal [44, 87, 56, 69, 3, 33, 68].

Valence (positive vs negative emotion) recognition using eye-based features is proposed in [56]. ER with EEG and pupillary responses acquired from five users is discussed in [86]. A deep unsupervised method for ER from raw EEG data is proposed in [44], and its effectiveness is shown to be comparable to manually extracted features. Differential entropy (DE) features from EEG data are extracted to train an integrated deep belief network plus hidden Markov model for ER in [87]. A DAE (differential auto-encoder) that learns shared representations from EEG and eye-based features is proposed for valence recognition in [47]. Almost all of these works employ lab-grade eye-trackers and EEG sensors which are bulky and intrusive, and therefore preclude naturalistic user behavior.

2.2 Gender Differences in Emotion Recognition

As facial emotions denote important non-verbal communication cues during social interactions, many social psychology studies have examined human recognition of emotional faces. Certain facial features are found to encode emotional cues better than others. When viewing a human face, most attention is drawn to the eyes, nose and mouth regions [78, 67]. [70] observed that visual attention is localized around the eyes for mildly emotive faces, but the nose and mouth regions also attract substantial eye fixations in highly emotive faces. A recent eye tracking study [61] notes that distinct eye fixation patterns emerge for different facial emotions. The mouth is most informative for the *joy* and *disgust* emotions, whereas eyes mainly encode information relating to *sadness*, *fear*, *anger* and *shame*. In a similar study [5], more fixations are noted on the upper face half for *anger* as compared to *disgust*, while no differences are observed on lower face half for the two emotions. However, humans can find it difficult to distinguish between some facial emotions that tend to have similar characteristics– examples are the high overlap rate between the *fear–surprise* and *anger–disgust* emotion pairs [67, 19].

Many studies have also identified gender differences during facial emotion processing. Females are generally found to be better at ER than males, irrespective of age [71]. Other studies examining the role of facial movements in ER [8, 49, 23] also note that females recognize facial emotions more accurately than males, even when only partial information is available. Some evidence also points to the fact that females are faster at ER than males [24, 57].

Differences in gaze patterns and neural activations have been found between males and females while viewing emotional faces. Women's tendency to fixate on the eyes positively correlates with their ER capabilities, while men tend to look at the mouth for emotional cues [71, 24]. Likewise, there exists evidence in terms of EEG ERPs that negative facial emotions are processed differently and rapidly by women, and do not necessarily entail selective attention towards emotional cues [46, 88].

An exhaustive review of GR methodologies is presented in [83], and the authors evaluate GR methods across different communities using metrics like *universality, distinctiveness, permanence* and *collectability*. While crediting bio-signals like EEG and ECG for their accuracy and trustworthiness, this work also highlights the invasiveness of bio-sensors. The sensors used in this work are nevertheless minimally intrusive and non-invasive, thereby ensuring naturalistic user experience, while also record-

ing meaningful emotion and gender-related information. Among the few works to attempt GR from user-centric cues, EEG and speech features are proposed for age and gender recognition in [52].

2.3 Analysis of Related Work

A close examination of related work reveals that (1) Many have attempted ER from user-centered cues, both conspicuous and implicit, and a number of works have also tried to isolate gender-specific differences based on gaze patterns and neural activations; nevertheless, very few works have that expressly attempted GR based on implicit user cues. Differently, this work employs implicit signals for GR and ER with a number of classifiers and reliable detection of both gender and valence (AUC score of $>$0.6) is achieved via our proposed methodology; (2) Another salient aspect of our approach is the use of low cost, off-the-shelf and easy-to-use sensors for acquiring implicit signals, which suffer from poor signal quality while enabling natural user behavior. We show how these sensors nevertheless capture meaningful information via the analysis of ERPs and fixation distribution patterns; (3) In contrast to most prior works have either analyzed (a) explicit user responses in terms of reaction times and recognition rates, or (b) implicit behavioral patterns to discover gender differences in ER, or (c) employed EEG and eye-based features for GR and ER without validation of the corresponding signals, our work touches upon all of these aspects. We show how explicit behavioral gender differences noted for negative emotions also reflect via EEG ERPs and fixation distributions, and GR performance. The multiple similarities among explicit and implicit behavioral patterns, along with the use of low-cost sensors validate the findings in this work and extend its ecological validity.

Chapter 3

Experimental Design and User Responses

3.1 Materials and Methods

We examined gender differences in visual emotional face processing when (a) the faces were fully visible, and (b) some facial features were occluded. Specifically, we investigated gender sensitivity to *high* and *low* intensity emotional faces and *high* intensity faces when the *eye* and *mouth* regions are occluded (see Fig.3.1), as the importance of these features towards conveying facial emotions has been highlighted by prior works [70, 61]. The above conditions are denoted as the ***HI***,***LI***, ***eye mask*** and ***mouth mask*** conditions throughout the paper.

Participants: 28 subjects of different nationalities (14 male, age 26.1 ± 7.3 and 14 female, age 25.5 ± 6), with normal or corrected vision, took part in our study.

Stimuli: We used emotional faces of 24 models (12 male, 12 female) from the Radboud Faces Database (RaFD) [41]. RaFD includes facial emotions of 49 models rated for *clarity*, *genuineness* and *intensity*, and the 24 models were chosen such that their Ekman facial emotions[1] were roughly matched based on these ratings. We then *morphed* the emotive faces from *neutral* (0% intensity) to *maximum* (100% intensity) to generate intermediate morphs in steps of 5%. Emotional morphs with 55–100% intensity were used as HI and 25–50% were used as LI morphs in this work. *Eye* and *mouth-masked* faces were generated upon automatically locating facial landmarks via [6] over the HI morphs. The *eye mask* covered the eyes and nasion, while the *mouth mask* covered the mouth and the nose ridge. The stimuli were 361×451 pixels in size, encompassing a visual angle of 9.1° and 11.4° about x and y at 60 cm screen distance.

Protocol: The protocol is outlined in Fig.3.1, and involved the presentation of *unmasked* and *masked* faces to viewers over two separate sessions, with a break in-between to avoid fatigue. We chose one face per model and emotion, resulting in 144 face images (1 morph/emotion $\times$ 6 emotions $\times$ 24 models). In

[1] **A**nger, **D**isgust, **F**ear, **H**appy, **S**ad and **Su**rprise.

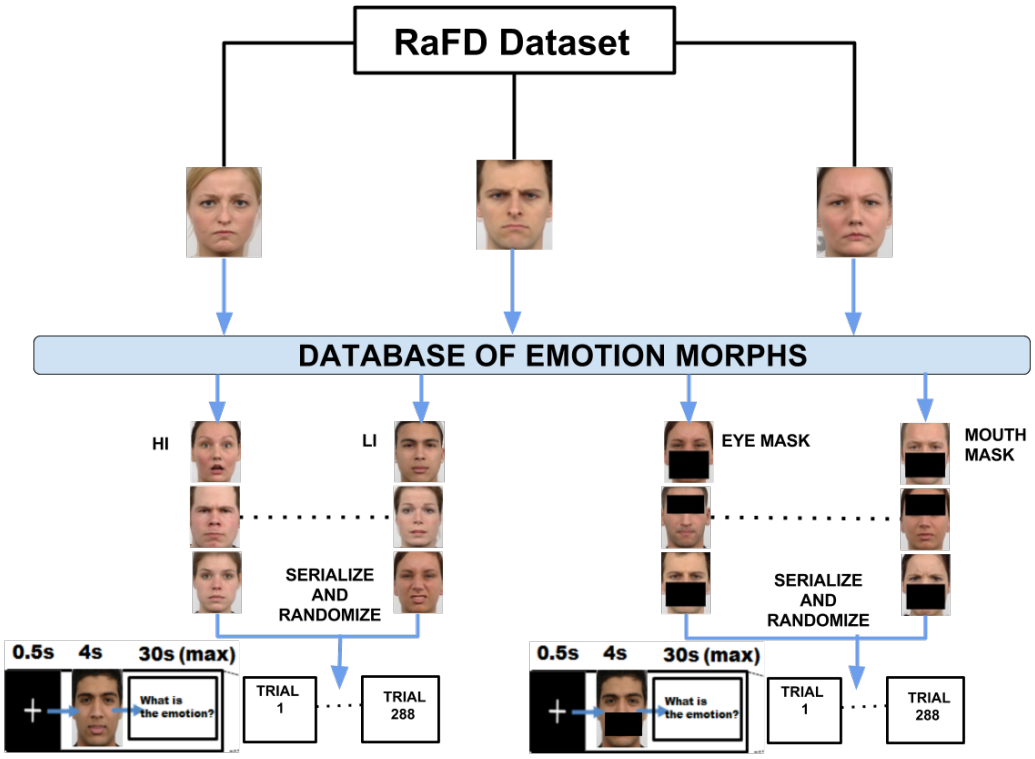

Figure 3.1 Experimental Protocol: Viewers were required to recognize the facial emotion from either an *unmasked* face, or with the *eye/mouth* masked. Trial timelines for the two conditions are shown at the bottom.

the first session, these faces were shown as is in random order (HI/LI condition), and were re-presented with an eye or mouth mask (eye/mouth mask conditions) in the second session. We ensured a 50% split of the HI and LI morphs in first session and eye/mouth-masked faces in the next session.

During each trial, the stimulus was displayed for 4s preceded by a fixation cross for 500 ms. The viewer then had a maximum of 30s to make one out of seven choices concerning the facial emotion (6 emotions plus neutral) via a radio button. Neutral faces were only utilized for morphing purposes and not used in the experiment. During the experiment, viewers' EEG signals were acquired via the 14-channel *Emotiv epoc* device, and eye movements were recorded with the *Eyetribe* tracker. The experiment was split into 4 segments to minimize acquisition errors, and took about 90 minutes.

3.2 User Responses

We first compare male and female sensitivity to emotions based on response times (RTs) and recognition rates (RRs), and then proceed to examine their implicit eye movements and EEG responses. RT and RR denote performance measures computed from *explicit* viewer responses.

Table 3.1 ANOVA summary for behavioral measures. df and Sig respectively denote degrees of freedom and significance level.

Response	Predictor	Morph Intensity	Emotion	Gender	MI*Emotion	MI*Gender	Error	Total
RT	df	1	5	1	5	5	32	160
	F	5.179	7.699	3.411	2.901	3.327		
	Sig	$p < 0.05$	$p < 0.001$	$p = 0.074$	$p < 0.05$	$p < 0.05$		
RR	F	531.1	59.83	7.37		7.359		
	Sig	$p < 0.001$	$p < 0.001$	$p < 0.05$		$p < 0.001$		

Figure 3.2 Emotion-wise RTs of females and males considering various morphs. y-axis denotes response time in seconds. Error bars denote unit standard error (best-viewed in color).

3.2.1 Response time (RT)

Overall RTs for the *HI*, *LI*, *eye mask* and *mouth mask* conditions were respectively found to be 1.44 $\pm$ 0.24, 1.52 $\pm$ 0.05, 1.17 $\pm$ 0.12 and 1.25 $\pm$ 0.09 seconds, implying that facial ER was fairly instantaneous, and that viewer responses were surprisingly faster with masked faces. A fine-grained comparison of male and female RTs in the four conditions (Fig.3.2) revealed that females ($\mu_{\text{RT}} = 1.40 \pm 0.10$s) were generally faster than males ($\mu_{\text{RT}} = 1.60 \pm 0.10$s) at recognizing *HI* emotions. Table 3.1 summarizes results from a 3-way ANOVA on RTs, with ***morph intensity*** (MI) and *facial* ***emotion*** as within-subject factors, and ***gender*** as the between-subject factor. ANOVA revealed the main effect of MI with generally faster RTs noted for HI morphs.The presented facial emotion also impacted RTs, and pairwise comparisons revealed that RT for *happy* (1.3 $\pm$ 0.9 s) was significantly faster than for *anger*, *fear*, *sad* and *surprise*. ANOVA also revealed a marginal gender effect with females (1.4 $\pm$ 0.1 s) generally responding faster than males (1.6 $\pm$ 0.1 s). Significant interaction effects were also noted between (a) morph intensity and emotion, and (b) gender and emotion. *Angry* faces produced the highest RTs for males (mean RT = 1.9 s) while females took longest to judge *surprise* (mean RT = 1.5 s). Conversely, lowest male RTs were noted for *happy* (mean RT = 1.3 s), and lowest female RTs for *disgust* (mean RT = 1.2 s). As evident from Fig. 3.2, females were significantly faster ($F(3, 96) = 3.33, p < 0.05$) at

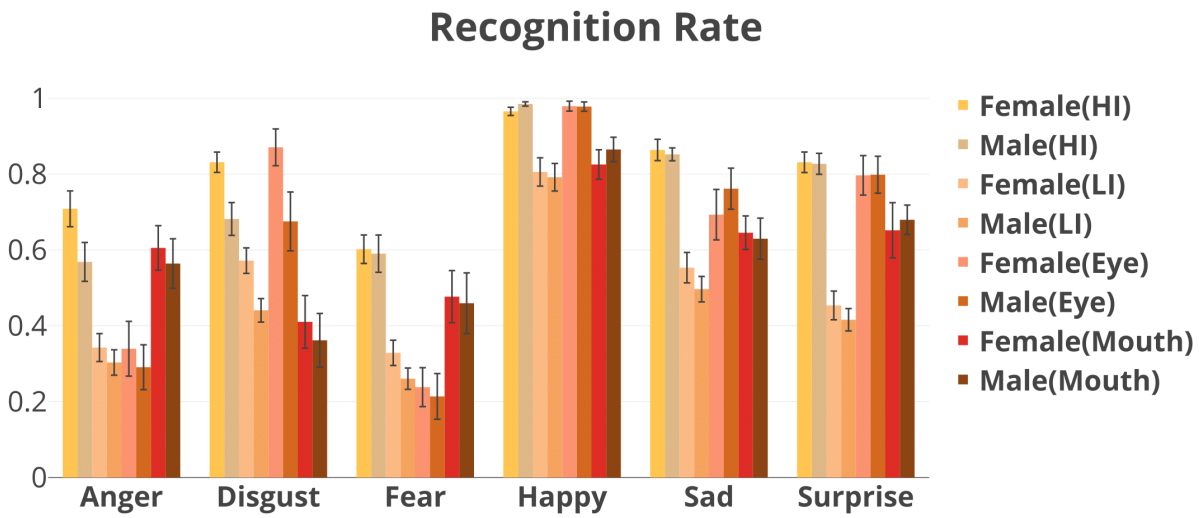

Figure 3.3 Emotion wise RRs for various morph conditions in females and males.

judging negative (*anger*, *disgust*, *fear* and *sad*) emotions. There was no significant change in the RTs for *LI* emotions. Females performed better than males except with a marginal increase in the RT for females(1.45 ± 0.41) and a decrease in males(1.58 ± 0.59). Female alacrity nevertheless decreased for masked faces, with males responding marginally faster for *eye masked* faces (μ_{RT}(male) $= 1.13 \pm 0.11$s vs μ_{RT}(female) $= 1.21 \pm 0.13$s), and both genders responding with similar speed for *mouth masked* faces (μ_{RT}(male) $= 1.24 \pm 0.10$s vs μ_{RT}(female) $= 1.25 \pm 0.09$s).

3.2.2 Recognition Rates

Overall, RRs for *HI* emotions ($\mu_{\text{RR}} = 77.6$) were expectedly higher than for *eye-masked* ($\mu_{\text{RR}} = 59.7$) and mouth-masked ($\mu_{\text{RR}} = 63.5$) emotions. The lowest RR was observed with *LI* emotions($\mu_{\text{RR}} = 49.1$). Happy faces were recognized most accurately in all four conditions. Specifically focusing on gender differences (Fig.3.3), females recognized facial emotions more accurately than males and this was particularly true for negative (A,D,F,S) emotions– male vs female RRs for these emotions differed significantly in the *HI* (μ_{RR}(male) $= 54.3$ vs μ_{RR}(female) $= 61.2, p < 0.05$) and *eye mask* conditions (μ_{RR}(male) $= 51.8$ vs μ_{RR}(female) $= 58.1, p < 0.05$), and marginally for *mouth mask* condition (μ_{RR}(male) $= 52$ vs μ_{RR}(female) $= 55.8, p = 0.08$). Males marginally outperformed females in *LI* condition (μ_{RR}(male) $= 49.8$ vs μ_{RR}(female) $= 47.7$). Overall, (a) HI emotion morphs were recognized way more accurately than LI morphs, (b) females recognized negative emotions better than males and (c) *Happy* was the easiest emotion to recognize.

Chapter 4

Analyzing Implicit Signals

4.1 EEG Analysis

4.1.1 EEG preprocessing

We extracted epochs for each trial (4.5s of stimulus-plus-fixation viewing time @ 128 Hz sampling rate), and the 64 leading pre-stimulus samples were used to remove DC offset. This was followed by (a) EEG band-limiting to within 0.1–45 Hz, (b) Removal of noisy epochs via visual inspection, and (c) Independent Component Analysis (ICA) to remove artifacts relating to eye-blinks, and eye and muscle movements. Muscle movement artifacts in EEG are mainly concentrated in the 40–100 Hz band, and are removed upon band-limiting and via inspection of ICA components. Finally, a 7168 dimensional ($14\times4\times128$) EEG feature vector was generated from concatenation of the 14 EEG channels over 4s of stimulus viewing.

4.1.2 ERP analysis

Having noted that females are quicker and better at recognizing (especially negative) emotions, we hypothesized that these behavioral differences should also manifest via implicit cues in the form of brain responses and eye movements during emotional face processing. This section describes the features extracted and methods employed to examine these cues.

Event Related Potentials (ERPs) are (averaged) time-locked neural responses related to sensory and cognitive events. ERPs occurring in the first 100ms post stimulus presentation are stimulus-related (or exogenous), while later ERPs are cognition-related (or endogenous). We examined the first second of EEG epochs for ERPs characteristic of emotion and gender.

Prior works have observed event-related potential[1](ERP) based gender differences from EEG data recorded with lab-grade sensors [46, 50, 88]. Specifically, [46] notes enhanced negative ERPs for females in response to negative emotional stimuli. However, capturing ERPs with commercial devices

[1]ERP denotes *average* EEG response over multiple trials. P300 (positive) and visual N100, N400 (negative) are examples of ERP components which respectively peak around 300, 100 and 400 ms post stimulus onset.

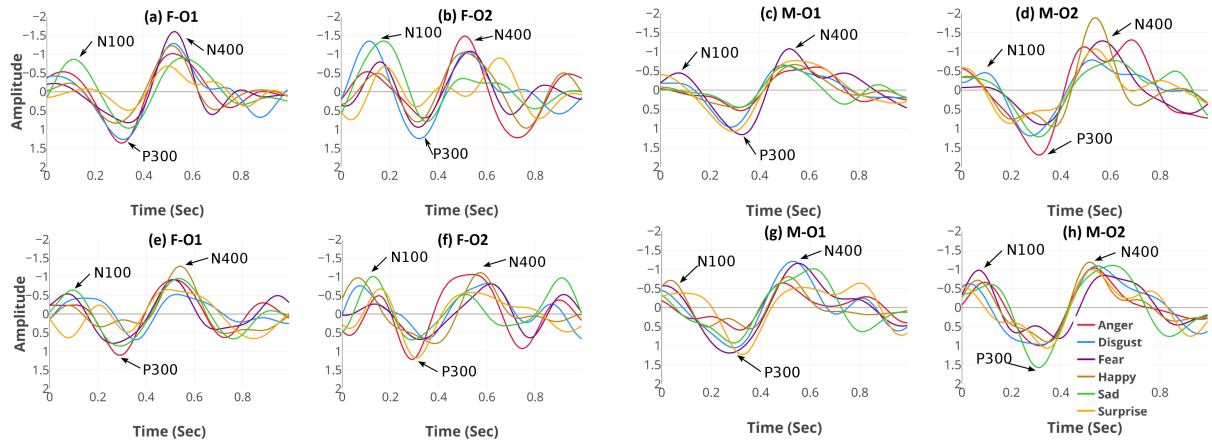

Figure 4.1 ERPs for HI morphs (a–d) and LI morphs (e–h): (from left to right) O1 and O2 ERPs for females and males.y-axis shows ERP in microvolts, refer (h) for legend. ERP data is plotted upside down (view in color and under zoom).

is challenging due to their low signal-to-noise ratio. Fig. 4.1 and 4.2 present the P300, visual N100 and N400 ERP components in the occipital O1 and/or O2 electrodes (see Fig. 1.1) corresponding to various face morphs. Note that the occipital lobe represents the *visual processing center* in the brain, as it contains the primary visual cortex.

Comparing O1/O2 male and female ERPs for positive (H,Su) and negative (A,D,F,Sa) emotions, no significant difference can be observed between male positive and negative ERP peaks for HI or LI faces (see columns 3,4). However, we observe stronger N100 and P300 peaks in the negative female ERPs for both HI and LI faces (columns 1,2). Also, a stronger female N400 peak can be noted for HI faces consistent with prior findings [46]. Lower male N100 and P300 latencies can be observed for positive HI emotions, with the pattern being more obvious at O2. Also, lower male N400 latencies can be noted at O2. The positive vs negative ERP difference for females is narrower for LI faces, which may be attributed to the greater difficulty in identifying LI emotions. This is further corroborated by the fact that LI faces generally produce weaker ERPs at O1 and O2 than HI faces.

In a related work studying gender differences in emotional processing, [46] also observed enhanced N100 and P300 ERP components in females for negative stimuli. Fig.4.2 shows female ERPs[2] observed in the occipital O2 electrode for the *HI* and *eye mask* conditions. Clearly, one can note enhanced N100 and P300 ERP components for negative emotions in the *HI* condition (Fig.4.2(a)). This effect is attenuated in the *eye mask* (Fig.4.2(b)) and *mouth mask* conditions, although one can note stronger N400 amplitudes for anger and disgust in the *eye mask* case. Such a pattern was not observable from male ERPs. Overall, the observed ERPs affirm that gender differences in emotional processing can be captured with the low-cost Emotiv EEG device.

[2]ERP data is plotted upside down as per convention.

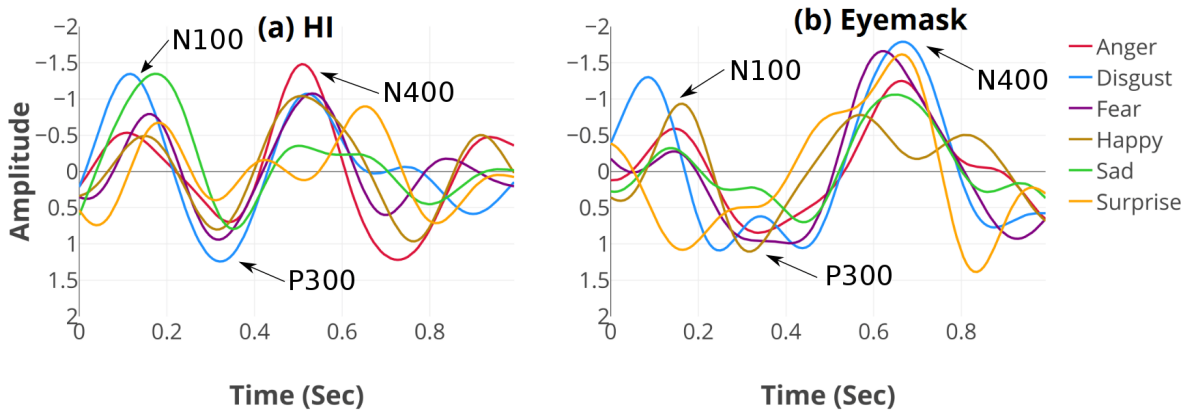

Figure 4.2 Female ERPs in the *HI* (a) and *eye mask* (b) conditions. Best-viewed in color and under zoom.

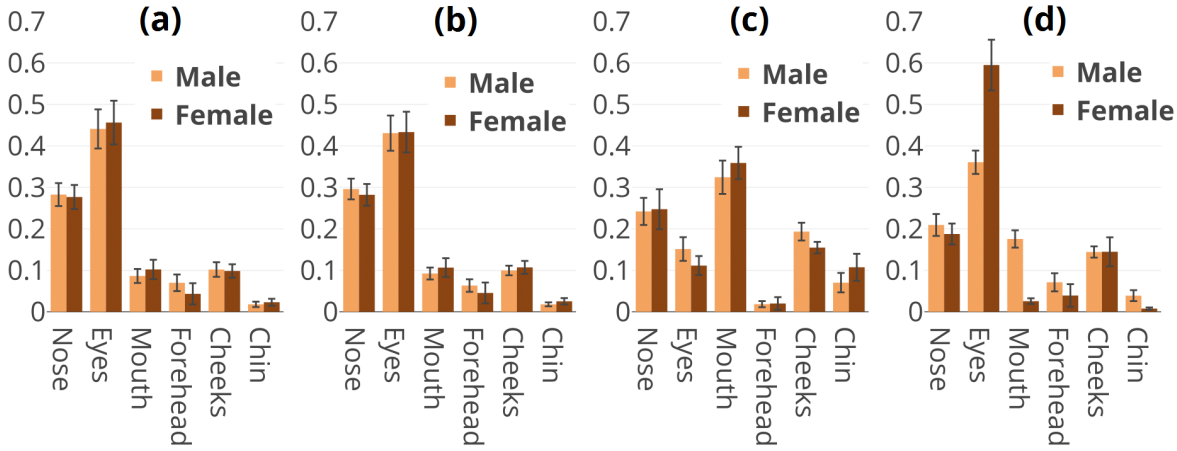

Figure 4.3 (a–d) Fixation duration distributions for males and females in the *HI*, *LI*, *eye-masked* and *mouth-masked* conditions. Best-viewed in color and under zoom.

4.2 Eye-tracking analysis

Gender differences in gaze patterns during emotional face processing have been observed by prior works [61, 71] using high-end eye trackers. In this work, we used the low-cost *Eyetribe* device with 30Hz sampling rate to synchronously record eye movements along with the EEG signals, and compare male and female eye movement patterns. To compute *fixations* from raw gaze positions estimated by the tracker, we used the EyeMMV toolbox [40] and considered the transition from one fixation to another as a *saccade*. Upon extracting fixations and saccades, we adopted the features employed for valence recognition in [56], namely, saccade orientation, top-ten salient coordinates, saliency map and histograms of a) saccade slope, b) saccade length, c) saccade velocity, d) fixation duration, e) fixation count and f) saliency to compute an 825-dimensional feature vector for our analyses.

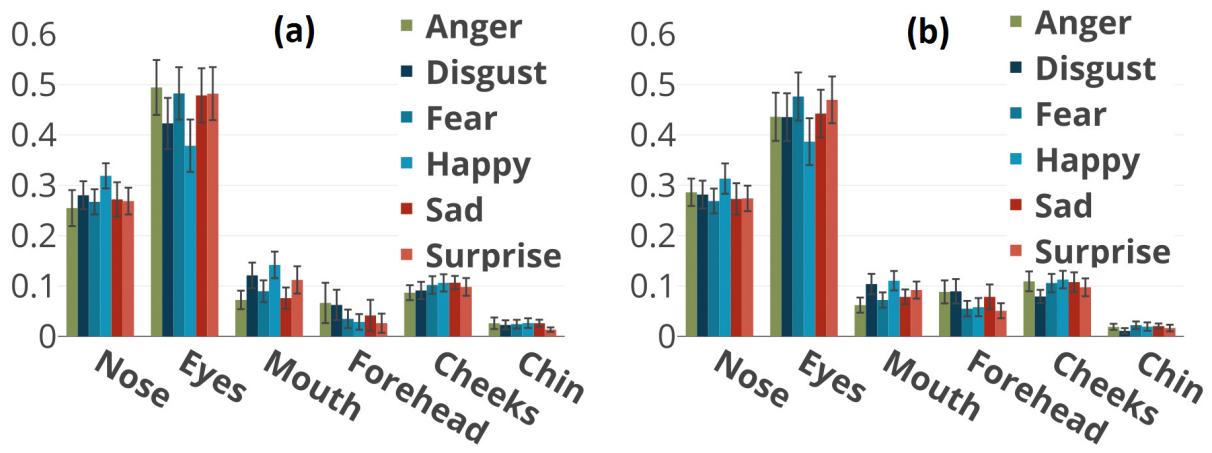

Figure 4.4 Fixation duration distributions: Region-wise fixation distribution across different emotions for (a) females and (b) males for both HI and LI conditions combined. Error bars denote unit standard error.

4.2.1 Fixation analysis

To validate observations on the fixating patterns of males and females, we computed the fixation duration distribution over six facial regions, namely, eyes, nose, mouth, cheeks, forehead and chin. Fig.4.3(a–d) show that the eye, mouth and nasal face regions attract maximum visual attention during facial emotion recognition, in line with prior studies such as [61, 70]. For both men and women, the time spent examining the eyes, nose and mouth account for over 80% of the total fixation duration, with eyes (≈45%) and nose (≈30%) attracting maximum attention as observed in [61]. No significant differences were observed between *HI* and *LI* morphs. We observe marginal increase in fixation duration in males for *LI* morphs than *HI* in nose region. A decrease in the difference between males and females for *LI* morphs when compared to *HI* morphs in eye, mouth and forehead regions. The fixation duration in mouth region is slightly higher for *LI* morphs than *HI*. Females have marginally higher fixation duration for *LI* morphs in cheeks region than *HI*. Focusing on the emotion-wise distributions(Figure-4.4), a high proportion of eye fixations is noted for all emotions except *happy*, while relatively higher visual attention is noted on the nose and mouth for this emotion consistent with the findings in [61] implying that these regions encode key emotional information concerning happiness. Overall, females were found to examine the eyes and mouth more for emotional cues, with males fixating more on the nose even though none of the differences were significant. The next section presents ER and GR results achieved with EEG and eye-based features.

Prior works such as [81] have also observed that females look up to the eyes for emotional cues, which is mirrored by longer female fixations around the eyes in the *HI* and *mouth mask* conditions. Fig.4.3(c) suggests that when eye information is unavailable, females tend to concentrate on the mouth and nasal regions for emotional cues.

Chapter 5

User Profiling with Implicit Signals

This section presents emotion recognition (ER) and gender recognition (GR) results obtained with EEG features, eye-based features and their combination, under the three considered conditions. Recognition experiments were performed on data compiled from trials where viewers correctly recognized the presented facial emotion.

5.1 Experimental settings

We considered (i) EEG features, (ii) attention-based features, (iii) concatenation of the two (termed *early fusion* or EF), and (iv) probabilistic fusion of the EEG and eye-based classifiers (*late fusion* or LF) for our analyses, in the three conditions during our recognition experiments. For LF, we employed the W_{est} technique [38] briefly described as follows. From the EEG and eye-based classifier outputs, the test set posterior probability for class j is computed as $\sum_{i=1}^{2} \alpha_i^* t_i p_i^j$, where i indexes the two modalities, p_i^j's denote posterior classifier probability for class j with modality i and $\{\alpha_i^*\}$ are the optimal weights maximizing test performance, as determined via a 2D grid search. If F_i denotes the training performance for the i^{th} modality, then the normalized training performance $t_i = \alpha_i F_i / \sum_{i=1}^{2} \alpha_i F_i$ for given α_i.

For benchmarking ER/GR performance, we considered the area under curve (**AUC**) performance measure. The AUC represents area under the ROC curve plotting true and false positive rates, and a random classifier will correspond to an AUC score of 0.5. Also, AUC is an appropriate metric while evaluating classifier performance on data with unbalanced positive and negative class sizes. As we are attempting recognition with few training data, we report ER/GR results over five repetitions of 10-fold cross validation (CV) (*i.e.*, total of 50 runs). CV is typically used to overcome the *over-fitting* problem on small datasets. In the following we discuss different classification approaches that are used in our analysis.

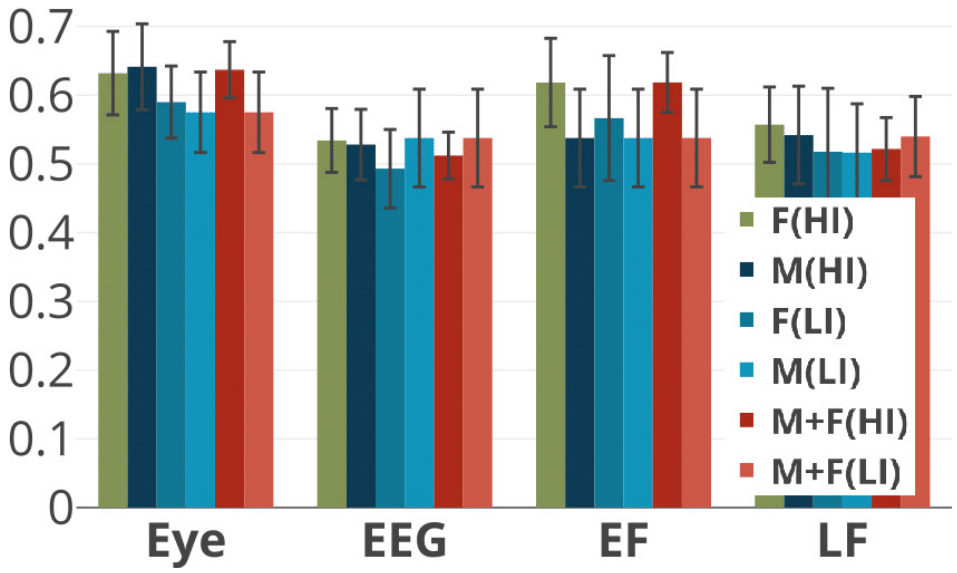

Figure 5.1 ER Results: Valence recognition results with HI and LI morphs.

5.2 Impact of Emotional Intensity

As explicit behavioral responses (Sec.3.2) and ERP-based (Sec.4.1.2) analyses suggest that (a) females are especially sensitive to negative facial emotions, and (b) there are cognitive differences in the processing of positive and negative emotions for females and males, we modeled ER as a binary classification problem, with an objective to distinguish between positive and negative emotional classes.

5.2.1 Valence Recognition

Fig. 5.1 presents valence recognition (VR) results obtained with male, female and male+female unimodal and multimodal features. Expectedly, better user-centered VR is achieved overall on examining responses to HI emotions (minimum AUC = 0.51 for HI vs 0.49 for LI morphs). Also, best VR performance is achieved with female data (mean AUC = 0.58 with HI morphs) as compared to male (mean AUC = 0.56, HI morphs) and male+female data (mean AUC = 0.543 for HI morphs) considering both unimodal and multimodal features. Interestingly, eye-based features (peak AUC = 0.64 with male data) are more discriminative for VR than EEG features (peak AUC = 0.53 with female data). Evidently, fusing the eye and EEG information is not beneficial even though a simple concatenation of the EEG and eye features is found to be superior to probabilistic fusion of unimodal results.

The obtained results are competitive against (and in some cases better than) prior approaches. Eye based features are found to achieve 52.5% valence recognition accuracy in [56], where emotions are induced by presenting diverse types of emotional scenes to viewers while only emotional faces employed in this work. Also, prior neural studies [3, 50] which employ music and movie content for emotion elicitation achieve only around 60% VR with lab-grade sensors. Our VR results are creditable especially

considering that low-cost EEG and eye tracking devices are used in this work and that our objective was not to induce emotions in viewers *per se*– the discriminability instead arises from the manner in which facial emotions are *perceived* by viewers. Superior performance of eye-based features may be attributed to the fact that in addition to the difference in fixation duration observable for *happy* (Fig. 4.4), the saccade and saliency based statistics are able to capture discrepancies in gazing patterns for positive and negative facial emotions [56]; in contrast, EEG data are vectorized and analyzed *as is* without any explicit feature extraction.

5.2.2 Gender Recognition

Table 5.1 presents eye and EEG-based GR results obtained with HI and LI emotions. Specifically, GR AUC scores with various modalities and classifiers[1] corresponding to *all* and *emotion-specific* trials are tabulated. Clearly better than chance GR is achievable with both EEG and eye-based features. An immediate observation from the tabulated results is that unlike for ER, very similar GR performance is noted for both HI and LI emotions, implying that gender specific differences are also effectively encoded in implicit viewer responses to mildly emotive faces. This is consistent with the explicit user behavioral results in Fig.3.3, where women are found to outperform men at recognizing LI facial emotions.

Focusing on specifics, EEG features considerably outperform eye-based features for GR and higher AUC scores are obtained with *emotion-specific* (exclusively negative emotion) features as compared to *emotion agnostic* features. This finding is consistent with Fig.3.3 and prior works such as [46, 88] which have noted significant differences in female neural responses to negative facial emotions. Particularly, the best EEG AUC scores for HI and LI morphs are obtained for the anger and disgust emotions respectively, for which the greatest difference in recognition rates (of 0.14 and 0.13) can be observed from Fig.3.3. These findings are salient and interesting, as Fig.3.3 and Table 5.1 present two different phenomena, namely, gender differences in ER and discriminability of male and female EEG patterns.

Surprisingly, multimodal fusion performs inferior to EEG/eye features. Also, marginally superior GR results are achieved with early fusion, suggesting that perhaps little complementary information is encoded between the EEG and eye-based features. Finally, RSVM is found to produce the best GR performance in most conditions implying that the extracted EEG and eye features are better separable in a higher-dimensional feature space.

5.2.3 Spatio-temporal EEG analysis

As EEG represents multi-channel time-series data, we also performed spatial and temporal analyses on the EEG signals to discover (i) the key EEG electrodes that capture gender differences in emotional face processing, and (ii) the critical time window for GR over the 4s of stimulus presentation (or visual emotional processing).

[1] with which optimal GR results are achieved for a given condition

Table 5.1 Table showing Gender Recognition results over different modalities and their fusion for HI and LI morphs using NB-NaiveBayes,RSVM-SVM with RBF kernel and LSVM-SVM with linear kernel.

AUC			HI	LI
All		EEG(NB)	**0.714 ± 0.002**	**0.600 ± 0.005**
		EYE(NB)	0.493 ± 0.013	0.481 ± 0.017
		Early fusion(RSVM)	0.522 ± 0.035	0.524 ± 0.022
		Late fusion(LSVM)	0.549 ± 0.022	0.523 ± 0.035
Emotion wise	**EEG (NB)**	A	**0.708 ± 0.064**	0.580 ±0.074
		D	0.673 ± 0.055	**0.696 ± 0.062**
		F	0.643 ± 0.059	0.596 ± 0.089
		H	0.696 ± 0.047	0.668 ± 0.046
		Sa	0.674 ± 0.048	0.634 ± 0.064
		Su	0.692 ± 0.048	0.636 ± 0.071
	EYE (NB)	A	**0.601 ± 0.021**	0.565 ± 0.031
		D	0.577 ± 0.011	**0.632 ± 0.009**
		F	0.595 ± 0.015	0.535 ± 0.029
		H	0.560 ± 0.021	0.538 ± 0.017
		Sa	0.539 ± 0.015	0.605 ± 0.030
		Su	0.555 ± 0.008	0.555 ± 0.018
	EF (RSVM)	A	0.555 ± 0.021	0.581 ± 0.037
		D	0.535 ± 0.024	**0.622 ± 0.025**
		F	**0.618 ± 0.011**	0.619 ± 0.041
		H	0.575 ± 0.017	0.597 ± 0.009
		Sa	0.540 ± 0.021	0.598 ± 0.024
		Su	0.579 ± 0.013	0.574 ± 0.014
	EF (RSVM)	A	0.543 ± 0.062	0.571 ± 0.081
		D	0.542 ± 0.044	**0.597 ± 0.093**
		F	0.519 ± 0.029	0.597 ± 0.170
		H	0.526 ± 0.031	0.508 ±0.017
		Sa	**0.573 ± 0.076**	0.584 ± 0.102
		Su	0.562 ± 0.073	0.568 ± 0.125

Fig.5.2(a,b) show the temporal variance in GR performance for HI and LI morphs when the 4s emotional face viewing time is split into four 1s-long windows. These results are computed considering emotion-specific EEG epochs and with information from all 14 electrodes. Overall, no consistent temporal trends concerning gender discriminability can be noted from the plots. However, higher AUC scores over temporal windows W1–W4 are achieved with LI morphs as compared to HI morphs. Also, slightly higher AUC metrics are noted for negative emotions especially with LI morphs, and the best and worst scores are obtained for the *fear* and *happy* emotions. Results of spatial analyses to identify the top three electrodes that best encode gender differences in emotion processing (EEG data over all 4s is used for this analysis.) are shown in Fig.5.2(c). Again, superior GR is noted with LI morphs in this scenario, and five of the six electrodes with the best AUC scores correspond to the *frontal* lobe. This

result is consistent with the findings in [48], which notes that gender differences in emotion processing are best encoded in the prefrontal cortex.

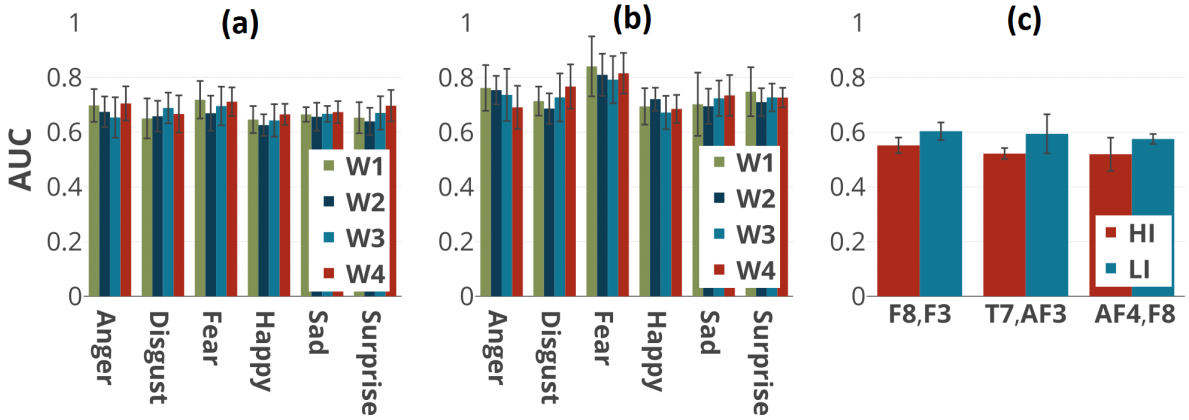

Figure 5.2 (a,b) Spatio-temporal EEG analyses: GR results for HI and LI morphs over temporal windows (W1–W4). (c) Top 3 electrodes based on AUC metric for HI and LI morphs. x-axis denotes the rank of the electrodes in decreasing order. Error bars denote unit standard error.

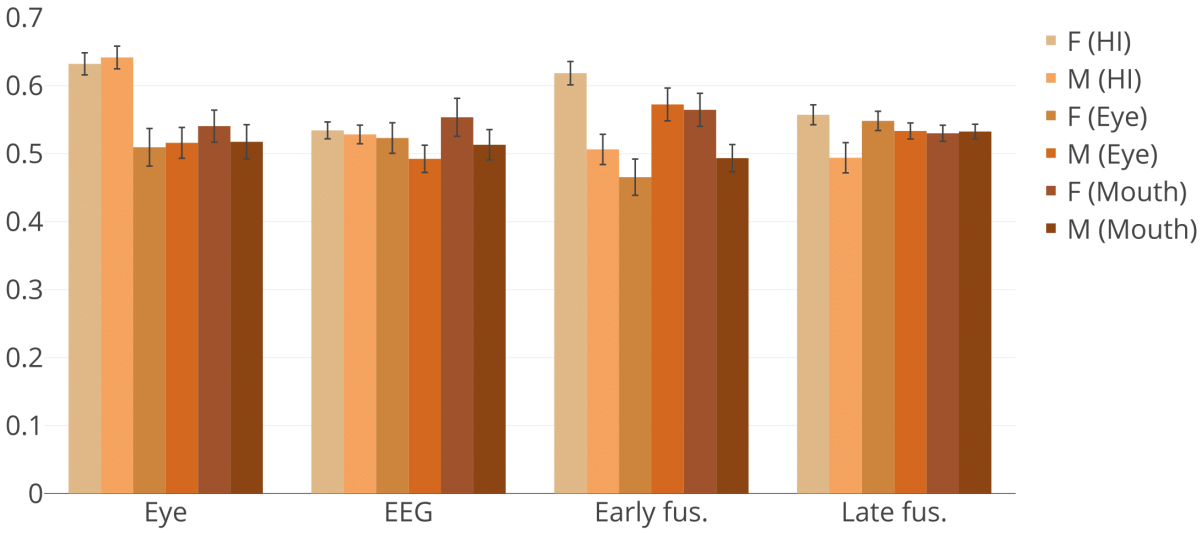

Figure 5.3 ER performance for *HI*, *eyemask* and *mouthmask* epochs for males and females.

5.3 Impact of facial regions

5.3.1 Valence Recognition

As explicit behavioral responses (Sec.3.2) and ERP-based (Sec.4.1.2) analyses suggest that (a) females are especially sensitive to negative facial emotions, and (b) there are cognitive differences in the

processing of positive and negative emotions for females, we attempted to classify positive (H, Su) and negative (A,D,F,S) emotions.Fig.5.3 presents ER results obtained with the LDA classifier for male and female data, with unimodal and multimodal features for the three conditions. The best ER results are expectedly achieved for the *HI* condition, and with eye-based features (AUC = 0.63 for females, and 0.64 for males), while minimal recognition is noted for the *eye mask* condition reiterating the importance of eyes towards encoding emotional cues. EEG-based features only produce slightly better than chance recognition with a peak AUC score of 0.54. Multimodal recognition is slightly better than unimodal, and higher AUC scores are achieved in the masked conditions suggesting the complementarity of EEG and eye-based features when only partial facial information is available.

While observed ER results are modest, they are still better or competitive with respect to prior eye and neural based ER approaches. Eye based features are found to achieve 52.5% valence recognition accuracy in [56], where emotions are induced in viewers by presenting a diverse types of emotional scenes to viewers as against emotional faces specifically employed in this work. Also, prior neural-based emotional studies [3, 50] achieve only around 60% valence recognition with lab-grade sensors. On the contrary, implicit behavioral cues are acquired with low-cost EEG and eye tracking devices in this work.

5.3.2 Gender Recognition

Table 5.2 presents GR results with features extracted for (a) *all* faces, and (b) *emotion-specific* faces for the three conditions. The classifier employed for GR is also specified with parenthesis. As with ER, the best GR performance (AUC = 0.71) is observed for the *HI* condition but with EEG features. With both EEG and eye-based cues, slightly higher GR performance is noted for the *eye mask* condition as compared to the *mouth mask* condition considering viewer responses for *all* as well as *emotion specific* faces. Overall, EEG features appear to encode gender differences better than eye-based features. Also, early or late fusion of the two modalities does not improve GR performance by much, even though one can note that fusion performs better than at least one of the modalities in the *eye mask* and *mouth mask* conditions.

A close examination of the emotion-wise results reveals that optimal GR is mainly achieved with viewer responses observed for negative facial emotions. This again is consistent with the observations from explicit user responses (Sec.3.2), where females were found to recognize negative facial emotions better than males. Overall, the obtained results point to systematic differences between males and females in the cognitive processing of emotional faces. Focusing on the standard deviation (sd) in AUC scores for emotion-wise GR, one can note much higher sd values in the *eye mask* and *mouth mask* conditions as compared to *no mask*, indicating the difficulty in recognizing facial emotions under occlusion conditions. However, relatively lower sd values are noted with LF implying that probabilistic fusion enables better exploitation of the complementarity in the eye and brain-based signals, resulting in more robust GR.

Table 5.2 GR with different conditions/modalities. The best classifier is denoted within parenthesis. Highest AUC for a given condition is shown in bold.

AUC			HI	Eye mask	Mouth mask
All		**EEG (NB)**	**0.71±0.01**	**0.69±0.03**	**0.65±0.04**
		EYE (NB)	0.49±0.01	0.47±0.06	0.52±0.05
		EF (RSVM)	0.52±0.03	0.52±0.07	0.52±0.06
		LF (RSVM)	0.55±0.02	0.54±0.05	0.61±0.07
Emotion Wise	**EEG (NB)**	A	**0.71±0.06**	0.61±0.16	**0.67±0.07**
		D	0.67±0.05	0.59±0.06	0.65±0.14
		F	0.64±0.06	0.60±0.14	0.56±0.12
		H	0.69±0.05	0.58±0.07	0.62±0.08
		Sa	0.67±0.05	**0.65±0.08**	0.59±0.08
		Su	0.69±0.05	0.63±0.08	0.65±0.09
	EYE (NB)	A	**0.60±0.02**	0.47±0.20	**0.59±0.16**
		D	0.57±0.01	0.48±0.14	0.52±0.27
		F	0.59±0.01	**0.68±0.25**	0.58±0.17
		H	0.56±0.02	0.55±0.13	0.54±0.12
		Sa	0.54±0.01	0.44±0.15	0.45±0.13
		Su	0.55±0.01	0.62±0.13	0.49±0.16
	EF (RSVM)	A	0.55±0.02	0.32±0.25	0.39±0.18
		D	0.53±0.02	0.52±0.18	0.50±0.24
		F	**0.62±0.01**	**0.70±0.22**	0.52±0.18
		H	0.57±0.02	0.57±0.13	**0.52±0.14**
		Sa	0.54±0.02	0.53±0.16	0.50±0.14
		Su	0.58±0.01	0.43±0.13	0.51±0.16
	LF (RSVM)	A	0.54±0.06	0.57±0.15	0.59±0.09
		D	0.54±0.04	0.59±0.10	0.57±0.14
		F	0.52±0.03	**0.64±0.16**	**0.69±0.12**
		H	0.52±0.03	0.56±0.09	0.55±0.10
		Sa	**0.57±0.07**	0.51±0.04	0.58±0.11
		Su	0.56±0.07	0.58±0.08	0.58±0.08

5.3.3 Spatio-temporal EEG analysis

As EEG represents multi-channel time-series data, we performed spatial and temporal analyses on the EEG signals to discover (i) the critical time window for GR over the 4s of stimulus presentation, and (ii) the key EEG electrodes encoding gender differences in emotional face processing.

Fig.5.4(a–c) show the temporal variance in GR performance for three considered conditions when the 4s presentation time is split into four 1s-long windows. These results are computed considering emotion-specific EEG epochs using RSVM. No significant difference in GR performance is observed over time for the *no mask* scenario, and an AUC score of around 0.65 can be consistently noted over time. Barring exceptions, the GR AUC scores for the *eye mask* and *mouth mask* conditions are lower than for the *no mask* condition. Also, maximum GR is predominantly noted for later temporal windows for mask conditions, implying that the discriminative cognitive processing between males and females happens later in time when only partial emotional information is made available to viewers. For the *eye mask* scenario, maximum GR AUC scores are achieved for anger (AUC = 0.71 in W3), disgust (AUC

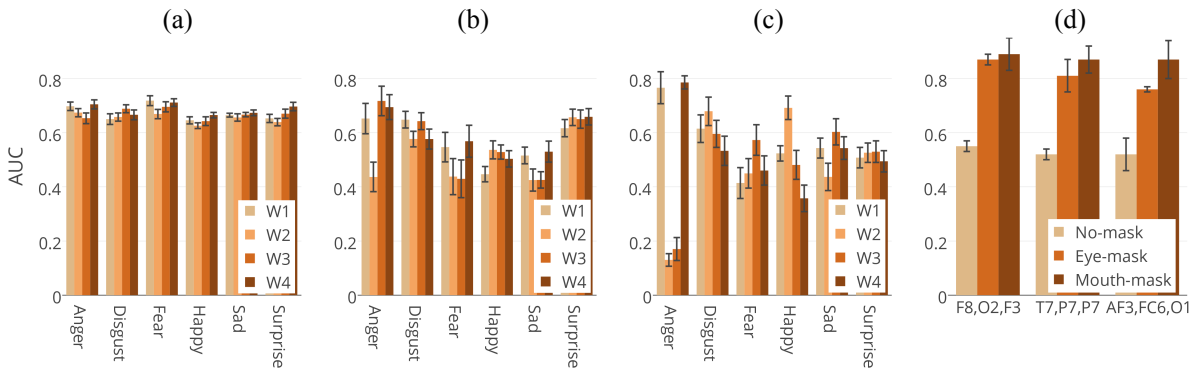

Figure 5.4 (a–c): Temporal variance in GR performance for the *no mask*, *eye mask* and *mouth mask* conditions. (d) Top three electrodes showing best GR performance for the three conditions.

= 0.65 in W1) and surprise (AUC = 0.65 in W2, W3 and W4). For the *mouth mask* condition, highest AUCs are achieved for anger (AUC = 0.76 in W1 and 0.78 in W4), disgust (AUC = 0.68 in W2) and happy (AUC=0.69 in W2). However, low GR performance is generally noted over all time windows for the fear and surprise emotions in this condition, attributable to the general difficulty in recognizing these emotions when mouth-related information is unavailable.

Results of spatial analysis to identify the (top three) electrodes that best encode gender differences in emotion processing[2] are shown in Fig.5.4(d). In the *no mask* condition, very similar AUC scores are observed for the frontal F8 (0.55), AF3 (0.52) and T7 (0.52) electrodes. Much higher numbers are noted for the *eye mask* and *mouth mask* conditions, suggesting that region-specific differences in cognitive processing for males and females manifest under occlusion conditions. For the *eye mask* condition, highest AUCs are noted for O2 (0.87), P7 (0.81) and FC6 (0.76), while for *mouth mask* F3 (AUC = 0.89), P7 (AUC = 0.87) and O1 (AUC = 0.87). An interesting observation is that a majority of the electrodes are located on the left hemisphere of the brain (see Emotiv electrode configuration for details).

5.4 Deep learning and AdaBoost

5.4.1 Deep learning Architecture

The previous part of the work has used preprocessed(for eye blinks, muscle movements and other external factors) EEG data for GR and ER. One of the foremost advantages of deep learning is the ability of the models to learn the features from raw data that are optimized for the task thereby avoiding the need to process the data for noise and human interference in constructing the features. Unlike domains like computer vision and natural language processing, the temporal dynamics of human brain are unclear and hence the definition of features for some tasks is weakly-robust due to our limited understanding of

[2]EEG data over all 4s is used for this analysis.

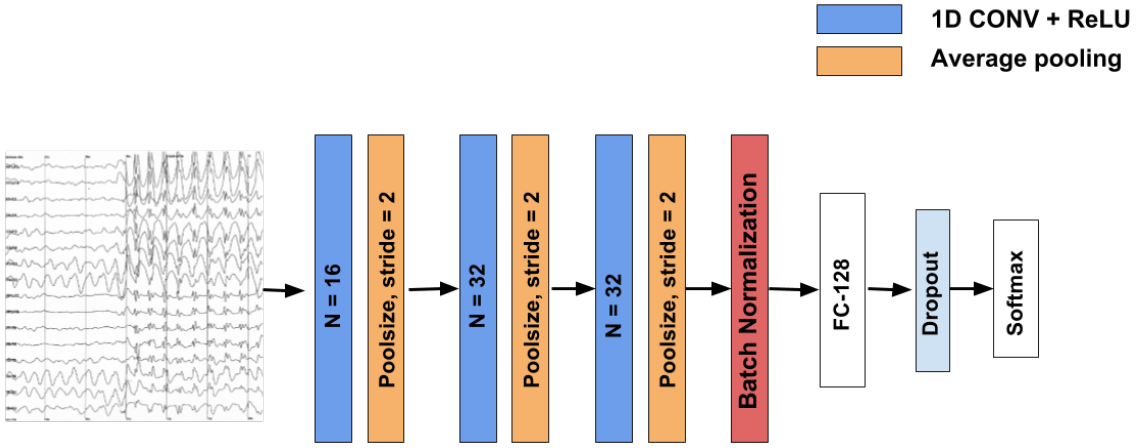

Figure 5.5 Deep learning architecture showing various layers in the model and parameters.

the same. The deep feature learning could possibly reject the eye-blink and muscle movement activities by projecting the noisy data to a robust feature space to preserving the task relevance and invariant to noise. These factors inspired us to use raw EEG data from epochs extracted as described in sec. 4.1.1 without subjecting them to filtering or later steps of preprocessing to train the deep learning model. All the above said advantages make deep learning one of the preferred methods to analyze brain signals(in conclusion-add eeg data scarcity and explainability).

Table 5.3 Parameters for the deep learning model

Parameter	Value
Learning rate	0.01
Kernel shape	3
Stride size	2
Pool size	2
Batch size	32
# Kernels(layer wise)	16,32,32
Momentum	0.9
Weight decay	0.0001
Dropout	0.1

We propose a 3-layer CNN model for classification of gender and emotional state. Given the sample size of the experiment, we found that a three layer model is the best fit. Three convolution together with ReLU activation function and average pooling layers are stacked together to extract the task specific features. The convolutions employed are 1-dimensional along time. To prevent over-fitting, we used batch normalization after feature learning and dropout after the fully connected layer with 128 neurons. We finally classify with *softmax* over 2 output neurons. The number of kernels increase with the depth of the convolution network as analogous to the VGG architecture[66]. We have optimized the network for categorical cross entropy using vanilla stochastic gradient descent with Nestrov momentum and weight decay. The learning rate was dropped to 0.1 times on over-fitting. The detail hyper-parameters

are specified in table-5.3. To decrease the dimensionality of the problem and avoid over-fitting without sacrificing the temporal dependencies in the EEG data, we downsampled the data to 32Hz from the initial sampling frequency of 128Hz. This helped reduce the dimensionality to 128(per channel) from the original 512 at 128Hz for 4 seconds of stimuli duration. Though this waives the typical advantage of high temporal resolution with EEG devices over other brain imaging methods, it can be considered an inspiration of explore further low-cost devices with low sampling rate for user profiling. A tenth of training data was used for validation to monitor over-fitting. The experiments were conducted in Keras[14] with Tensorflow backend on a 16GB CPU machine with an intel i5 processor.

5.4.1.1 Training Strategy

The deep learning model defined above was trained for GR and ER problems in different ways. The model was trained for GR problem in the traditional sense with *He normal* initializers(draws weights from a normal distribution with zero mean and std=$\sqrt{(2/fan_{in})}$). We adopted a two stage training process for ER. In first stage, the model was trained for ER with EEG signals corresponding to all four morph conditions. In second stage, it was trained with the data corresponding to given morph condition. The intuition was to enable the model extract valence specific features irrespective of the morph intensity and occlusions in the first stage and fine-tune it in the second stage to learn morph specific features to enable better ER. In all cases it was made sure that the train and test data do not overlap.

5.4.2 AdaBoost

The gaze features are often sparse because of strong localization of emotional importance to certain face regions and the experimental design of utilizing a low-cost device with low sampling frequency. It is mostly redundant in employing all the features corresponding to various face regions for GR and ER. Hence we considered to prune for the necessary features for task. One approach towards feature selection is based on sequential addition(or removal) to obtain best performance as described in [56] using sequential forward(SFS) or backward(SBS) selection. We used AdaBoost, an ensemble method known popularly for face detection [76] and facial expression recognition [65] for GR and ER. [65] also emphasized on the importance of local features for expression recognition and utilizes Gaussian and Gabor features from images extracted at scales to generate composite features for the recognition task using AdaBoost. The boosting algorithm combines a large number of weak classifiers (typically decision stumps) trained using individual features, each performing slightly better than chance to achieve better performance. The final ensemble decision is based on the weighted output of each weak classifier. Since the gaze features we described in section- 4.1.1 are similar to the ones utilized in [65] in capturing local dependences using simple representations such as histograms of basic elements like fixations, saccades and saliency, this inspired us to use AdaBoost to analyze the gaze features. We have used the SAMME (Stagewise Additive Modeling using a Multi-class Exponential loss function) algorithm [26] for training. SAMME overcomes the pitfalls of multi-class classification with vanilla AdaBoost[20] and is equivalent

to the same for a binary classification problem. Our choice of refraining from the early fusion method is due to the fact that CNN input features are spatio-temporally dependent unlike our approach in table-5.2 where the features were acquired after PCA on concatenated channel data. The results are tabulated as table-5.4.

5.4.3 Emotion Recognition

An initial and quick observation from our ER results reveals that gaze features perform significantly better than the brain signals. They have achieved near perfect performance for most morph conditions barring exception of *LI* morphs in female subjects. The highest performance was noticed for *HI* morphs in males(AUC = 0.99) similar to our analyses with traditional methods in [9]. However, the ER in females(AUC = 0.98) for the same was extremely close. Similar effect was noticed with EEG signals though the performance was observed to be near chance level. The difference was not unshunnable with the late fusion strategy due to the fall in female ER performance. While the ER with gaze features for *LI* morphs in females is noticeably lower than males(AUC = 0.96), the females outperform males in fusion(0.82 vs 0.81) with a minimal margin. The fusion method helped the ER outperform the unimodal methods(females-0.97 vs 0.98, males-0.98 vs 0.99) modestly for *eyemask* morphs in both males and females and noticeably in males(0.91 vs 0.94) for *mouthmask* morphs. The best performance with EEG was observed with *mouthmask* morphs with females outperforming males(0.70 vs 0.74). Though no significant difference is noticed between *eyemask* and *mouthmask* morphs in ER with gaze features, the *mouthmask* morphs exceeding the *eyemask* morphs performance could be due to the fact that eyes are more informative in emotion analysis than mouth and nasal region. This also reaffirms our observations in [10].

To summarize, gaze features outperform the EEG signals in ER during facial emotion recognition, EEG signals could give a reliable ER with mouthmasked morphs and probabilistic fusion of features is helpful for *LI* morphs with females and marginally helpful with other conditions. Emotional faces with partial information due to occlusion of mouth region provide better emotional cues than the ones with occlusion of eye region. The task of ER could be explored through emotional perception unlike previous works like [3, 50] that explicitly invoke emotions for the same.

The obtained results are better than prior approaches. We achieved an ER(two class valence recognition) (AUC = 0.99) with emotion perception task when compared to [9, 10] where the highest performance was reported to be 0.63 AUC score for no mask faces(*HI*) with gaze features. Eye based features are found to achieve 52.5% valence recognition accuracy in [56], where emotions are induced by presenting diverse types of emotional scenes to viewers while only emotional faces employed in this work. Also, prior neural studies [3, 50] which employ music and movie content for emotion elicitation achieve only around 60% VR with lab-grade sensors. Our objective was not to induce emotions in viewers *per se*– the discriminability instead arises from the manner in which facial emotions are *perceived* by viewers. Superior performance of eye-based features may be attributed to the fact that the saccade and saliency based statistics are able to capture discrepancies in gazing patterns for positive and nega-

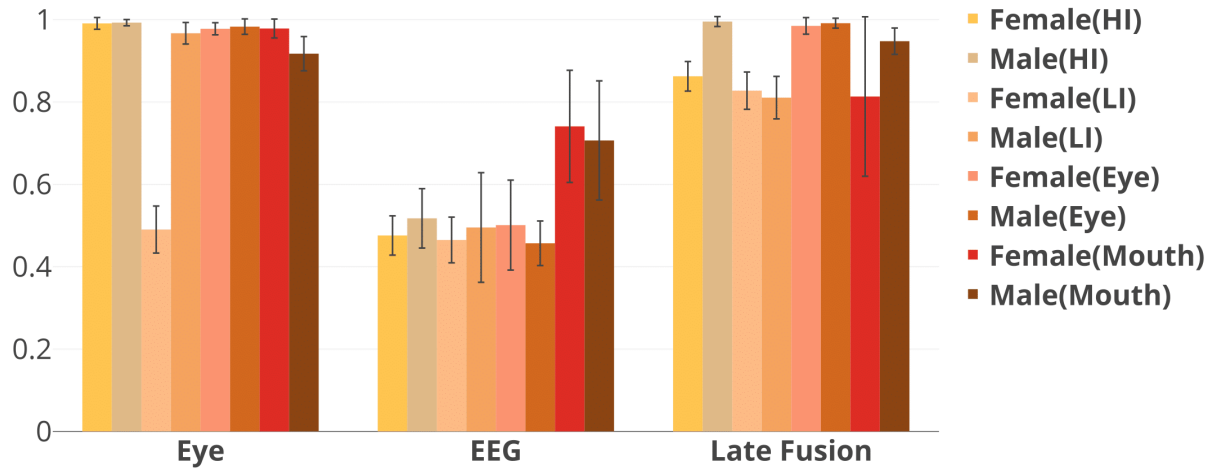

Figure 5.6 ER Results: valence recognition results with HI and LI morphs.

tive facial emotions [56]; in contrast, EEG data is purely based on the subject's neural activity during recognition of the emotions.

5.4.4 Gender Recognition

The deep learning model and AdaBoost algorithm have significantly outperformed the SVMs and other traditional classifiers summarized in table-5.2. While EEG performs better than gaze features for *unmasked* morphs(HI-0.93 vs 0.52 and LI-0.88 vs 0.85), the gaze features significantly outperform EEG for *masked* morphs(eyemask-0.97 vs 0.89 and mouthmask-0.97 vs 0.88). Late fusion method performs close to the best modality for all morphs and better for *eyemask* morphs. For emotion-wise GR with EEG, *happy* morphs stimulate most differences in emotional perception between males and females however, it is to be noted that Sad and Happy perform closely. The emotion-wise comparison of EEG and gaze based GR confirms our previous observation of EEG outperforming gaze for HI and LI morphs but the inverse is true for *masked* morphs. It is interesting to note that the GR is better for HI morphs than LI using EEG however, no such significant difference is observed with gaze. Gaze features reveals that *anger* and *sad* are the most discriminable emotions for *unmasked* and *masked* morphs respectively which is similar to our analysis with EEG where both *anger* and *sad* are close to *happy*.

The late fusion strategy did not significantly improve the performance however, it outperformed the best of the either modalities for *mouthmask* morphs for *disgust* and *surprise* emotions. Both EEG and late fusion strategy confirmed that LI morphs are least helpful in GR, masked morphs help in invoking better gender differences in emotional perception with *mouthmask* morphs doing it the best. This has already been established through our initial evaluation of fixation duration(fig.4.3(d)), we observe a significant difference between males and females for *mouthmask* morphs.

Table 5.4 Gender Recognition results over different modalities and their fusion for HI, LI, Eyemask and Mouthmask morphs using CNN(EEG) and AdaBoost(Gaze).

AUC			HI	LI	Eyemask	Mouthmask
All		EEG(CNN)	**0.931 ± 0.011**	**0.882 ± 0.020**	0.892 ± 0.038	0.886 ± 0.023
		EYE(Ada)	0.521 ± 0.016	0.538 ± 0.023	0.969 ± 0.015	**0.966 ± 0.014**
		Late fusion	0.920 ± 0.015	0.850 ± 0.045	**0.974 ± 0.017**	0.946 ± 0.037
Emotion wise	EEG (CNN)	A	0.757 ± 0.113	0.719 ± 0.086	0.626 ± 0.165	0.676 ± 0.157
		D	0.758 ± 0.047	0.644 ± 0.080	0.594 ± 0.161	0.673 ± 0.108
		F	0.738 ± 0.083	0.618 ± 0.107	0.542 ± 0.128	0.585 ± 0.199
		H	**0.790 ± 0.049**	0.701 ± 0.054	**0.714 ± 0.082**	**0.751 ± 0.051**
		Sa	0.739 ± 0.089	**0.737 ± 0.078**	0.689 ± 0.097	0.731 ± 0.079
		Su	0.720 ± 0.076	0.636 ± 0.042	0.649 ± 0.098	0.652 ± 0.067
	EYE (Ada)	A	**0.651 ± 0.064**	**0.651 ± 0.079**	0.840 ± 0.160	0.957 ± 0.038
		D	0.591 ± 0.038	0.600 ± 0.068	0.950 ± 0.045	0.910 ± 0.108
		F	0.569 ± 0.078	0.528 ± 0.157	0.950 ± 0.067	0.881 ± 0.075
		H	0.574 ± 0.090	0.529 ± 0.050	0.931 ± 0.049	0.929 ± 0.051
		Sa	0.587 ± 0.087	0.576 ± 0.068	**0.959 ± 0.051**	**0.958 ± 0.041**
		Su	0.530 ± 0.093	0.546 ± 0.052	0.938 ± 0.067	0.921 ± 0.107
	LF	A	0.720 ± 0.082	0.651 ± 0.113	0.835 ± 0.167	0.860 ± 0.132
		D	0.676 ± 0.083	0.557 ± 0.153	**0.939 ± 0.070**	**0.967 ± 0.105**
		F	0.675 ± 0.140	0.470 ± 0.164	0.887 ± 0.139	0.869 ± 0.097
		H	**0.770 ± 0.050**	**0.683 ± 0.088**	0.897 ± 0.042	0.926 ± 0.060
		Sa	0.729 ± 0.062	0.555 ± 0.107	**0.939 ± 0.062**	0.843 ± 0.133
		Su	0.723 ± 0.090	0.604 ± 0.105	0.826 ± 0.142	0.943 ± 0.058

5.4.5 Spatio-temporal EEG analysis

Table 5.5 Spatial performance to evaluate the importance of channels in decreasing order. The parentheses contain the channel for which the score was achieved.

	HI	LI	Eyemask	Mouthmask
1	0.8266 (AF4)	0.8008 (F3)	0.7160 (F4)	0.7135 (F3)
2	0.8200 (F3)	0.7772 (O2)	0.7143 (F3)	0.7124 (F4)
3	0.8131 (O2)	0.7745 (AF4)	0.7055 (AF4)	0.7107 (AF3)

As EEG represents multi-channel time-series data, we also performed spatial and temporal analyses on the EEG signals to discover (i) the key EEG electrodes that capture gender differences in emotional face processing, and (ii) the critical time window for GR over the 4s(downsampled) of stimulus presentation (or visual emotional processing).

A reliable GR was achieved for shorter time windows at various instants of stimuli presentation suggesting faster(less than 4s of stimuli presentation) GR possible and existence of gender differences throughout the stimuli duration(since the ERPs in sec. 4.1.2 only suggest gender differences in the first second from stimuli onset). Highest performance was observed with HI morphs while the least with

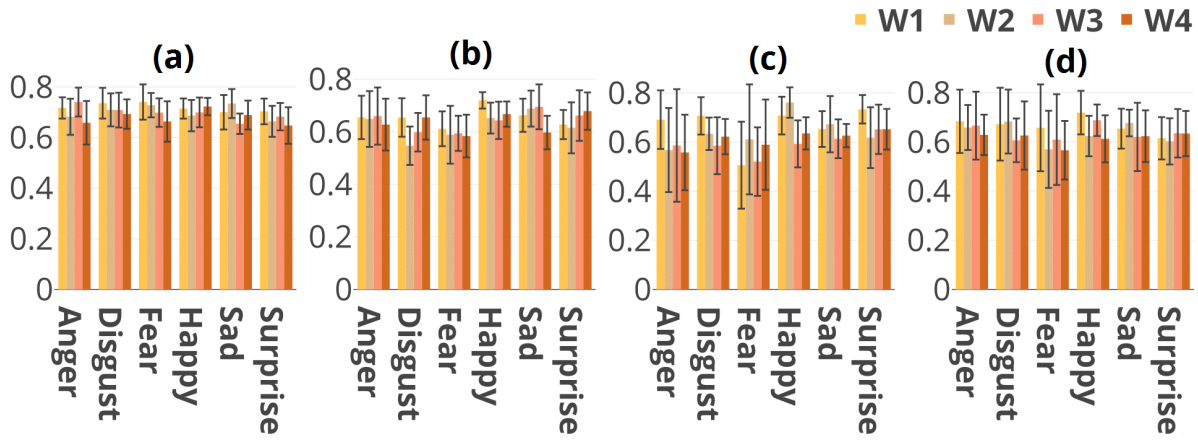

Figure 5.7 Temporal EEG analyses: GR results for HI, LI, Eyemask and Mouthmask morphs respectively over temporal windows.

eyemask which is consistent with the trend in table-5.4. The temporal analysis suggested no significant differences in GR among the windows however, best performance was observed for W-1,2 for all morph conditions suggesting early elicitation of differences in emotion processing consistent with our ERP analysis sec.4.1.2. This difference is most observable for eyemask and marginally for mouthmask morphs(D,F,H,Sa). The masked morphs report higher standard deviation than full face morphs. This might be due to the difficulty induced due to occlusion of face regions in emotion recognition.

We evaluated each channel for their contribution to GR with the deep learning architecture configuration (fig.5.5) except with a slight modification of reducing the number of input channels to fit only one channel. The spatial analysis suggests possible GR with fewer channel devices that are available in market. Consistent with the findings in [9, 10, 48] we observed best performance with channels corresponding to Frontal and Occipital lobes of brain. The single electrode GR followed the trend as exhibited by table-5.4 *i.e., HI*, *LI*, *eyemask* and *mouthmask* morphs in descending order of GR. We observed a symmetric pattern among the best performing channels .i.e AF3 and AF4, and F3 and F4. While psychologists suggest the importance and utility for the brain hemispheres[75] in emotion recognition, its discussion deviates from the scope of our work.

Chapter 6

Conclusions

Of late, there has been considerable interest in the use of technologies such as EEG and eye-tracking to study the cognitive and emotional behavior of users via implicit behavioral cues. Even as potential applications range from affect estimation [37, 3], prediction of erroneous actions in collaborative settings [74] and evaluating the design of data visualization interfaces [54], many of these studies have been limited to lab settings with bulky and intrusive sensors constraining the manner in which users can express themselves.

A key impediment towards employing off-the-shelf sensing devices, which enable naturalistic user behavior in interactive applications has been their poor signal quality, and skepticism regarding the nature of 'useful' information they actually encode. In this regard, we believe that this work adds on to a small body of works such as [74] that have demonstrated the encoding of meaningful information with commercial sensors.

In particular, our work employs the wireless and wearable *Emotiv* EEG device along with the low-cost *Eyetribe* eye tracker to discover gender differences in cognitive and gazing behavior during a facial emotion recognition task. The validity of our study is endorsed by the multiple correlations observed between *explicit* user responses and their *implicit* behavioral patterns. To begin with, women are found to recognize negative facial emotions quickly and more accurately in our experiment, and correspondingly, higher female ERP amplitudes are noted for negative emotions; good GR performance with EEG, eye features and their fusion is also noted for negative emotions. The maximum difference in female and male recognition rates for HI and LI morphs are noted for the *angry* and *disgust* emotions, and the peak(or close to peak) GR AUC scores are noted for anger and sad. This work expressly examined the utility of implicit behavioral cues, in the form of EEG brain signals and eye movements to encode gender-specific information during visual processing of emotional faces. In addition to the use of unoccluded (*HI* and *LI*) emotive faces, we also employed occluded (*eye* and *mouth masked*) faces to this end.

The existence of gender differences in emotional experience could be attributed to several factors like social structure, environment and evolution, way of living and social stereotypes[18]. Though our current task does not involve explicit invoking of emotions through stimuli i.e the task of emotional

experience where gender differences were found in emotional experience of negative emotions. We have encountered a qualitative existence of gender differences in recognition(or analysis) of emotions similar to [18]. A near perfect GR with EEG (peak AUC of 0.93), eye features (peak AUC of 0.97) and their fusion (peak AUC of 0.97), as well as valence recognition (peak AUC of 0.99 with eye data) is achieved using the deep learning framework and AdaBoost algorithm respectively. The use of minimally intrusive and low-cost devices for studying user behavior makes our findings ecologically valid, and enhances the possibility of large-scale user profiling without traditional preprocessing steps(in case of EEG). Emotion and gender recognition experiments with EEG and eye-based features, as well as their combination revealed that optimal recognition of emotional valence is achieved with eye-based features, while best GR performance is achieved with EEG with all machine learning methods used. Though the best GR with emotion-specific features is observed with *happy* in case of EEG, the negative emotions(*anger* and *sad*) perform close to it. Also the gaze features confirm that the best GR is achieved with negative emotions which is consistent with our observations from table-5.2. EEG features generally outperform eye-based features for the purpose of GR. The fusion of EEG and gaze features has not significantly contributed to the performance however it is to be noted that in case of eyemask morphs, the performance is higher than that of either of the modalities. Similarly for emotion-wise GR, the mouthmask morphs outperformed the best of the modalities.

The gender differences in facial emotion perception have been exploited to achieve excellent performance in terms of ER and GR performance, especially considering that implicit user behavioral cues are acquired with low cost sensors, and also consistent with prior observations. Lithari *et al.* [46] noted amplitude and phase-related differences for positive and negative emotions in males and females, and some of these findings are mirrored in our ERP analyses. Several eye tracking works (*e.g.*, see [23]) have noted enhanced female sensitivity for negative emotions, while other neural studies [48] have noted that gender differences in emotion processing are best encoded by the frontal and occipital brain lobes. As strong GR (max AUC = 0.97) is achievable with the considered features. This does not involve manual intervention in preprocessing the data for eyeblinks, muscle movements and electrical interference. Thus suggesting a real-time utilization of such methods of user profiling using low-cost devices with fewer electrodes and low sampling frequency(as demonstrated from our spatial analyses results and downsampling to 32Hz for deep learning model). This also offers a chance to further exploit the low-cost devices for more complex BCI applications that could make the BCIs affordable to general population and more importantly, the people with disabilities.

The larger aim of this work is to predict user demographics via implicit and anonymous behavioral cues so as to implement intelligent advertising, recommendation systems and mental health monitoring systems(for disorders like Alexithymia).Some of the applications are described here. A recommendation system could be built by evaluating user engagement in advertisements/movies– movie reviews can be highly impacted by certain sequences in it, such reviews are often biased and do not provide a holistic feedback of the entire movie. Using our emotion recognition framework, one can evaluate movie sequences for distraction, for a better rating system and a feedback system for an immersive cinematic

experience. Emotional human faces (especially happy and anger) are ubiquitous on e-commerce websites specifically in categories like clothing and accessories. These can act as stimuli to obtain EEG response and predict the gender to recommend appropriate products. As the window based analysis has shown, the stimulus time could be reduced to close to a second which is practically more feasible system so as to accelerate the GR and provide a smooth customer experience. Furthermore, computers need to understand and adapt to user emotions for effective human-computer interaction, and gender-specific emotional aspects can be effectively utilized by affective interfaces for manipulating user behavior.

Related Publications

- **Maneesh Bilalpur**, Syed Mostafa Kia, Manisha Chawla, Tat-Seng Chua, Ramanathan Subramanian, "Gender and Emotion Recognition with Implicit User Signals", International Conference on Multimodal Interaction (ICMI), 2017.
- **Maneesh Bilalpur**, Syed Mostafa Kia, Tat-Seng Chua and Ramanathan Subramanian, "Discovering Gender Differences in Facial Emotion Recognition via Implicit Behavioral Cues", Affective Computing & Intelligent Interaction (ACII), 2017.
- **Maneesh Bilalpur**, Syed Mostafa Kia, Mohan Kankanhalli, Ramanathan Subramanian, "Implicit Signals for Gender and Emotion Recognition", submitted to the IEEE Transactions on Affective Computing.
- Viral Parekh, **Maneesh Bilalpur**, Jawahar. C.V., Ramanathan Subramanian, "EEG-based Cognitive Load Estimation Across Visualizations".

Appendix A

EEG-based Cognitive Load Estimation Across Visualizations

Visual representation of data is perceivably more intuitive representation of information over other methods such as data sonification. Evaluating such representations has led to the development of ***visual analytics*** (VA). Knowledge from multiple domains of HCI, data mining, and behavior and psychology are employed for this purpose [1]. With the growing data needs and analytics, an efficient way of evaluating the data representations is the need of the hour. This work focuses on utilizing the cognitive workload induced by the visualizations to evaluate their effectiveness.

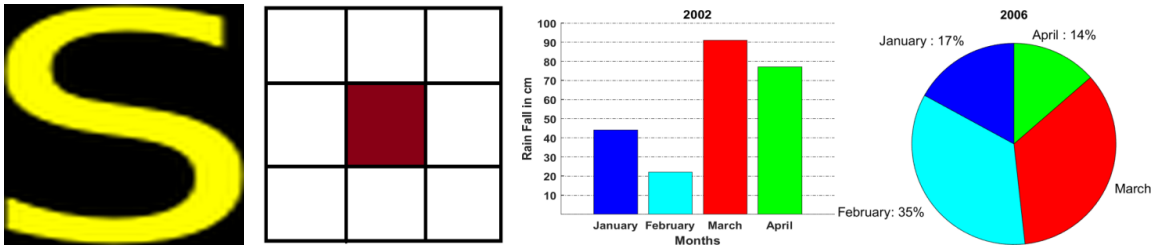

Figure A.1 Various visualizations used in the problem *character*, *position*, *bar* and *pie*.

The purpose of using visual interfaces is to express information in an easily understandable way, to put it in scientific terms, to convey information using means which would invoke lower cognitive load when compared to traditional methods. Hence, cognitive load has been a crucial concept for evaluating task difficulty. Cognitive load [13] is described as amount of mental resources consumed to perform a task. This could be due to intrinsic or extrinsic factors. The intrinsic factors are typically associated with task design while in the case of visual interfaces, the extrinsic factors are the visualizations used. This work aims at using the implicit responses towards replacing the traditional user experience studies for evaluating tasks or designs, in particular, visual interfaces [60, 54].

Utilizing implicit responses through eye movements [31, 59] and brain imaging methods such as EEG [4, 80] or fNIRS [54] for Viz evaluation has been a significant part of Neuroergonomics. These methods are non-invasive and do not require an explicit response from the user in addition to being light-weight commercial devices that make easy deployment feasible. However,the drawback lies in the

fact that these methods are often visualization dependant and difficult to generalize over different Viz UIs.

The aim of the current work is to investigate the possibility of a generalizable framework for workload classification, particularly for visual interfaces. The applications for a successful framework could be to compare multiple visual interfaces to decide upon the best possible representation. This could further replace traditional user based evaluation that are today's standards for evaluating such applications.

Traditional Machine learning methods often required hand-tuned features, with the advent of the deep convolutional neural networks, the idea of a generalizability of CLE towards disentangling the visualizations with cognitive load has received a boost. The CNNs effectively learn features across visualizations that could improve generalizability of the CLE model. This could possibly provide a generic model for evaluating Viz UIs. Towards this we use *character*, *spatial pattern* or *position*-based, *bar graph* and *pie chart* visualizations for invoke various levels of mental workload through a traditional n-back task [54, 7, 34, 80]. We compared our work with the state-of-the-art proximal SVM(pSVM) [80] method used for within Viz CLE. We infer that though pSVM performs better for same(within) Viz CLE, CNN outperforms pSVM for across Viz CLE even though the results are relatively lower than the within Viz CLE.

A.1 Related Work

Traditional methods for evaluating Viz UIs have been user studies. The opinions regarding the visualization techniques could be answered with focused studies [39]. However, they are inadequate for evaluating them, partly because VA tools need to provide the human expert insights by presenting the same data from multiple perspectives, and subjective human opinions are more indicative of holistic experience, and cannot effectively capture the UI's effectiveness at supporting data exploration. [60] supports the use of quantitative measures for evaluation of Viz UIs, and also how physiological signals could be used for detection of critical moments for insight discovery. [21] also supports the use of more cognitive theories towards evaluating them over traditional user studies.

A.1.1 Implicit signals for workload evaluation

Previous works [35] suggest that cognitive load changes are reflected in the spectral power of EEG namely in the frequency bands of α and θ. This is validated by [4] where Emotiv heaset is used to collect the EEG data for cognitive load invoking task. To summarize the effect of cognitive load on α power, the alpha power decreases as the workload increases. The deoxygenated blood level monitored through fNIRS is also one of the preferred methods for workload evaluation specifically non-invasively. The fNIRS devices further are light-weight and do not require any fluids to provide conductivity between the scalp and the device(which is the case with EEG). This has inspired [54] to use fNIRS to evaluate the

workload for bar and pie-chart visualizations. Despite such advantages, fNIRS lacks temporal resolution over EEG. Hence, EEG is the preferred imaging method for such dynamic scalp electrical activity.

As our work uses visual stimuli for workload invoking, a traditional *n*-back task is our approach to invoke load in Visual Working Memory(VWM). The VWM is responsible for storage and retrieval of information related to a task. The *n*-back requires the users to continuously update the VWM and respond. To avoid any confounding factors with the workload, we make the motor activity uniform for all visualizations. Previous works [80] use morphological and spectral features for this purpose, proximal SVM(pSVM) was then adopted to evaluate the classification of the load. The feature learning capability of Deep CNNs have inspired us to use for the same, however, unlike traditional CNNs using temporal features or vectorized inputs, we use the multi-spectral images preserving the topography of the channels for CLE.

A.1.2 Analysis of related work

Recent VA works reveal the community prefers quantitative evaluation of Viz UI, especially using the brain imaging methods such as EEG and fNIRS. While several works use these methods to validate, only few works have utilized them for predicting the cognitive load. The prediction works assume that the underlying perceptual differences invoked by different visualizations are minimal when compared to the task difficulty.

[54] has validated the existence of deoxygenated hemoglobin level changes in blood based on the cognitive load for bar graph and pie chart visualizations however, they have not utilized such differences to perform a CLE using the fNIRS data. Later works [80] have attempted to evaluate the load based on the char and spatial visualizations. They assumed that the different visualizations induce minimal confounds when compared to the workload induced by the task as described above.While [34] suggests feature selection towards CLE, they do not perform it across visualizations.

Inspite of having a wide range of works employing different modalities towards validating and predicting CLE, the question of determining the best visualization given the task type and difficulty remains unanswered. A part of the reason could be the large intra and inter subject variations in the cognitive activity [7]. However, a CLE model that works well for a particular set of users for a given task and visualization type would not have any validity ecologically [34], and also could severely restrict the generalization of CLE.

Contribution: We exploit the implicit EEG signals for CLE based on the similar *n*-back task for different visualizations. Hence also perform an inter-subject CLE for generalization across the population.

A.2 Materials and Methods

We describe in detail, all the specifics of our data collection protocol and method.

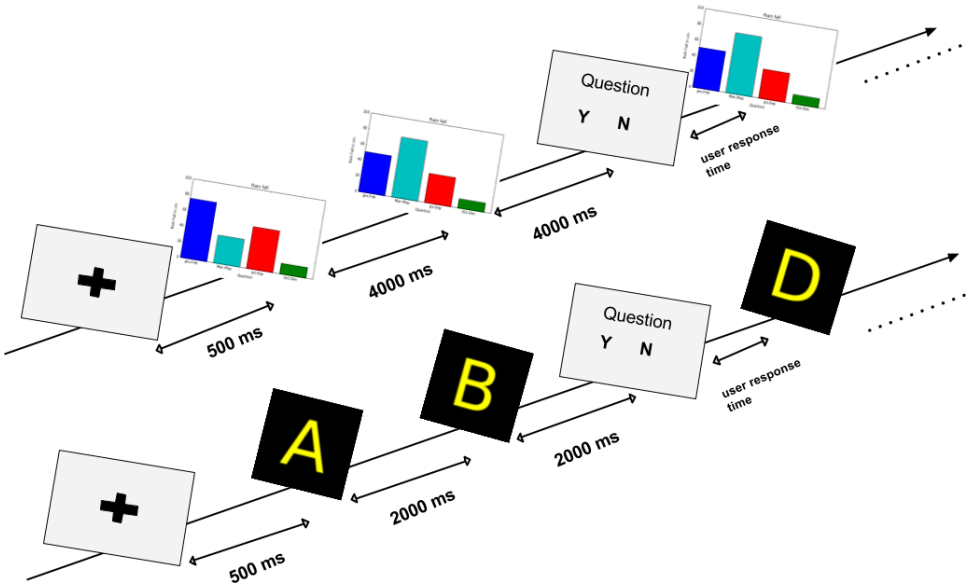

Figure A.2 Sample protocol settings for 1-back task.

A.2.1 Experimental Design

A.2.1.1 Stimuli

We employed four visualization types, namely, character (*char*), spatial pattern or position-based (*pos*), bar graph (*bar*) and pie chart (*pie*) as stimuli for our study. Sample images are shown as Figure A.1. These visualizations have been used in prior studies (*e.g.*, [7, 54, 80]), and are designed to utilize the visual sensory pathway and working memory.

Response	Predictor	Viz Type	N-back	Viz Type*N-back	Error	Total
	df	3	3	9	304	316
RT	**F**	3.96	27.17	4.33		
	Sig	$p < 0.001$	$p < 0.000001$	$p < 0.000001$		
RA	**F**	74.04	24.75	1.62		
	Sig	$p < 0.000001$	$p < 0.000001$	n.s.		

Table A.1 ANOVA summary for RTs and RAs. df and Sig respectively denote degrees of freedom and significance level.

The visualizations shown in A.1 were some of the samples of the ones that are used in the experiment. The bar and pie chart visualizations were generated from naturally available rainfall which is obtained from an open source website. The pos stimuli is a 3x3 grid with one of them highlighted to convey spatial information. 20 individuals(9 female, 4.4$\pm$2.1) have voluntarily participated in the experiment. None of them had any kinds of visual impairments. They were financially compensated for their time and effort. All the participants agreed and were monitored by the institute ethics committee.

A.2.1.2 Protocol

Having adopted a typical *n*-back test, our experiment is split over two sessions for equal duration. The entire data collection took about 1 hour and the participants were given the liberty of opting for a short break to avoid fatigue. This also helps us in re-calibrating the sensors and acquire reliable user data. The users are required to express there answer with mouse button press, this further minimizes the muscle movement activity. All stimuli were screen centered and adjusted to meet the participants requirements for easy interpretation.

The entire experiment is sub-divided into 24 blocks for equal number of stimuli. Each block corresponds to a particular visualization category with a particular *n*-back condition(varies from 0–3 backs). Instructions were presented before each block to convey the *n*-back type and the visualization type. All participants were required to undertake a practice session before beginning the actual experiment. No sensor information was collected during the practice session. The number of low load blocks(0, 1-back) were balanced with the number of high load stimuli(2, 3-back). Similarly, the pos/char visualization is always followed by a bar/pie respectively an vice-versa. Thus the experiment was highly balanced in terms of the user experience and no external factors confounded with the experiment outcomes. Each block comprises of 12 slides of same visualization. The participants were required to answer a simple 'yes' or 'no' question for stimuli presented based on the instructions corresponding to the block. For a 0-back task, they were required to compare the current stimuli with a standard mentioned along with the instructions for the block. However, for a n>0 task, they were required to compare the current stimuli with the stimuli presented n slides ago. This requires the participants to continuously store, retrieve and refresh their working memory thus invoking different levels of cognitive load. Since the experiment was designed keeping the user experience in mind, we encountered imbalance in the number of user responses and thus the data for classification.

Figure A.2 shows the timeline from our experiment design. Each block begins with an instruction slide describing the *n*-back type and the visualization. If its a 0-back case, we also specify the condition to be compared with. A fixation cross preceeds each stimuli for 0.5sec. The fixation cross usually is employed to clear any residues of previous cognitive process and help the users focus. It is also used to calibrate the device for resting state activity. This is followed by the stimuli. Based on the visualization type, the stimuli duration varies as 2sec(char and pos) or 4sec(bar and pie). The increased time for bar and pie are due to the fact that they require the participants to perform mental arithmetic to infer the quantity of interest. Additional care was taken such that the quantity of interest differs between any stimuli of a given visualization by a significant amount and not marginal.

Each user was given a maximum time limit of 10sec to respond. In case of non-response, the trial would be marked a invalid and not used for any data analysis. This helped us in particularly maintain the total experiment time, the 10 sec parameter was found be acceptable with participants from our initial investigation. The user implicit activity was recorded using 14-channel Emotiv Epoc device, a low-cost consumer grade EEG device. To summarize, our experiment was designed to investigate 4 cognitive load settings of 4 different visualizations.

A.2.1.3 Hypotheses

We explain our initial hypothesis that inspired us in designing the experiment.

1. ***Bar and pie visualizations are challenging over char and pos:*** While char and pos stimuli involve intuitively east tasks, the bar and pie visualizations require the user to perform mental calculations to estimate the measure of interest in addition to the intrinsic difficulty because of the visualization.

2. ***Higher n-back leads to decrease in user performance:*** Irrespective of the visualization, the performance in terms of response times and accuracy decreases with task difficulty, in our case the *n*-back . This is based on the assumption that difficult tasks consume more mental resources and effort.

3. ***Similar cognitive process exists between the stimuli*** Leaving the protocol similarities, the bar-pie and char-pos stimuli are lot similar than other combinations. This led us to believe that these pairs could achieve better generalization because of the stimuli similarities among them.

A.3 User Response Analysis

The sensitivity of users' in terms of the explicit responses are summarized as response times(RTs) and response accuracy(RAs):

A.3.1 Response times

Figure A.3(left) shows RTs for the four visualizations for 0-3 *n*-back tasks. The responses for char and pos in 0 and 1 back cases remain same with marginal differences for the 2 and 3 back. The mean RT for char, pos, bar and pos are 0.72±0.15s, 0.88±0.14s, 1.17±0.22s, 1.26±0.23s respectively. Consistent with out hypothesis-2, the RT increases with increasing *n*-back as greater mental resources are consumed for higher *n*-back type. Table A.1 summarises the two-way ANOVA results on RTs along with the effects of visualization, *n*-back type and their interactions. The mean RTs are significantly different for pos and pie and marginally different for char and pie. We also observe a marginally lower RT for bar over pie.

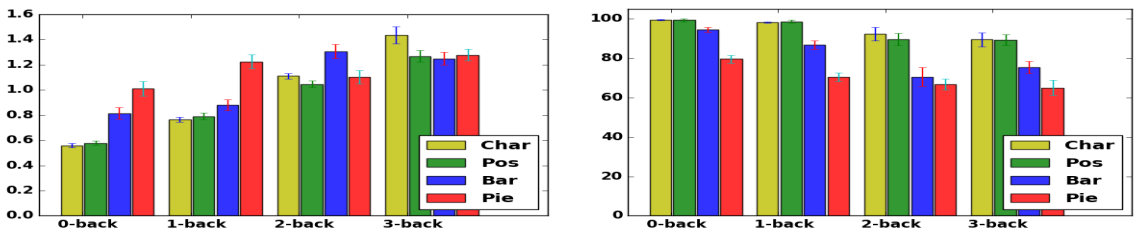

Figure A.3 (left) Response times(RTs in seconds) and (right) response accuracy(RAs) for various stimuli.

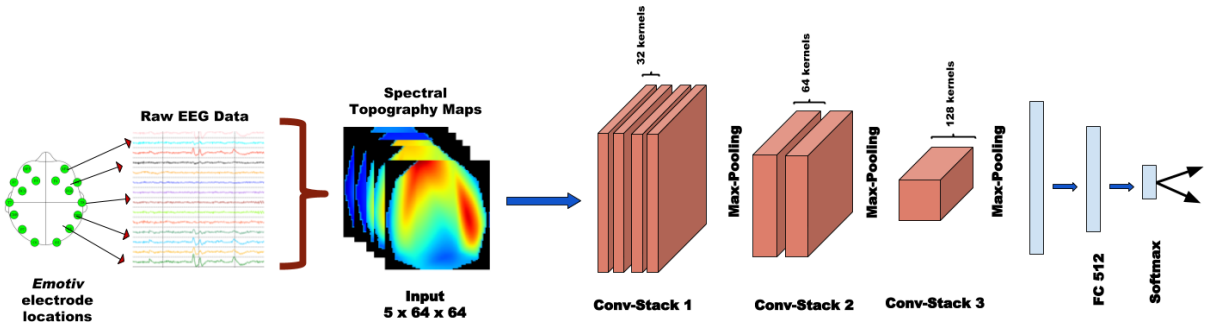

Figure A.4 Deep learning architecture.

A.3.2 Response accuracies

Similar to our observation of RTs, we observe a decreasing performance with increasing n-back as shown in Figure A.3(right). The pos, char for 0 and 1 back have achieved a ceiling performance. Interestingly, $< 80\%$ accuracy was observed for the rest. In addition, the RA decreases for high workload case of 2 and 3-back with the accuracy following the trend $char > pos > bar > pie$. Table A.1 reveals the ANOVA results. The RA is significantly different for bar and pie despite our hypothesis-3. The RAs for char and pos is significantly higher when compared to the rest of the visualizations.

A.3.3 Explicit responses

The RTs and RAs results from the above sections are consistent with our hypothesis 1 and 2. As the n-back increases, the user performance decreases. We also observe that there exists a similarity between char and pos and also between bar and pos visualization in terms of explicit user responses of RTs and RAs. The difficulty increases with n-back and our results are consistent with [15],e an estimation task, but differ with respect to the work of [54] who found that the bar and pie are user specific but not generic.

In the rest of the work, we see if the changes in explicit responses are also observed in implicit signals. We frame a CLE task using CNNs across visualizations and compare the same with the previous state-of-art pSVM.

A.4 EEG analysis for Cognitive load

A.4.1 Preprocessing

Recent works [80, 11, 4] have demonstrated the effectiveness of Emotiv EEG heaset for cognitive workload capturing. We extracted epochs for each trial (2s of stimulus-plus-pre response time @ 128 Hz sampling rate), and the 64 leading samples were used to remove DC offset. This was followed by (a) EEG band-limiting to within 0.1–45 Hz, (b) Removal of noisy epochs via visual inspection, and (c)

Independent Component Analysis (ICA) to remove artifacts relating to eye-blinks, and eye and muscle movements. Muscle movement artifacts in EEG are mainly concentrated in the 40–100 Hz band, and are removed upon band-limiting and via inspection of ICA components. Finally, a 7168 dimensional (14×4×128) EEG feature vector was generated from concatenation of the 14 EEG channels over 2s of pre-response event.

Method	Raw EEG Input	Extracted Features	What is classified?	What is output?
Deep CNN	14×2×128 samples from 2 sec pre-response epochs.	14× 5 electrode-wise spectral measurements projected to form 5×64×64 spectral topography maps via interpolation and projection.	512 dimensional FC layer output vector	One of two input labels as predicted by the softmax layer.
pSVM		14×47 electrode-wise spectral, statistical, morphological and entropy features.	Top 10 among 658 features as determined by feature selection.	One of two input labels as predicted by the pSVM classifier.

Table A.2 Summary of the features and outputs for CNN and pSVM methods.

A.4.2 Deep learning for Cognitive Load Estimation

One of the recent works on CLE [7] using convolutional neural networks is inspired from image classification works. The philosophy is centered about the automatic feature learning for a specific task. This has alleviated the problem of using handcrafted features towards the machine learning task.

Unlike traditional methods [80] that use statistical, morphological features, [7] uses mulitspectral topograhic images generated from multiple EEG spectral bands namely, $\alpha(8-13Hz)$, $\beta(13-30Hz)$, $\gamma(30-45Hz)$, $\delta(0.5-4Hz)a$ and $\theta(4-7Hz)$ bands. Some works use vectorized EEG data as features, however, these neglect the spatial and temporal relations that exist among the EEG data. The CNN architecture is shown in figure A.4. The spectral components are generated with Fast Fourier Transform(FFT) over the two second EEG epochs. The signal power in each of these components is used to generate the spectral maps that are similar to the RGB channels of a traditional image,

We have generated an image like input to the CNN with 5 channels(each corresponding to one of the frequency bands). The FFT signal power is projected onto a 2D surface using Azimuthal Equidistant projection(AEP) to preserve the spatial 3-D relation between the channels. The inputs were generated at various sizes–32, 64 and 128. We observe that our analysis was best achieved with 64 image dimensions. Similar to VGG architecture [66], the CNN avoids the normalization layers and uses max-pooling with convolutions. The number of kernels follow an increasing trend of 32, 64 and 128(increasing with depth) to learn more high level abstract representations. The convolutional layers are followed by the FC layers and a softmax layer. Despite large number of design similarities between VGG [66], this is one of the first deep learning works for CLE.

A.4.3 pSVM based Cognitive Load Estimation

[80] has assessed the workload using Emotiv EEG headset. They have performed extensive pre-processing with ICA based algorithm(which was not the same algorithm used in our preprocessing) to

Type	0-back	1-back	2-back	3-back	Total
Char	356	199	364	167	1086
Pos	377	175	165	376	1093
Bar	203	391	194	390	1178
Pie	193	401	393	192	1179
Total	1129	1166	1116	1125	**4536**

Table A.3 Data distribution split with respect to the back and the visualization type.

generate 47 features per electrode. The features can be broadly classified as spectral, statistical, morphological and wavelet based. The 14x47 electrode specific features were concatenated to form a feature vectors. These were used for CLE using proximal SVM algorithm, a faster alternative to standard SVM. Table A.2 summarizes the results of CNN and SVM approaches.

A.4.4 Experiments and Results

An ideal CLE task would generate the cognitive load score for an implicit response corresponding to a given visualization. This would be regression task, however, most recent works [11, 7, 80] focus on classification of cognitive load into high and low categories. Despite being a more appropriate task(over regression), generalization of CLE has not been achieved with a coarse classification task. As a first step towards generalization, we perform classification between each of the different workload combinations for each of the visualizations and cross visualization combinations. Though [7] have used multiframe approach(representing the epoch data of multiple windows of equal size as input to the CNN) towards classification, we have evaluated the generalization only for singleframe(entire epoch data represented as one input to the CNN) approach.

We hypothesized to observe some intuitive outcomes in our results. To elaborate, we investigated if the performance of a 0 vs 3-back classifier would be better than a 0 vs 1-back classifier. We have used user independent data for testing(no isolation of users for validation/testing was used to) to learn the task specific cognitive load features and ignore the user dependencies. We evaluated the classifier performance with accuracy and F1-score metrics considering the data imbalance(as higher correct responses were observed for low cognitive load scenario when compared to high load, and also bar and pie are relatively difficult visualizations). Table A.3 illustrates the data imbalance after preprocessing, it also shows the data distribution used for classification task.

A.4.4.1 Classification settings

From table- A.3 we observe that fewer than a couple of hundred are available for classification. We explain the classifier settings used for training as follows. We used 5 repetitions of 10 fold cross-validation of epoch data for classification. The classifier performance over independent training and test sets remain similar. We also experienced a reasonable standard deviation in accuracy and F1 scores thus reaffirming a stable classifier. With relatively fewer data for deep learning training, our primary concern

		char		pos		bar		pie	
		Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
char	**0 vs 1**	0.63	0.63	0.53	0.68	0.37	0.52	0.48	0.61
	0 vs 2	0.66	0.66	0.59	0.73	0.56	**0.71**	0.47	0.59
	0 vs 3	0.71	0.69	0.59	0.68	0.46	0.54	0.62	0.63
	1 vs 2	0.71	0.71	0.64	**0.74**	0.41	0.54	0.46	0.59
	1 vs 3	0.65	0.64	0.43	0.49	0.54	0.62	0.63	**0.68**
	2 vs 3	0.78	**0.80**	0.46	0.58	0.27	0.35	0.54	0.63
pos	**0 vs 1**	0.61	**0.68**	0.67	0.64	0.43	0.52	0.38	0.44
	0 vs 2	0.57	0.65	0.76	**0.72**	0.63	**0.69**	0.43	0.48
	0 vs 3	0.52	0.62	0.60	0.60	0.54	**0.69**	0.64	**0.70**
	1 vs 2	0.54	0.65	0.64	0.64	0.50	0.61	0.44	0.54
	1 vs 3	0.50	0.58	0.68	0.68	0.48	0.60	0.43	0.53
	2 vs 3	0.47	0.52	0.63	0.60	0.60	0.67	0.40	0.47
bar	**0 vs 1**	0.39	0.53	0.47	0.63	0.63	0.63	0.54	0.66
	0 vs 2	0.52	0.58	0.67	**0.75**	0.65	0.61	0.42	0.48
	0 vs 3	0.48	0.58	0.60	0.74	0.83	**0.83**	0.71	**0.77**
	1 vs 2	0.35	0.41	0.43	0.46	0.73	0.69	0.45	0.53
	1 vs 3	0.56	**0.64**	0.44	0.54	0.79	0.78	0.62	0.70
	2 vs 3	0.26	0.37	0.51	0.65	0.66	0.65	0.51	0.63
pie	**0 vs 1**	0.40	0.48	0.43	0.52	0.57	**0.66**	0.59	0.54
	0 vs 2	0.51	**0.63**	0.48	0.62	0.47	0.62	0.60	0.58
	0 vs 3	0.56	0.61	0.58	**0.67**	0.49	0.58	0.73	0.68
	1 vs 2	0.46	0.56	0.52	0.62	0.50	0.62	0.60	0.59
	1 vs 3	0.56	0.56	0.39	0.39	0.55	0.57	0.79	**0.74**
	2 vs 3	0.58	0.60	0.37	0.43	0.35	0.39	0.73	0.68

Table A.4 Deep CNN based cognitive load estimation. Rows represent the training visualization while columns represent the test.

		char		pos		bar		pie	
		Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**
char	**0 vs 1**	0.74	0.72	0.52	0.54	0.27	0.27	0.47	0.47
	0 vs 2	0.66	0.65	0.54	0.56	0.48	0.51	0.52	0.52
	0 vs 3	0.90	0.83	0.42	0.48	0.43	0.48	0.41	0.51
	1 vs 2	0.84	0.82	0.71	**0.71**	0.34	0.38	0.45	0.45
	1 vs 3	0.79	0.77	0.41	0.44	0.55	**0.62**	0.50	**0.64**
	2 vs 3	0.90	**0.86**	0.44	0.46	0.24	0.25	0.40	0.57
pos	**0 vs 1**	0.49	0.53	0.79	0.76	0.52	0.53	0.28	0.28
	0 vs 2	0.56	0.59	0.81	0.74	0.49	**0.65**	0.45	0.47
	0 vs 3	0.40	0.47	0.72	0.70	0.43	0.44	0.41	0.42
	1 vs 2	0.64	**0.66**	0.83	**0.80**	0.53	0.57	0.50	**0.50**
	1 vs 3	0.50	0.50	0.79	0.77	0.49	0.49	0.37	0.40
	2 vs 3	0.46	0.49	0.82	0.79	0.43	0.45	0.31	0.34
bar	**0 vs 1**	0.29	0.29	0.59	**0.60**	0.72	0.70	0.45	0.47
	0 vs 2	0.48	0.52	0.50	0.59	0.72	0.64	0.42	0.42
	0 vs 3	0.47	0.53	0.49	0.53	0.82	**0.81**	0.54	**0.59**
	1 vs 2	0.39	0.39	0.50	0.57	0.78	0.72	0.49	0.51
	1 vs 3	0.55	**0.58**	0.54	0.55	0.69	0.67	0.44	0.50
	2 vs 3	0.26	0.26	0.33	0.40	0.76	0.72	0.43	0.44
pie	**0 vs 1**	0.45	0.45	0.23	0.23	0.48	0.50	0.75	0.73
	0 vs 2	0.55	0.56	0.40	0.41	0.41	0.41	0.78	0.77
	0 vs 3	0.44	**0.62**	0.50	**0.57**	0.55	**0.59**	0.86	0.70
	1 vs 2	0.49	0.50	0.49	0.50	0.40	0.43	0.63	0.62
	1 vs 3	0.47	0.61	0.44	0.46	0.45	0.57	0.79	0.65
	2 vs 3	0.42	0.58	0.28	0.28	0.48	0.49	0.84	**0.78**

Table A.5 pSVM based cognitive load estimation. Rows represent the training visualization while columns represent the test

was overfitting. However, the above points confirm that the training was stable and no occurrence of overfitting during the training process. We believe that this could be partly because of fine tuning of [7] network over training the model from scratch with our limited and imbalanced data.

A.4.4.2 Comparison of results

CLE for both within and cross-visualization stimuli using CNN and pSVM are presented as tables A.4 and A.5 respectively. The metrics tabulated are accuracy and F1 scores. Having observed the imbalance in classification, F1 score is more appropriate metric to draw inferences. Within visualization performance is tabulated as diagonal elements of the 4 x 4 elements of each table. The highest F1 score for each of the within and cross visualization are highlighted in bold. We observe that the highest performance is observed for high(2, 3-back) vs low(0, 1-back) cognitive load case. This trend is observed in 13 out of 16 cases of within and cross visualization cases. This validates our hypothesis-2. pSVM outperforms CNN for within visualization CLE.

We denote the mean F1 score over the 5 repetitions of 10 fold cross-validation with $\overline{\text{F1}}$. In terms of within visualization performance, the char based pSVM performs the best(within $\overline{\text{F1}}$=0.86) with pos(within $\overline{\text{F1}}$=0.80) with $\overline{\text{F1}}$=0.71. We observe that pos(within $\overline{\text{F1}}$=0.80) based visualization performs best with char with $\overline{\text{F1}}$=0.66, however, this is closely associated with bar visualization performance of $\overline{\text{F1}}$=0.65. The bar visualization(within $\overline{\text{F1}}$=0.81) performs better than chance level with all the other visualizations, however char performs relatively worse over the rest under all the cases. The pie visualization(within $\overline{\text{F1}}$=0.78) performs similar to bar in terms of cross-visualization testing, however, it generalizes relatively well over the char when compared to the same model when trained with bar visualizations. Over all the visualizations, the cross visualization performance is drastically lower than the within classification performance.

Contrast to the pSVM results described above, we observe better cross-visualization performance with the CNN architecture. The char based visualization(within $\overline{\text{F1}}$=0.80) performs best with pos $\overline{\text{F1}}$=0.74. In addition, we observe better than chance level performance over the rest of the visualizations. Surprisingly, the pos visualization (within $\overline{\text{F1}}$=0.72) performs best with pie $\overline{\text{F1}}$=0.70. We also notice a close performance of within and cross-visualization classification. Consistent with our hypothesis, the bar(within $\overline{\text{F1}}$=0.83) performs best with pie with $\overline{\text{F1}}$=0.77. We observe significantly low performance when trained on bar and tested on char. The pie visualization(within $\overline{\text{F1}}$=0.66) better generalizes over bar $\overline{\text{F1}}$=0.66 and pos $\overline{\text{F1}}$=0.67. Overall, the CNN model better generalizes when compared to the traditional pSVM method based on the hand-tuned features.

A.5 Conclusions and Future works

Examining the utility of user interfaces through cognitive sensing mechanisms has been an interest to the HCI community in recent times. Though certain works have examined the utility of such methods

using low cost devices such as EEG and other sensing methods such as fNIRS, as [27, 34] has described, the generalizability of such models over different visualizations or interfaces was never investigated. Our work is first of its kind to evaluate the generalizability of cognitive load for visual interfaces(to the best of our knowledge). We use the low-cost EEG device to acquire implicit signals to evaluate the workload induced various visualizations. We overcome the inter-subject dependency of the implicit signals by using data irrespective of the users(no data leak) for training and testing. This work could be of its own importance in alleviating the need for user based surveys towards interface utility evaluation, a traditional method adopted by the industry.

The *n*-back task used in our study is standard workload inducing test. The experiment design requires the users' to perform a first order comparison(equal/same or not task) for char and pos visualizations but second order comparisons(greater than or less than task) for bar and pie visualizations. The explicit response analysis reveals that the user response deteriorates with the task difficult and bar and pie visualizations are relatively difficult. This could be because these involve the users performing mental arithmetics to answer the simple yes/no question. This is different from the previous works that require size ratio comparison in [15] or size comparison in [54] with as margin of 10%.

Having employed the two current state-of-the-art works(CNN [7] and pSVM [80]) for workload measurement, our observations are tow fold–(a) the best performance was noticed for high vs low cognitive load conditions(wither of 0 vs 2-back or 0 vs 3-back or 1 vs 2-back) (b) pSVM performs best for within visualization classification but CNNs perform better in terms of cross visualization based classification. The prominent causes for CNNs outperforming the SVM is due to the ability to learn task-agnostic feature representations; also the input to the CNNs is the multispectral topographical maps preserve spectral and spatial relations of EEG while the other method used vectorized statistical features .

The hypotheses specified in the previous sections were validated from our machine learning results. Though the pSVM performance in terms of cross visualization was relatively poor, we found that there could exist a common cognitive activity between similar tasks/visualizations namely, the char-pos and bar-pie pairs. The CNN results reemphasize that through a significantly similar performance among multiple visualizations. In specific, we observe that the model trained on pos performs equally well on the other visualizations not just the char(as hypothesized). Similar was the case with bar and pie visualizations. Another unique aspect of our work–ecological validity of low-cost EEG device makes large scale user study a relatively easier [63] task when compared to lab-grade devices.

Despite finding a reasonable level of agreement between the hypotheses and our results, we believe that some of the questions described here would warrant attention as parts of future problems in this direction. The low-cost deviced helping us with simpler experiment design, acquisition comes at the cost of extensive preprocessing pipeline. While most parts of the preprocessing are straight forward, some parts of them such as visual rejection would require a trained eye. Hence this would hinder a possible chance of building an end-to-end online system. The transfer learning used over the VGG net like architecture further leaves scope for latest architectures such as ResNet [28] or training approaches such as stochastic depth [30]. We also believe that attention along spatial dimension(with spatio-temporal

data unlike multispectral images as input) would be a more reasonable mechanism towards capturing dependencies among channels unlike pooling approaches which were meant to capture neighbourhood relations in data such as images. Further for fine-grained classification tasks such as ours, models such as siamese [36] and triplet networks has received the interest of the community in recent times. While most of the above said problems require large quantities of data thus requiring large scale and crowdsourced user studies, the penetration of EEG devices would be a hurdle to overcome in this direction.

Bibliography

[1] Solving problems with visual analytics. *Procedia Computer Science*, 7:117 – 120, 2011.

[2] Biometric security poses huge privacy risks, 2013.

[3] M. Abadi, R. Subramanian, S. Kia, P. Avesani, I. Patras, and N. Sebe. DECAF: MEG-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, 2015.

[4] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva. A user study of visualization effectiveness using eeg and cognitive load. In *Computer Graphics Forum*, volume 30, pages 791–800. Wiley Online Library, 2011.

[5] H. Aviezer, R. R. Hassin, J. Ryan, C. Grady, J. Susskind, A. Anderson, M. Moscovitch, and S. Bentin. Angry, disgusted, or afraid? *Psychological Science*, 19(7):724–732, 2008.

[6] T. Baltrušaitis, P. Robinson, and L.-P. Morency. Openface: an open source facial behavior analysis toolkit. In *IEEE Winter Conference on Applications of Computer Vision*, 2016.

[7] 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.

[8] J. N. 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*, 37(11):2049–2058, 1979.

[9] M. Bilalpur, S. M. Kia, M. Chawla, T.-S. Chua, and R. Subramanian. Gender and emotion recognition with implicit user signals. In *International Conference on Multimodal Interaction*, 2017.

[10] M. Bilalpur, S. M. Kia, T.-S. Chua, and R. Subramanian. Discovering gender differences in facial emotion recognition via implicit behavioral cues. *arXiv preprint arXiv:1708.08729*, 2017.

[11] A.-M. Brouwer, M. A. Hogervorst, J. B. Van Erp, T. Heffelaar, P. H. Zimmerman, and R. Oostenveld. Estimating workload using eeg spectral power and erps in the n-back task. *Journal of neural engineering*, 9(4):045008, 2012.

[12] P. Campisi and D. La Rocca. Brain waves for automatic biometric-based user recognition. *IEEE Transactions on Information Forensics and Security*, 9(5):782–800, 2014.

[13] P. Chandler and J. Sweller. Cognitive load theory and the format of instruction. In L. E. Associates, editor, *Cognition and Instruction*, pages 292–332. Taylor & Francis, 1991.

[14] F. Chollet et al. Keras. 2015.

[15] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):pp. 531–554, 1984.

[16] M. Czerwinski, D. S. Tan, and G. G. Robertson. Women take a wider view. In *Conference on Human Factors in Computing Systems (CHI)*, pages 195–202, 2002.

[17] E. Dalmaijer. Is the low-cost eyetribe eye tracker any good for research? Technical report, PeerJ PrePrints, 2014.

[18] Y. Deng, L. Chang, M. Yang, M. Huo, and R. Zhou. Gender differences in emotional response: Inconsistency between experience and expressivity. *PloS one*, 11(6):e0158666, 2016.

[19] N. L. Etcoff and J. J. Magee. Categorical perception of facial expressions. *Cognition*, 44(3):227–240, 1992.

[20] Y. Freund and R. E. Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory*, pages 23–37. Springer, 1995.

[21] F. L. Greitzer, C. F. Noonan, and L. Franklin. *Cognitive Foundations for Visual Analytics*. 2011.

[22] X. Gui, C. Chuansheng, L. Zhong-Lin, and D. Qi. Brain imaging techniques and their applications in decision-making research. *Xin li xue bao. Acta psychologica Sinica*, 42(1):120, 2010.

[23] J. A. Hall and D. Matsumoto. Gender differences in judgments of multiple emotions from facial expressions. *Emotion*, 4(2):201–206, 2004.

[24] J. K. Hall, S. B. Hutton, and M. J. Morgan. Sex differences in scanning faces: does attention to the eyes explain female superiority in facial expression recognition? *Cognition and Emotion*, 24(4):629–637, 2010.

[25] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Transactions on multimedia*, 7(1):143–154, 2005.

[26] T. Hastie, S. Rosset, J. Zhu, and H. Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009.

[27] J.-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523–534, 2006.

[28] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, 2015.

[29] B. D. Homer, E. O. Hayward, J. Frye, and J. L. Plass. Gender and player characteristics in video game play of preadolescents. *Computers in Human Behavior*, 28(5):1782–1789, 2012.

[30] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger. Deep networks with stochastic depth. In *European Conference on Computer Vision*, pages 646–661. Springer, 2016.

[31] W. Huang. Using eye tracking to investigate graph layout effects. In *Int'l Asia-Pacific Symposium on Visualization*, pages 97–100, 2007.

[32] H. Joho, J. Staiano, N. Sebe, and J. M. Jose. Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents. *Multimedia Tools and Applications*, 51(2):505–523, 2011.

[33] H. Katti, R. Subramanian, M. Kankanhalli, N. Sebe, T.-S. Chua, and K. R. Ramakrishnan. Making computers look the way we look: exploiting visual attention for image understanding. In *ACM Int'l conference on Multimedia*, pages 667–670, 2010.

[34] Y. Ke, H. Qi, F. He, S. Liu, X. Zhao, P. Zhou, L. Zhang, and D. Ming. An eeg-based mental workload estimator trained on working memory task can work well under simulated multi-attribute task. *Frontiers in Human Neuroscience*, 8:703, 2014.

[35] W. Klimesch. EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain research. Brain research reviews*, 29(2-3):169–195, 1999.

[36] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *Deep Learning Workshop (ICML)*, 2015.

[37] S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdan, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A Database for Emotion Analysis Using Physiological Signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.

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

[39] R. Kosara, C. G. Healey, V. Interrante, D. H. Laidlaw, and C. Ware. Thoughts on User Studies: Why, How, and When. *Computer Graphics and Applications*, 23(4):20–25, 2003.

[40] V. Krassanakis, V. Filippakopoulou, and B. Nakos. Eyemmv toolbox: An eye movement post-analysis tool based on a two-step spatial dispersion threshold for fixation identification. *Journal of Eye Movement Research*, 7(1), 2014.

[41] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg. Presentation and validation of the radboud faces database. *Cognition and Emotion*, 24(8):1377–1388, 2010.

[42] C. M. Lee and S. S. Narayanan. Toward detecting emotions in spoken dialogs. *IEEE transactions on speech and audio processing*, 13(2):293–303, 2005.

[43] M. Li, K. J. Han, and S. Narayanan. Automatic speaker age and gender recognition using acoustic and prosodic level information fusion. *Computer Speech and Language*, 27(1):151–167, 2013.

[44] X. Li, P. Zhang, D. Song, G. Yu, Y. Hou, and B. Hu. Eeg based emotion identification using unsupervised deep feature learning. 2015.

[45] W. L. Lim, O. Sourina, and L. Wang. Mind-an eeg neurofeedback multitasking game. In *Cyberworlds (CW), 2015 International Conference on*, pages 169–172. IEEE, 2015.

[46] C. Lithari, C. A. Frantzidis, C. Papadelis, A. B. Vivas, M. A. Klados, C. Kourtidou-Papadeli, C. Pappas, A. A. Ioannides, and P. D. Bamidis. Are females more responsive to emotional stimuli? a neurophysiological study across arousal and valence dimensions. *Brain Topography*, 23(1):27–40, 2010.

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

[48] E. B. McClure, C. S. Monk, E. E. Nelson, E. Zarahn, E. Leibenluft, R. M. Bilder, D. S. Charney, M. Ernst, and D. S. Pine. A developmental examination of gender differences in brain engagement during evaluation of threat. *Biological psychiatry*, 55:1047–55, 2004 Jun 1 2004.

[49] B. Montagne, R. P. C. Kessels, E. Frigerio, E. H. F. de Haan, and D. I. Perrett. Sex differences in the perception of affective facial expressions: do men really lack emotional sensitivity? *Cognitive Processing*, 6(2):136–141, 2005.

[50] C. Muhl, B. Allison, A. Nijholt, and G. Chanel. A survey of affective brain computer interfaces: principles, state-of-the-art, and challenges. *Brain-Computer Interfaces*, 1(2):66–84, 2014.

[51] C. B. Ng, Y. H. Tay, and B.-M. Goi. Vision-based human gender recognition: A survey. *CoRR*, abs/1204.1611, 2012.

[52] P. Nguyen, D. Tran, X. Huang, and W. Ma. Age and gender classification using eeg paralinguistic features. In *Neural Engineering (NER), 2013 6th International IEEE/EMBS Conference on*, pages 1295–1298. IEEE, 2013.

[53] D. Passig and H. Levin. The interaction between gender, age, and multimedia interface design. *Education and Information Technologies*, 6(4):241–250, 2001.

[54] E. M. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 473–482. ACM, 2013.

[55] R. W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

[56] H. R.-Tavakoli, A. Atyabi, A. Rantanen, S. J. Laukka, S. Nefti-Meziani, and J. Heikkil. Predicting the valence of a scene from observers eye movements. *PLoS ONE*, 10(9):1–19, 2015.

[57] Q. Rahman, G. D. Wilson, and S. Abrahams. Sex, sexual orientation, and identification of positive and negative facial affect. *Brain and Cognition*, 54(3):179–185, 2004.

[58] R. P. Rao. *Brain-computer interfacing: an introduction*. Cambridge University Press, 2013.

[59] M. Raschke, T. Blascheck, M. Richter, T. Agapkin, and T. Ertl. Visual analysis of perceptual and cognitive processes. In *Information Visualization Theory and Applications*, pages 284–291, 2014.

[60] N. H. Riche. Beyond system logging: human logging for evaluating information visualization. In *BELIV*, 2010.

[61] M. W. Schurgin, J. Nelson, S. Iida, H. Ohira, J. Y. Chiao, and S. L. Franconeri. Eye movements during emotion recognition in faces. *Journal of Vision*, 14(13):1–16, 2014.

[62] J. D. Schwark, I. Dolgov, D. Hor, and W. Graves. Gender and personality trait measures impact degree of affect change in a hedonic computing paradigm. *International Journal of Human Computer Interaction*, 29(5):327–337, 2013.

[63] P. Shenoy and D. Tan. Human-aided computing: Utilizing implicit human processing to classify images. In *SIGCHI Conference on Human Factors in Computing Systems*, 2008.

[64] A. Shukla, S. S. Gullapuram, H. Katti, K. Yadati, M. Kankanhalli, and R. Subramanian. Affect recognition in ads with application to computational advertising. *arXiv preprint arXiv:1709.01683*, 2017.

[65] P. Silapachote, D. R. Karuppiah, and A. R. Hanson. Feature selection using adaboost for face expression recognition. Technical report, MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE, 2005.

[66] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[67] M. L. Smith, G. W. Cottrell, F. Gosselin, and P. G. Schyns. Transmitting and decoding facial expressions. *Psychological Science*, 16(3):184–189, 2005.

[68] R. Subramanian, D. Shankar, N. Sebe, and D. Melcher. Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *Journal of vision*, 14(3):1–18, 2014.

[69] R. Subramanian, J. Wache, M. Abadi, R. Vieriu, S. Winkler, and N. Sebe. ASCERTAIN: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, 2016.

[70] R. Subramanian, V. Yanulevskaya, and N. Sebe. Can computers learn from humans to see better?: Inferring scene semantics from viewers' eye movements. In *ACM International Conference on Multimedia*, pages 33–42, 2011.

[71] S. Sullivan, A. Campbell, S. B. Hutton, and T. Ruffman. What's good for the goose is not good for the gander: age and gender differences in scanning emotion faces. *Journals of Gerontology, Series B: Psychological Sciences and Social Sciences*, pages 1–6, 2015.

[72] M. van Vliet, A. Robben, N. Chumerin, N. V. Manyakov, A. Combaz, and M. M. V. Hulle. Designing a brain-computer interface controlled video-game using consumer grade eeg hardware. In *Biosignals and Robotics for Better and Safer Living (BRC)*, pages 1–6, 2012.

[73] M. van Vliet, A. Robben, N. Chumerin, N. V. Manyakov, A. Combaz, and M. M. Van Hulle. Designing a brain-computer interface controlled video-game using consumer grade eeg hardware. In *Biosignals and Biorobotics Conference (BRC), 2012 ISSNIP*, pages 1–6. IEEE, 2012.

[74] C. T. Vi, I. Jamil, D. Coyle, and S. Subramanian. Error related negativity in observing interactive tasks. In *Conference on Human Factors in Computing Systems*, pages 3787–3796, New York, NY, USA, 2014. ACM.

[75] G. Vingerhoets, C. Berckmoes, and N. Stroobant. Cerebral hemodynamics during discrimination of prosodic and semantic emotion in speech studied by transcranial doppler ultrasonography. *Neuropsychology*, 17(1):93, 2003.

[76] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vision*, 57(2):137–154, May 2004.

[77] V. Vonikakis, R. Subramanian, J. Arnfred, and S. Winkler. A probabilistic approach to people-centric photo selection and sequencing. *IEEE Transactions on Multimedia*, 2017.

[78] G. J. Walker-Smith, A. G. Gale, and J. M. Findlay. Eye movement strategies involved in face perception. *Perception*, 6:313–326, 1977.

[79] H. L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE Transactions on circuits and systems for video technology*, 16(6):689–704, 2006.

[80] S. Wang, J. Gwizdka, and W. A. Chaovalitwongse. Using wireless eeg signals to assess memory workload in the n-back task. *IEEE Transactions on Human-Machine Systems*, 46(3):424–435, 2016.

[81] L. J. Wells, S. M. Gillespie, and P. Rotshtein. Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze. *PloS one*, 11(12), 2016.

[82] S. Wiedenbeck, V. Grigoreanu, L. Beckwith, and M. Burnett. Gender HCI: What about the software? *Computer*, 39:97–101, 2006.

[83] Y. Wu, Y. Zhuang, X. Long, F. Lin, and W. Xu. Human gender classification: A review. *arXiv preprint arXiv:1507.05122*, 2015.

[84] S. Yang and F. Deravi. On the effectiveness of eeg signals as a source of biometric information. In *Emerging Security Technologies (EST), 2012 Third International Conference on*, pages 49–52. IEEE, 2012.

[85] W. Zhang, M. L. Smith, L. N. Smith, and A. R. Farooq. Gender and gaze gesture recognition for human-computer interaction. *Computer Vision and Image Understanding*, 149:32–50, 2016.

[86] W.-L. Zheng, B.-N. Dong, and B.-L. Lu. Multimodal emotion recognition using EEG and eye tracking data. In *Engineering in Medicine and Biology Society (EMBC)*, pages 5040–5043, 2014.

[87] W.-L. Zheng, J.-Y. Zhu, Y. Peng, and B.-L. Lu. EEG-based emotion classification using deep belief networks. In *International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2014.

[88] M. D. Zotto and A. J. Pegna. Processing of masked and unmasked emotional faces under different attentional conditions: an electrophysiological investigation. *Frontiers in psychology*, 6:1691, 2015.