



UNIVERSITÀ DEGLI STUDI
DI TRENTO

DEPARTMENT OF INFORMATION ENGINEERING AND COMPUTER SCIENCE
ICT International Doctoral School

**ANALYSIS OF USERS’
PSYCHO-PHYSIOLOGICAL PARAMETERS IN
RESPONSE TO AFFECTIVE MULTIMEDIA
A MUTLIMODAL AND IMPLICIT APPROACH FOR
USER-CENTRIC MULTIMEDIA TAGGING**

Mojtaba Khomami Abadi

Advisor

Prof. Nicu Sebe

Università degli Studi di Trento

April 2017

Abstract

The affective state of a user, during an interaction with a computer, is a great source of information for the computer in order to (i) employ the information for adapting an interaction, make the interaction flawless, leading in adaptive affective interfaces. The computer may also use emotional responses of a user to some affective multimedia content (ii) to tag the multimedia content with affective labels. The second is very useful to create affective profiles of users within real world applications for user-centric multimedia retrieval. Affective responses of users could be collected either explicitly (i.e. users directly assess their own emotions through computer interfaces) or implicitly (i.e. via sensors that collect psycho-physiological signals such as facial expressions, vocal clues, neuro-physiological signals, gestures and body postures). The major contributions of this thesis are as follows: (i) We present (and made publicly available) the very first multimodal dataset that includes the MEG brain signals, facial videos and some peripheral physiological signals of 30 users in response to two sets of affective dynamic stimuli. The dataset is recorded via cutting-edge lab equipments in highly controlled lab environments, facilitating proper analysis of MEG brain responses for affective neuro-science research. (ii) We then present two other multimodal datasets that we recorded using off-the-shelves market-available sensors for the purpose of analyzing users' affective responses to video clips and computer-generated music excerpts. The stimuli are selectively chosen to evoke certain target emotions. The first dataset also includes the BigFive personality traits of individuals and we show that it is possible to infer users' personality traits given their spontaneous reactions to affective videos. Both multimodal datasets are acquired via commercial sensors that are prone to noise artifacts that lead to some noisy uni-modal recordings. We made both datasets publicly available together with quality-assessments of

each signal recording. Within the research on the second dataset we present a multimodal inference system that jointly considers the quality of signals and ends up with highly signal noise tolerance. We also show that peripheral physiological signals include patterns that are similar across user. We develop a cross-user affect recognition system that is successfully validated via a leave-one-subject-out cross-validation scheme on the second dataset. (iii) We also present a crowdsourcing protocol for the collection of time-continuous affect annotations for videos. We collect a dataset of affective annotations for 12 videos with the contribution of over 1500 crowd-workers. We introduce algorithms to extract high quality time-continuous affect annotations for the 12 videos from the noisy crowd annotations. We observe that, for the prediction of time-continuous affect annotations given low-level multimedia content, higher regression accuracies are achieved when the crowd sourced annotations are employed as labels than expert annotations. The study suggests that expensive expert annotations for large affective video corpora developments could be replaced by crowdsourcing annotation techniques. Finally, we discuss opportunities for future applications of our research, and conclude with a summary of our contributions to the field of affective computing.

Keywords

affective computing, emotion recognition, neuro-physiological signals, personality, crowdsourcing

Acknowledgements

In the name of God, the Most Gracious, the Most Merciful

*I would first like to thank **my wonderful mother** whose soul is resting in peace now, Sheida Abed Lati, and **my beloved father**, Ali Khomami Abadi, without their continuous love, support and encouragement I never would have been able to achieve any of my goals. This one is for you mom and dad! I love you two for ever! I could never put into words how much I love you!*

*Second, a special thank you to **my lovely wife**, Negar Rostamzadeh. Words cannot describe how lucky I am to have you in my life. She has selflessly given more to me than I ever could have asked for. I love you very much, and I look forward to the lifelong journey ahead of us.*

*My special and profound thanks go to **my amazing brothers**, Mr. Koorosh Khomami Abadi and Mr. Morteza Khomami Abadi, who offered invaluable support, training, and love over my whole life. You have been always there for me and you are in my heart for ever!*

*I am very grateful to **my awesome advisor, Prof. Nicu Sebe**, for guiding me very well through my PhD journey and showing me the right way to become a good scientist. He taught me how to be persistent while tackling hard research problems. He supported me by all means in the development of top datasets that will serve greatly the field of affective computing. He kindly encouraged me to go through a great entrepreneurial experience that lead to the creation of Sensaura Inc. He is both a great advisor and an amazing friend. Thank you very much for giving me the great opportunity!*

I am also thankful to Prof. Paolo Avesani, Prof. Ioannis Patras, Ostad Hamidreza Bayat, Prof. Ramanathan Subramanian, Prof. Nathan Weisz, Dr. Gianpaolo Demarchi, Dr. Gianpiero Monittola who provided me with their valuable advice

and supports. Also thanks to my friends for their supports and mentorship at the ICT Doctoral school and the MHUG group at the university of Trento, NILAB at FBK, MEG lab at CIMEC, and the MMV group at Queen Mary University of London; especially to Dr. Francesca Belton, Dr. Andrea Stenico, Dr. Manuel Zucchellini, Dr. Seyed Mostafa Kia, Dr. Jacopo Staiano, Dr. Azad Abad, Dr. Julia Wache, Dr. Radu-Laurentiu Vieriu, Dr. Juan Abdn Miranda Correa, Dr. Fabio Morreale, Dr. Heng Yang, Dr. Thomas Hartmann, Dr. Emanuele Olivetti, and Dr. Philipp Ruhnau.

I thank the Semantics and Knowledge Innovation Lab (SKIL) , part of the Joint Open Labs network of Telecom Italia, that provided fund and supports for the first three years of my PhD. My thanks to my advisors, colleagues and friends at the SKIL specially to Dr. Michele Vescovi, Dr. Fabrizio Antonelli, Dr. Michele Caraviello, Mattia Pasolli, Silvana Bernaola and Roberto Larcher for the great time we had and their mentorship.

I thank all the participants who helped me with the research experiments. Your time and efforts are appreciated and yes! you have contributed to some great scientific achievements.

I thank my friends and colleagues at TandemLaunch Inc. who provided me a unique entrepreneurial research opportunity within the Sensaura project, especially to Dr. Helge Seetzen, Matthew Smith, Emilie Boutros, Jean-Philip R. Poulin, Jesus Cardenes Cabre, Robert Johnson, Prof. Björn W. Schuller, Prof. Rafael A. Calvo, Prof. Kyunghyun Cho, Prof. Christopher Pal, Seth Tropper, Fahd Benchekroun, Amir Bahador Gahroosi, Stefan M. Orzechowski, Claudia Torregrosa, Dr. Rishabh Gupta and Dr. Tara Akhavan.

The last but not the least I want to thank my dear father in-law Ahmadali Rostamzadeh, my dear mother in-law, Nayyereh Mabood Mojdehi, my dear brother in-law Dr. Pooya Rostamzadeh, my lovely aunt Marzieh, my dear cousins Rozita, Maria and Hoorieh who has been always there for me with their supports and encouragements. I appreciate all!

Contents

1	Introduction	1
1.1	Elicitation of emotions in users and stimuli selection	3
1.2	Implicit Characterisation of Users' Emotions Based on Spontaneous Responses	4
1.3	Implicit Characterisation of Users' Personality Based on Spontaneous Responses	6
1.4	Affective Multimedia Retrieval via Implicit Affective Tagging	8
1.5	Crowd sourcing affective multimedia tags:	9
1.6	Structure of the Thesis	10
2	MEG-based Multimodal Database for Decoding Affective Physiological Responses	11
2.1	Introduction	13
2.2	Related Work	15
2.3	Stimuli Selection	17
2.4	Experiment Setup	18
2.4.1	MEG, peripheral physiological signals, and NIR facial videos	18
2.4.2	Experimental set-up	22
2.5	Rating Analysis	25
2.5.1	Self-assessments: Music vs movie clips	25
2.6	Data Analysis	27

2.6.1	MEG preprocessing and feature extraction	27
2.6.2	Peripheral physiological feature extraction	29
2.6.3	Facial Expression Analysis	33
2.6.4	Multimedia features	33
2.7	MEG correlates with user ratings	33
2.8	Experimental Results	36
2.8.1	Single-trial Classification: MEG versus EEG	37
2.8.2	Classification procedure and results	38
2.8.3	Discussion of classification results	39
2.9	Continuous Emotion Estimation	42
2.10	Conclusion	45
3	Emotion and Personality Recognition using Commercial Sensors	47
3.1	The ASCERTAIN Dataset and Research	48
3.1.1	Related Work	51
3.1.2	ASCERTAIN Overview	55
3.1.3	Descriptive Statistics	61
3.1.4	Personality measures vs user ratings	67
3.1.5	Physiological correlates of emotion and personality	71
3.1.6	Recognition results	75
3.1.7	Discussion	79
3.2	Signal Quality Matters - QAMAF	81
3.2.1	Introduction	82
3.2.2	Affect Classifier Development	85
3.2.3	Quality Adaptive Multimodal Fusion	86
3.2.4	Results	88
3.2.5	Discussion and Conclusion	90
3.3	Conclusion	91

4	Crowdsourcing Continuous Affective Annotations for Video Tagging	93
4.1	Related work	95
4.1.1	Crowdsourcing	95
4.1.2	Affective movie analysis	96
4.1.3	Crowdsourcing for affective media tagging	96
4.1.4	Multi-task learning	97
4.2	Experimental Protocol	97
4.2.1	Dataset	97
4.2.2	Experimental Protocol	97
4.2.3	Annotation Mechanism	100
4.3	Multimedia Feature Extraction	101
4.3.1	Video Features	101
4.3.2	Audio Features	103
4.4	A Conditioned Crowd is Better Than the Expert	104
4.4.1	Quality control without prior knowledge	105
4.4.2	Quality control with prior knowledge	108
4.4.3	Aggregation of accepted annotations	110
4.4.4	Agreement Between Annotators	111
4.4.5	Wisdom of Crowd vs. Experts	113
4.4.6	Feature Preprocessing	114
4.4.7	Label Extraction, aggregation of continuous annotations	114
4.4.8	Regression Algorithm and Inner-loop parameter optimization	115
4.4.9	Information Fusion: Early vs. Late	116
4.4.10	Prediction Post-processing	117
4.4.11	Results	117
4.5	Applying a Multi-task Learning Framework	118
4.5.1	Data Analysis and Experiments	120

4.5.2 Experiments and Results	121
4.6 Conclusion and future work	123
5 Conclusion and Future Work	125
Bibliography	131

Chapter 1

Introduction

Giant leaps in Human-computer interaction (HCI) research over the past decade have made computers an integral part of human life. Like humans, intelligent agents endowed with cognitive capabilities can learn from user behavior to predict their actions and present information in a user-centric manner. Nevertheless, human actions are guided by both cognition and *emotion*, and a person's emotional state can provide significant clues to his/her behavior [25]. It would indeed be beneficial if HCI systems can interpret and learn from the user's emotional response- *e.g.*, knowing when the user is satisfied/frustrated with a system's output is valuable feedback, which can help improve the system's usability and user experience. *Affective computing* relates to the development of systems that can recognize human emotions during interactions, and formulate appropriate responses.

Recognition of a user's emotion given the psycho-physiological signals of the user is a fairly hard problem that we approached through our research in chapter 2 and chapter 3. Two major issues when dealing with supervised emotion recognition problem are (i) the presence of noise over input data and (ii) the presence of noise over affective labels where the labels are often provided by external *expert* observers or via users's self-assessments about their own feelings

The first issue is most evident when the train/test datasets are collected using commercial, portable sensors instead of specialized lab equipments (see chapter 3).

The second issue is one of the main challenges in the field of affective computing[16] that is usually tackled by employing affective stimuli that are not *controversial*, meaning, they evoke consistent emotions across users with various backgrounds (such as age, gender, race, culture and personality). The stimuli selection step is a fundamental steps in any user-centric affect analysis study that we also cover(See chapter 2).

In any supervised pattern recognition problem, providing a reliable ground-truth is one of the most critical steps to create a good model from 2 aspects: (i) it is very time consuming to collect high quantity of annotation labels, and (ii) hiring an expert to get high quality labels is expensive. Although providing reliably annotated big-datasets is very expensive and in some cases it is not feasible, there are trends in patter recognition community toward using big dataset, e.g. ImageNET[24], to achieve a more accurate recognizer. The idea of having huge and well annotated datasets is to include variety of information encoded in the predicting system. Russel et al. [137] provides an image database where the labels for an image segmentation task are collected via a web-based tool namely, LabelMe. Pattern analysis in multi media (e.g. action recognition, anomaly detection, emotion recognition and video quality assessment) is among the most interesting tasks in patter recognition. With the huge growth of multimedia databases (e.g. YouTube) the automatic content analysis (e.g. action recognition or emotion recognition) of multimedia has become a very remarkable and challenging problem. In chapter 4 we propose a framework for collection of high quality affective labels for vidoes.

1.1 Elicitation of emotions in users and stimuli selection

Many affective studies have been conducted with image stimuli, and there exist standard datasets such as [60] for researchers to conduct experiments and evaluate their findings. However, there exist few affective video datasets, in spite of studies confirming that reliable emotion elicitation is feasible with video stimuli such as movies [38]. An affective music video dataset, comprising 40 music videos, was recently presented in [57]. Our endeavor was to create a large-sized affective movie dataset along those lines owing to the following reasons:

1. The importance of context in emotion perception has been acknowledged by many studies (*e.g.*, [7]). Temporal context can be conveyed effectively by both audio and visual content in movies, whereas context in music videos is predominantly conveyed by the audio, which is supplemented by the visual information.
2. As a result, movies can effectively elicit a larger range of emotions (*e.g.*, including surprise/shock and fear) as compared to music videos.

In part of our research (chapter 2), we investigated the suitability of different types of video stimuli for emotion elicitation. For a study examining viewers' emotional responses to be successful, the employed stimuli should effectively elicit the emotions targeted by the study. While some works have attempted to identify appropriate video stimuli for studying affect [32, 8], different authors have employed different stimuli for emotion elicitation. [68] presents an affect characterization study using 21 movie clips, while the authors in [57] elicit emotions through music videos.

While affective content creators *intend* to convey a certain emotion (or a set of emotions) through the created stimulus, the *actual* emotion induced upon perceiving the stimulus is influenced by a number of psychological and contextual factors, and can therefore be highly subjective. Consequently, correlating the

1.2. IMPLICIT CHARACTERISATION OF USERS' EMOTIONS BASED ON SPONTANEOUS RESPONSES

observed emotional response with the *expected* response is a non-trivial problem which is typically simplified in practice employing the following ideas: (1) Most affective studies assume that the entire gamut of human emotions can be represented as a set of points on the valence-arousal¹ plane as demonstrated by Greenwald *et al.* [31], and (2) To largely ensure that the elicited and expected emotions are consistent, the presentation stimuli are carefully selected based on previous studies, or based on 'ground truth' valence-arousal ratings compiled from a large population that evaluates the stimuli prior to the actual experiment.

Emotional states have been found to produce specific types of physiological responses- *e.g.*, excitement is associated with increased heart-beat and respiration rates, and this correlation is exploited in a number of physiology-based affect studies. Heart-rate, skin temperature and conductance level, blood pressure and facial EMG are recorded as subjects view affective imagery in [103]. Their experiments indicate that the responses for anger and fear are uniquely different from responses to neutral images.

1.2 Implicit Characterisation of Users' Emotions Based on Spontaneous Responses

Humans perceive emotions from the environment through visual and auditory stimuli- characterized by speech, audio/video music clips, images and movies in the digital world. While many studies have investigated how speech and image signals can effectively elicit emotions in people [84, 103], research on isolating emotional content in music and movie videos began only recently. Past works such as [37, 41] have attempted to identify emotions either by (i) analyzing the content to develop models that link low-level image and audio features to *valence* (emotion type) and *arousal* (emotion intensity) or (ii) ana-

¹*Valence* indicates the type of emotion induced by the stimulus in the viewer (*e.g.*, pleasant or unpleasant), while *arousal* denotes the intensity of emotion (*e.g.*, exciting or boring) [37].

lyzing the viewer’s facial activity/expressions and correlating these responses with the presented content. While content-based analysis enables discovery of video highlights (typically high-arousal segments), it is inherently not suited for tagging content on the valence-arousal plane. Conversely, while facial expressions can provide some insight regarding emotional video content, they can easily be controlled by the viewer and are therefore, not always reliable.

The above shortcomings have prompted researchers to investigate emotional response to affective stimulus through physiological responses such as (i) Electroencephalogram (EEG), which measures electrical activity along the scalp, (ii) Electromyogram, measuring electrical activity of skeletal muscles, (iii) heart rate, (iv) galvanic skin response (GSR) measuring skin conductance and (v) skin temperature, *etc.* These signals² have been found to effectively encode emotional responses [68, 57] and are more primitive than facial expressions, which typically denote the conscious manifestation of an emotion.

In chapter 2, we examine the feasibility of employing the Magnetoencephalogram (MEG) signal for measuring emotional responses to affective music and movie videos. MEG localizes activated superficial parts of the brain. When a group of neurons is activated, electrical currents along the neurons generate tiny, orthogonally oriented magnetic fields. The sum of these magnetic fields generates a change in magnetic field around the activated part, and constitutes the MEG response. While many EEG studies (*e.g.*, [68, 57, 109]) have successfully decoded affective viewer response to videos, there are no such MEG-based studies. However, the fact that MEG can effectively encode affective responses, similar to EEG, is demonstrated in [90] employing image stimuli. Their results are obtained on analyzing event-related magnetic fields (ERF), where an individual’s brain responses are acquired over many trials and averaged. In contrast, *we present MEG could be employed for single-trial clas-*

²EEG is the response from the central nervous system, while the remaining are responses from the peripheral nervous system, and are therefore termed peripheral physiological signals.

sification of affective viewer responses to videos.

1.3 Implicit Characterisation of Users' Personality Based on Spontaneous Responses

The need to recognize the affective state of users for effective human-computer interaction has been widely acknowledged. Nevertheless, affect is a highly subjective phenomenon influenced by a number of contextual and psychological factors including *personality*. The relationship between individuals' personality traits and emotional responses has been actively studied by social psychologists ever since a correlation between the two was proposed in Eysenck's personality model [28]. Eysenck posited that (i) Extraversion, the personality dimension that describes a person as being either talkative or reserved, is accompanied by low cortical arousal— *i.e.*, extraverts require more external stimulation than introverts, and ii) Neurotics, characterized by negative feelings such as depression and anxiety, become very easily upset or nervous due to minor stressors, while emotionally stable persons remain composed under pressure.

While multiple factors such as stimuli used for emotion elicitation, quality of recruits and number of experimental trials have been found to be influencing user-centric affect recognition [109], the critical influence of personality differences on emotion perception has not been examined by prior affect recognition works to the best of our knowledge. In chapter 3, we describe a novel methodology to decode users' Big-five personality factors [21]— Extraversion, Neuroticism, Conscientiousness, Agreeableness and Creativity, as they watch emotional movie clips. Their physiological responses to the affective stimuli are acquired in the form of Electrocardiogram (ECG), Galvanic Skin Response (GSR), Electroencephalogram (EEG) and facial response signals, and these are fused employing both feature and decision fusion techniques to predict personality dimensions. As a first step, we examine correlations between explicit

ratings of users obtained post clip viewing, and their personality scores through statistical analysis. Then, we analyze the correlations between physiological features, emotional responses and personality measures. Finally, classification results for the five personality dimensions are presented. Chapter 3 makes the following research contributions:

1. To our knowledge, our research outputs that are presented in chapter 3 cover the first work to attempt personality profiling based on implicit user responses to affective multimedia. Personality assessment has traditionally been achieved through the use of questionnaires, or by analyzing users' behavior in videos and social media. The methodology we propose can enable simultaneous and automated annotation of affective content and personality dimensions.
2. Among studies that have analyzed the relationship between personality traits and affective responses, the study is the first work to attempt prediction of all the big-five factors as well as use movie stimuli for eliciting emotions. Furthermore, the use of off-the-shelves portable sensors for measuring user responses enhances the applicability of our profiling framework in real-life settings.
3. Employing off-the-shelves portable sensors, datasets are prone to noise artifacts. Among studies that have studied the emotion recognition at the presence of noise in the signals, our research is the first work to attempt fusing the information content of various modalities and merge them efficiently to make the recognition system highly robust to noise artifacts.

1.4 Affective Multimedia Retrieval via Implicit Affective Tagging

Affective media tagging has evoked considerable interest among multimedia researchers lately. Varied methodologies have been adopted for characterizing affective media including analysis of the content [37], or the behavior of users viewing emotional content in terms of facial expressions [41] and physiological responses such as brain activity, heart beat rate and skin conductance [68, 57, 109, 50]. While content-based methods have been unable to bridge the semantic gap between low-level audiovisual features and high-level emotion, user-based approaches have only achieved moderate success due to prevalent differences between the *expected* emotion (which the content creator or director intends to convey), and the *actual* emotion evoked in different users.

The burgeoning number of platforms for online multimedia streaming/storage (e.g. YouTube, Netflix) has resulted in generation of a huge database of multimedia content online. In 2015, approximately 400 hours of videos were uploaded every minute on YouTube³. This enormous amount of data is not limited to the online realm, the availability of portable devices storing thousands of music tracks and pictures has brought massive amounts of multimedia information in our pockets. However, the huge amount of generated multimedia information needs to be indexed to be searchable and retrievable by the users.

Most of existing multimedia retrieval systems that rely on user-generated labels for indexing are potentially biased by subjective judgements and/or intentions[85]. Moreover, manual tagging of multimedia content interrupts the user experience process. Therefore, it is necessary to automate the process of multimedia indexing through implicit tagging. The classical multimedia indexing relies on cognitive indexing procedures which are based on concepts to characterize the multimedia content, such as locations, objects, and events.

³www.reelseo.com/hours-minute-uploaded-youtube/

Whereas, a recent approach, the so-called affective indexing, depends on the emotions generated by the multimedia content [109]. The implicit affective indexing technique is expected to provide more detailed and meaningful information regarding users experience with multimedia [85, 52]. Previously, affective tags have been used for indexing multimedia content for improving information retrieval and recommendation systems [98, 52, 134].

1.5 Crowd sourcing affective multimedia tags:

With the proliferation of multimedia content on the web, the need to tag or index audio and video based on the type of information (*sports, documentary*) and emotions (*funny, exciting*) they convey has become essential— Hanjalic and Xu [37] term the former as cognitive and the latter as affective categorization. Furthermore, since movie genres are expressly defined by the emotions they evoke (*e.g., comedy, thriller, romance*), the need to develop automated methods for affective movie categorization is paramount. However, content-centric approaches that attempt to estimate the scene emotion continuously over time [37], and user-centered methods that utilize physiological responses to estimate the general scene emotion [52, 57] have only been moderately successful. The limited success has been partly due to the (1) inherent difficulty in representing emotion and (2) non-availability of extensively labeled training data for this purpose.

Given that data label quality improves with the number of annotations [99], in chapter 4, we suggest that emotion tags compiled from a crowd should be comparable in quality to that obtained from a few experts. To validate this hypothesis, we obtained time-continuous valence and arousal annotations for 12 Hollywood movie clips from (a) seven experts in controlled lab conditions, and (b) numerous crowd workers in uncontrolled conditions. We then systematically condition the crowdsourced annotations using a series of filters. Process-

ing both the expert and (cleaned) crowd-annotated data, we arrive at a representative time-continuous emotion characterization for each clip. Finally, we predict the valence and arousal for a given clip (i) based on a generalized linear model (GLM) trained with low-level, audio-visual features and (expert or crowd) annotations for the remaining clips and (ii) a multi-task learning approach. Comparison of prediction results obtained with expert vs crowd labels showed that the *crowd model outperformed the expert model*, thereby confirming our hypothesis.

1.6 Structure of the Thesis

This thesis is structured as follows: in Chapter 2 we report our research on a multimodal MEG-based dataset for user-centric affect recognition that is collected within very controlled lab environments; in Chapter 3 we present two datasets that are collected via commercial sensors in user-friendly environments. We present it is possible to recognize personality of users given their psychophysiological signals. It is also possible to perform cross-user affect recognition using a novel signal noise tolerant system; in Chapter 4 we propose a crowdsourcing solution to obtain high-quality time-continuous affect annotations for affective multimedia tagging. Finally, in Chapter 5, we summarize the thesis and we elaborate the possible future research directions.

Chapter 2

MEG-based Multimodal Database for Decoding Affective Physiological Responses

Humans feel emotions when looking at movies or when listening to or watching music videos. One way to capture these emotions is using facial expressions, but these are easily controllable and not always reliable. In addition, the literature has investigated facial expressions and psychological signals in depth. On the other hand, the brain signals seem to be a more reliable way of capturing the genuine emotions. To the best of our knowledge, particularly the Magnetoencephalogram (MEG) responses to dynamic stimuli has not been investigated in any emotion recognition studies. The hypothesis is that by measuring these signals we are able to capture reliably the emotions felt by the users. By doing a comprehensive study under several stimuli we want to validate the hypothesis and show that indeed by using MEG, one can get a good estimate of the emotions.

Upon reviewing related literature, one can make the following observations:

1. All these studies, apart from DEAP [57], derive their conclusions from experiments involving a relatively small number of stimuli. This is because such studies are inherently hard to conduct. One needs to take into account

the time required for subject preparation, stimulus viewing and recording user ratings while designing the experiment protocol. Also, the fact that fatigue strongly influences the quality of emotional responses discourages lengthy experiments with many stimuli.

2. While all these approaches have been generally successful in isolating physiological correlates of specific emotions arising from the presented stimuli, no comparison studies have been made to determine which stimulus is ideally suited for affect computation, given the experiment hypotheses and duration. This research targets one of the first steps in that direction.
3. There are no available datasets having recorded MEG signals, physiological signals and video of the face of the subjects while they were stimulated by emotional video clips other than the one we developed in our research.

The gap that we aim to fill is the one left by the deficit of a multi-modal emotion recognition system, that has the capability to analyze MEG brain signals, standard physiological signals, facial expressions, and the verbal and non-verbal behaviors in response to dynamic stimuli. Another contribution of this study is to find the mappings between different modalities. Although each aspect of this problem (except MEG) has been considered as an interesting problem and has been investigated previously, studying all of these together gives us a proper scale for comparing these modalities regarding their ability to encode emotions.

This chapter¹ covers the development of a dataset² to answer the following research questions (RQs):

- RQ1: Doed MEG brain signals encode affective brain responses to affective video contents?

¹The research is an extension to our previous works [49, 50] and it is published [53] in IEEE Transactions on Affective Computing, Jan. 2015.

²The developed dataset is publicly available to the research community:
mhug.disi.unitn.it/wp-content/DECAF/DECAF.html

- RQ2: Which class of affective video stimuli works better in eliciting consistent target emotions across users? Music video clips or video movie clips?

2.1 Introduction

Affect recognition is a necessity in human-computer interaction. Users' demands can be implicitly inferred from their emotional state, and systems effectively responding to emotional inputs/feedback can greatly enhance user experience. However, affect recognition is difficult as human emotions manifest both explicitly in the form of affective intonations and facial expressions, and subtly through physiological responses originating from the central and peripheral nervous system. Given that the majority of multimedia content is created with the objective of eliciting emotional reactions from viewers, representing, measuring and predicting emotion in multimedia content adds significant value to multimedia systems [1]. Approaches to predict affect from multimedia can be categorized as (i) *content-centric* [37, 126], using primitive audio-visual features which cannot adequately characterize the emotion perceived by the viewer, or (ii) *user-centric*, employing facial expressions [41] and speech intonations [84], which denote a conscious and circumstantial manifestation of the emotion, or peripheral physiological responses [68], which capture only a limited aspect of human emotion.

Recently, cognition-based approaches employing imaging modalities such as fMRI and EEG to map brain signals with the induced affect [38, 57, 109] have gained in popularity, and brain signals encode emotional information complementary to multimedia and peripheral physiological signals, thereby enhancing the efficacy of user-centric affect recognition. However, acquisition of high-fidelity brain signals is difficult and typically requires the use of specialized lab equipment and dozens of electrodes positioned on the scalp, which impedes

naturalistic user response. Magnetoencephalogram (MEG) is a non-invasive technology for capturing functional brain activity, which requires little physical contact between the user and the sensing coil (Fig. 2.2), and therefore allows for (1) recording meaningful user responses, with little psychological stress and (2) compiling affective responses over long time periods. Also, MEG responses can be recorded with higher spatial resolution as compared to EEG.

In this chapter, we present **DECAF**— a MEG-based multimodal database for **decoding affective** user responses. Benefiting from facile data acquisition, DECAF comprises affective responses of 30 subjects to 36 movie clips (of length $\mu=80s$, $\sigma=20$) and 40 1-minute music video segments (used in [57]), making it one of the largest available emotional databases³. In addition to MEG signals, DECAF contains synchronously recorded near-infra-red (NIR) facial videos, and horizontal Electrooculogram (hEOG), Electrocardiogram (ECG), and trapezius-Electromyogram (tEMG) peripheral physiological responses⁴. A major limitation of affective computing works [68, 57, 109] that DECAF seeks to address is the lack of benchmarking with respect to stimuli and sensing modalities. DECAF facilitates comparisons between (1) MEG vs. EEG modalities for affect sensing via their performance on the DEAP database [57], and (2) music-video vs. movie clips concerning their suitability for emotion elicitation.

We present analyses concerning (i) participants’ self-assessment ratings for *arousal* and *valence* for music and movie stimuli, (ii) correlations between user ratings (explicit feedback) and implicitly observed MEG responses, and (iii) single-trial classification of *valence*, *arousal* and *dominance* from MEG, peripheral responses, facial activity, content-based audio visual features and fusion of these modalities. Finally, *time-continuous* emotion annotations useful for dynamic emotion analysis, were compiled from seven experts for the movie clips— as an application, we show dynamic emotion prediction on time-contiguous

³<http://disi.unitn.it/~mhug/DECAF.html>

⁴DECAF represents a significant extension of the dataset reported in [50], which only contains MEG and peripheral physiological responses of 18 subjects.

snippets from the movie clips with a model trained using these annotations and audio-visual/MEG features.

The chapter is organized as follows: Section 2.2 overviews related work. Methodology adopted for movie clip selection is described in Section 2.3, while the experimental protocol is detailed in Section 2.4. Analysis of users' self assessments is presented in Section 2.5, while features extracted for affect recognition are described in Section 2.6. Correlations between self-assessments and physiological responses along with single-trial classification results are presented in Sections 2.7 and 2.8. Dynamic emotion estimation is detailed in Section 2.9, and conclusions are stated in Section 2.10.

2.2 Related Work

Creating a stimulus database for eliciting emotions is crucial towards understanding how affect is expressed in controlled lab conditions. The *actual* emotion induced upon perceiving a stimulus designed to elicit an *intended* emotion is influenced by a number of psychological and contextual factors, and can therefore be highly subjective. Consequently, ensuring that the *actual* affective response is in agreement with the *intended* response is non-trivial, and is typically achieved in practice as follows: (1) Many affective studies assume that the entire gamut of human emotions can be represented on the valence-arousal-dominance⁵ (VAD) space as proposed by Bradley [13], and (2) To largely ensure that the elicited and intended emotions are consistent, presentation stimuli are carefully selected based on literature, or based on 'ground truth' V-A ratings acquired from a large population that evaluates them prior to the actual study.

Gross and Levenson's seminal work on affective database creation [32] eval-

⁵*Valence* indicates emotion type (*pleasant* or *unpleasant*), while *arousal* denotes the intensity of emotion (*exciting* or *boring*). Dominance measures the extent of control on viewing a stimulus (feeling *empowered* or *helpless*) [57]. We mainly use the VA-based affect representation, shown to account for most emotional responses by Greenwald *et al.* [31].

uates the responses of 494 subjects to 250 movie clips for identifying 16 movie clips capable of evoking eight target emotions. Content-based affect recognition works [37, 126] also perform emotion analysis on movie clips/scenes. User-centric emotion recognition works have employed a variety of stimuli to elicit emotions— Joho *et al.* [41] use a combination of movie and documentary clips to evoke facial activity, which is then used for highlights detection. Use of physiological responses for recognizing affect, pioneered by Sinha and Parsons [103] to distinguish between neutral and negative imagery, has gained popularity recently. Lisetti and Nasoz [68] use movie clips and mathematical equations to evoke emotions, which are decoded from users' skin conductance, heart rate, temperature, EMG and heat flow responses. Kim and André [56] use audio music clips to induce emotions, recognized through heart rate, EMG, skin conductivity and respiration changes.

Among cognition-based approaches, the DEAP dataset [57] is compiled to develop a user-adaptive music recommender system. It contains EEG, galvanic skin response (GSR), blood volume pressure, respiration rate, skin temperature and EOG patterns of 32 viewers watching 40 one-minute music video excerpts. The MAHNOB-HCI database [109] is compiled to model emotional responses of users viewing multimedia stimuli. It contains face and upper-body video, audio, physiological and eye-gaze signals of 27 participants watching 20 emotional movie/online clips in one experiment, and 28 images and 14 short videos in another. Analyses on the DEAP and MAHNOB-HCI datasets confirm that EEG effectively encodes emotional information, especially arousal.

Examination of related works reveals that user-centered affect recognition has been achieved with diverse stimuli, reflecting the fact that human affect sensing is multimodal. However, indigenous stimuli and signals employed by each of these works provides little clarity on (1) which stimulus most effectively elicits consistent emotional responses across users, in order to maximize our understanding of affect perception and expression, and (2) which modality best

characterizes user emotional responses— answers to these questions can increase the efficacy of affect recognition approaches. DECAF is compiled with the aim of evaluating both stimuli and sensing modalities for user-centered affect recognition.

2.3 Stimuli Selection

One of our objectives was to compile a large database of affective movie stimuli (comparable in size to DEAP [57]) and user responses for the same. This section describes how the 36 movie clips compiled to this end were selected. Based on previous studies that have identified movie clips suited to evoke various target emotions [32, 8], we initially compiled 58 Hollywood movie segments. These clips were shown to 42 volunteers, who self-assessed their emotional state on viewing each video to provide: valence level (very negative to very positive), arousal level (very calm to very excited), and the most appropriate tag that describes the elicited emotion (Table 2.1).

These annotations were processed to arrive at the final set of 36 clips as follows:

- (1) To ensure that the annotations are comparable, we transformed all V and A annotations using the z -score normalization.
- (2) To better estimate the affective perception of annotators, we discarded the outliers from the pool of annotators for each video clip as follows: Along the V-A dimensions, we thresholded the annotations at zero to associate *high* (H_i) and *low* (L_i) video sets to each annotator ($i = 1 \dots 42$). We then computed Jaccard distances D_H, D_L (42×42 matrices) between each pair of annotators i, j for the *high*, *low* sets, *e.g.*, $D_H(i, j) = 1 - \frac{|H_i \cap H_j|}{|H_i \cup H_j|}$, where $|\cdot|$ denotes set cardinality, and cumulative distance for each annotator from peers as the sum of each row. Finally, we derived Median Absolute Deviation of the cumulative distance distribution, and those annotators more than 2.5 deviations away from

the median were considered outliers as per [65]. In all, 5 and 2 outlier annotators were respectively removed for the V and A dimensions.

(3) Similar to [57], we computed μ/σ from the inlier V-A ratings for each movie clip as plotted in Fig. 2.1, and chose 36 clips such that (a) their ratings were close to the corners of each quadrant, (b) they were uniformly distributed over the valence-arousal plane, and (c) only one clip per movie was chosen from each quadrant to avoid priming effects. Table 2.1 contains descriptions of the selected movie clips, while Fig. 2.1 presents the distribution of μ/σ ratings for the original 58 clips and highlights the 36 selected clips. The mean V-A ratings listed in Table 2.1 are considered as *ground truth* annotations in our work. The chosen movie clips were 51.1–128.2s long ($\mu = 80, \sigma = 20$) and were associated with diverse emotional tags. For benchmarking affective stimuli, we also recorded emotional responses to 40 one-minute music video used in the DEAP study [57].

2.4 Experiment Setup

In this section, we present a brief description of (a) MEG, peripheral physiological and facial signals recorded in the study before detailing the (b) experimental set-up and protocol.

2.4.1 MEG, peripheral physiological signals, and NIR facial videos

To collect users’ implicit affective responses, we recorded (i) Magnetoencephalogram (MEG), (ii) horizontal Electrooculogram (hEOG), (iii) Electrocardiogram (ECG), (iv) Trapezius Electromyogram (tEMG) and (v) Near Infra-red (NIR) facial video signals that are described below.

MEG: MEG technology enables non-invasive recording of brain activity and is based on SQUIDS (Super-conducting Quantum Interference Devices), which enables recording of very low magnetic fields. Magnetic fields produced by the

CHAPTER 2. MEG-BASED MULTIMODAL DATABASE FOR DECODING AFFECTIVE PHYSIOLOGICAL RESPONSES

Table 2.1: Description of movie clips selected for the DECAF study with their duration in seconds (L), most frequently reported emotion tag and statistics derived from 42 annotators. Introductory videos are marked with **.

Emotion	ID	Source Movie	L	Valence		Arousal		Scene Description
				μ	σ	μ	σ	
Amusing	01	Ace-Ventura: Pet Detective	102.1	1.22	0.53	1.03	1.00	Ace Ventura successfully hides his pets from the landlord
	02	The Gods Must be Crazy II	67.1	1.56	0.50	1.20	0.96	A couple stranded in the desert steal ostrich eggs for food
	04	Airplane	85.2	0.99	0.83	1.15	0.88	Woman and co-passengers react as pilot struggles to control aircraft
	05	When Harry Met Sally	100.2	1.05	0.61	1.08	1.02	Sally shows Harry how women fake orgasms at a restaurant
	**	Modern Times	106.4	0.87	0.69	-0.35	0.86	Bewildered factory worker in an assembly line
Funny	03	Liar Liar	55.1	0.95	0.65	0.56	0.96	Prosecution and defense discuss a divorce case in court
	06	The Gods Must be Crazy	52.1	1.26	0.56	0.81	1.15	Man tries to get past an unmanned gate on a brakeless jeep
	07	The Hangover	90.2	0.95	0.70	0.85	1.06	Group of friends on the morning after a drunken night
	09	Hot Shots	70.1	0.98	0.66	0.81	0.90	A hilarious fight sequence
Happy	08	Up	67.1	1.42	0.43	0.35	1.18	Carl- a shy, quiet boy meets the energetic Elle
	10	August Rush	90.1	0.76	0.68	-1.17	1.02	A son meets his lost mother while performing at a concert
	11	Truman Show	60.1	0.90	0.50	-1.98	0.69	Truman and his lover go to the beach for a romantic evening
	12	Wall-E	90.2	1.41	0.53	-0.82	0.91	Wall-E and Eve spend a romantic night together
	13	Love Actually	51.1	1.03	0.70	-1.38	0.80	Narrative purporting that 'Love is everywhere'
	14	Remember the Titans	52.1	0.79	0.58	-0.99	0.82	Titans win the football game
	16	Life is Beautiful	58.1	1.10	0.42	-0.16	0.79	Funny Guido arrives at a school posing as an education officer
	17	Slumdog Millionaire	80.1	0.94	0.35	-0.34	0.85	Latika and Jamal unite at the railway station
18	House of Flying Daggers	77.2	0.84	0.56	-1.79	0.88	Young warrior meets with his love with a bouquet	
Exciting	15	Legally Blonde	51.1	0.64	0.37	-0.62	0.80	Elle realizes that she has been admitted to Harvard Law School
	33	The untouchables	117.2	-0.70	0.60	1.05	0.70	Shoot-out at a railway station
Angry	19	Gandhi	108.1	-0.50	0.67	-1.00	0.92	Indian attorney gets thrown out of a first-class train compartment
	21	Lagaan	86.1	-0.98	0.49	-0.69	0.71	Indian man is helpless as a British officer threatens to shoot him
	23	My Bodyguard	68.1	-0.81	0.59	-1.35	0.79	Group of thugs provoke a teenager
	35	Crash	90.2	-1.56	0.45	0.45	0.95	A cop molests a lady in public
Disgusting	28	Exorcist	88.1	-1.52	0.64	1.71	0.90	An exorcist inquires a possessed girl
	34	Pink Flamingos	60.2	-1.95	0.61	0.18	0.83	A lady licks and eats dog faeces
Fear	30	The Shining	78.1	-0.85	0.49	1.01	0.95	Kid enters hotel room searching for his mom
	36	Black Swan	62.2	-1.07	0.35	1.00	0.73	A lady notices paranormal activity around her
	**	Psycho	76.2	-1.23	0.73	0.44	1.01	Lady gets killed by intruder in her bath tub
Sad	20	My girl	60.1	-0.85	0.62	-0.82	1.06	Young girl cries at her friend's funeral
	22	Bambi	90.1	-0.95	0.37	-0.43	1.07	Fawn Bambi's mother gets killed by a deer hunter
	24	Up	89.1	-0.99	0.45	-0.97	0.76	Old Carl loses his bedridden wife
	25	Life is Beautiful	112.1	-0.62	0.41	-0.16	0.81	Guido is caught, and shot to death by a Nazi soldier
	26	Remember the Titans	79.1	-0.84	0.53	-0.55	0.87	Key <i>Titans</i> player is paralyzed in a car accident
	27	Titanic	71.1	-0.98	0.57	-0.30	0.99	Rescuers arrive to find only frozen corpses in the sea
31	Prestige	128.2	-1.24	0.73	1.20	0.88	Lady accidentally dies during magician's act	
Shock	29	Mulholland Drive	87.1	-1.13	0.55	0.82	0.97	Man shocked by suddenly appearing frightening figure
	32	Alien	109.1	-0.99	0.71	1.22	0.76	Man is taken by an alien lurking in his room

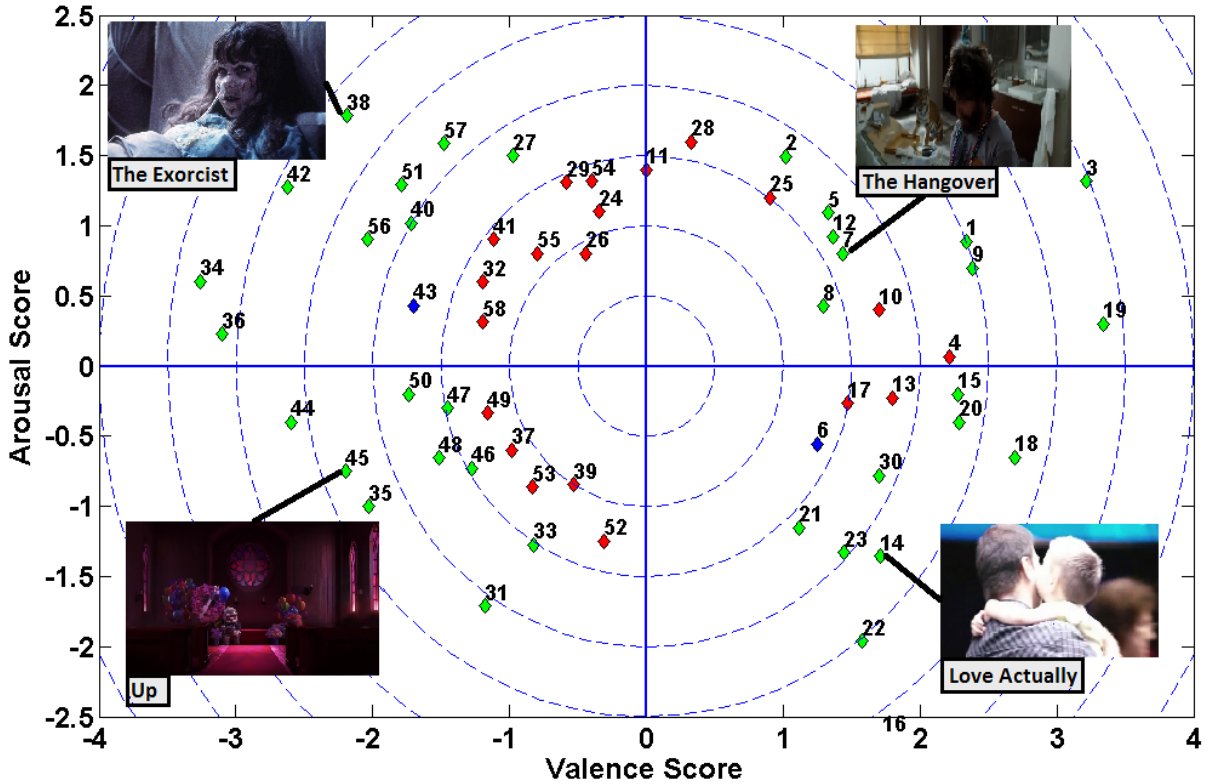


Figure 2.1: Distribution of videos' μ/σ ratings in the V-A plane. The 36 selected videos are highlighted in green, while two introductory videos are highlighted in blue.

human brain are in the order of femtotesla (fT) and since sensors are really sensitive to noise, the MEG equipment is located in a magnetically shielded room insulated from other electrical/metallic installations. A multiple coils configuration enables measurement of magnetic fields induced by tangential currents, and thus, brain activity in the sulci of the cortex can be recorded. We used the *ELEKTA Neuromag* device which outputs 306 channels (corresponding to 102 magnetometers and 204 gradiometers, as in Fig. 2.5) with a sampling frequency of 1 KHz.

Unlike in EEG, MEG sensors do not touch the subject's head and the participant can potentially make head movements during the recordings. However, due to high spatial resolution, even small head movements will cause a sensor to sense another part of the brain and induce changes in the MEG signal. There-

fore, we asked subjects to not move their head during the recordings. To compensate for inadvertent head movements, before each recording, we attached five Head Position Indicator (HPI) coils to accurately determine the subject's head pose. Two HPI coils were attached behind the ears without being in the hair, while three coils were interspersed on the forehead. Prior to the experiment, we also recorded the subject's skull shape by sampling the 3D positions of 210 points uniformly distributed around the skull⁶.

ECG: ECG is well known for its relevance in emotion recognition [56, 57, 109]. ECG signals were recorded using three sensors attached to the participant. Two electrodes were placed on the wrist, and a reference was placed on a boney part of the arm (ulna bone). This setup allows for precise detection of heart beats, and subsequently, accurate computation of heart rate (HR) and heart rate variability (HRV).

hEOG: Electrooculography denotes the measurement of eye movements, fixations and blinks. In this study, we used hEOG which reflects the horizontal eye movement of users by placing two electrodes on the left and right side of the user's face close to the eyes. Zygomatic muscle activities produce high frequency components in the bipolar EOG signal, and hence the EOG signal also captures facial activation information.

tEMG: Different people exhibit varying muscle movements while experiencing emotions. However, some movements are involuntary— *e.g.*, nervous twitches produced when anxious, nervous or excitable. Trapezius EMG is shown to effectively correlate with users' stress level in [128]. We placed the EMG bipolar electrodes above the trapezius muscle to measure the mental stress of users as in [56, 57]. The ECG reference electrode also served as reference for hEOG and tEMG.

NIR Facial Videos: As the MEG equipment needs to be electrically shielded,

⁶While DECAF contains HPI information, HPI-based MEG signal compensation will be attempted in future work. Since head-movement can induce noise in the MEG data, HPI MEG compensation can be useful for discarding noise and improving signal-to-noise ratio.

traditional video cameras could not be used for recoding facial activity, and we therefore used a near infra-red camera for the same. Facial videos were recorded as *avi* files at 20 fps.

The ELEKTA Neuromag device accurately synchronizes MEG signals with the peripheral physiology signals. Synchronization of the NIR videos was handled by recording the sound output of the stimulus presentation PC with the user’s facial videos, and using this information to determine stimulus beginning/end.

2.4.2 Experimental set-up

Materials: All MEG recordings were performed in a shielded room with controlled illumination. Due to sensitivity of the MEG equipment, all other devices used for data acquisition were placed in an adjacent room, and were controlled by the experimenter. Three PCs were used, one for stimulus presentation, and two others for recording NIR videos and MEG, physiology data as seen in Fig. 2.2. The stimulus presentation protocol was developed using MATLAB’s Psychtoolbox (<http://psychtoolbox.org/>) and the ASF framework [96]. Synchronization markers were sent from the stimulus presenter PC to the MEG recorder for marking the beginning and end of each stimulus. All stimuli were shown at 1024×768 pixel resolution and a screen refresh rate of 60 Hz, and this display was projected onto a screen placed about a meter before the subject inside the MEG acquisition room (Fig. 2.2). All music/movie clips were played at 20 frames/second, upon normalizing the audio volume to have a maximum power amplitude of 1. Participants were provided with a microphone to report their emotional state and communicate with the experimenters.

Protocol: 30 university graduate students (16 male, age range 27.3 ± 4.3) participated in the experiments. Data acquisition for each participant was spread over two sessions— movie clips were presented in one session, and music videos in the other (Fig. 2.3). The presentation order of the music and movie clips

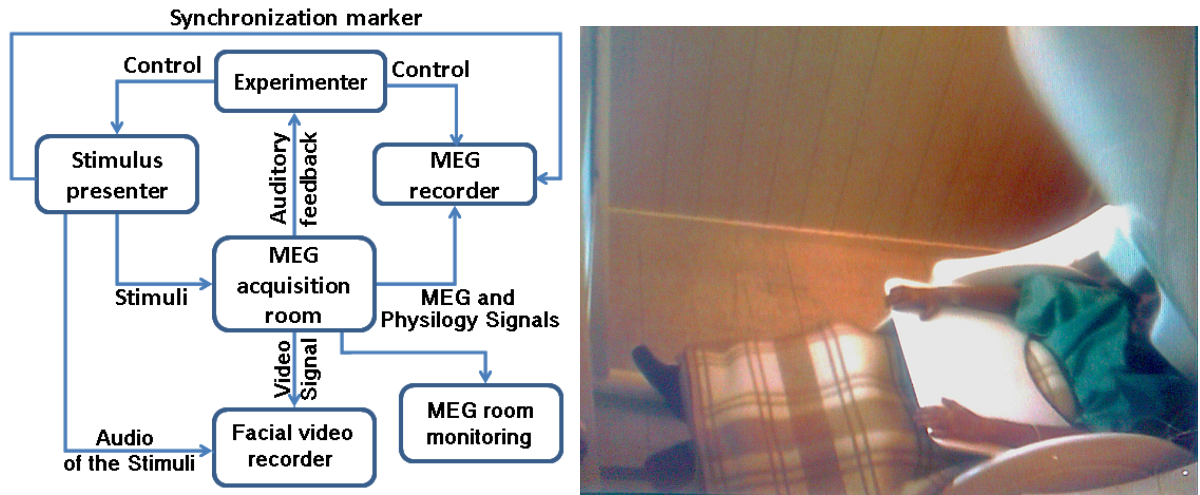


Figure 2.2: (Left) Illustration of the experimental set-up. (Right) A subject performing the experiment– the stimulus is presented on the screen to the left, while the subject is seated under the MEG equipment on the right.

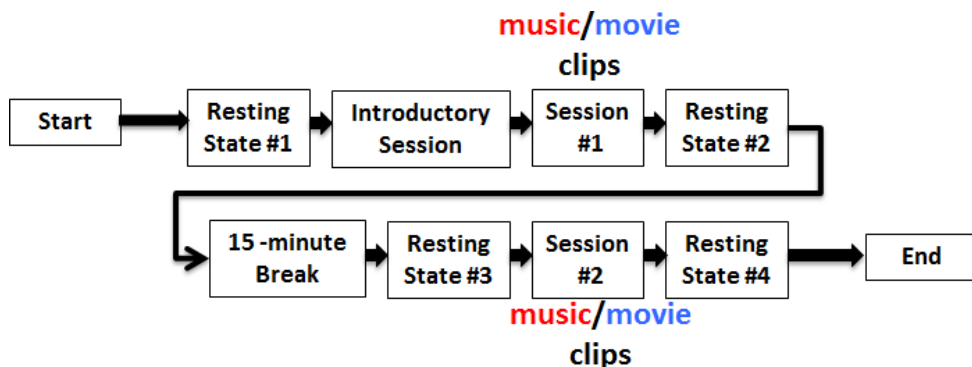


Figure 2.3: Timeline for experimental protocol.

was counterbalanced across subjects. During each session, music/movie clips were shown in random order, such that two clips with similar valence, arousal characteristics did not follow one another. To avoid fatigue, each recording session was split into two halves (20 music/18 movie clips shown in each half) and lasted one hour. We recorded the resting state brain activity for five minutes at the beginning of each session, and for one minute at the end or before/after breaks.

Subject Preparation: To ensure the absence of metallic objects near the

MEG equipment, prior to each recording session, participants had to change their clothing and footwear— those wearing glasses were given suitable metal-free replacements. First, participants were briefed about the experiment and asked to provide written informed consent. HPI coils were placed on their head and their head shapes and coil positions were registered as explained in section 2.4.1. Once inside the MEG room, electrodes of physiological sensors were attached to participants, and by checking the impedance level of the electrodes from the MEG recorder, we made sure that they were comfortable and were positioned correctly under the MEG sensor. Participants were provided with a desk pad, pillows and blanket to relax during the experiment. We then recorded five minutes resting state brain activity while the subject was fixating on a cross at the middle of the screen. Then, two practice trials (with the videos highlighted in blue in Fig 2.1, and denoted using ** in Table 2.1) were conducted to familiarize subjects with the protocol.

Each acquisition session involved a series of trials. During each trial, a fixation cross was first shown for four seconds to prepare the viewer and to gauge his/her rest-state response. Upon stimulus presentation, the subject conveyed the emotion elicited in him/her to the experimenter through the microphone. Ratings were acquired for (i) Arousal ('How intense is your emotional feeling on watching the clip?') on a scale of 0 (very calm) to 4 (very excited), (ii) Valence ('How do you feel after watching this clip?') on a scale of -2 (very unpleasant) to 2 (very pleasant), and (iii) Dominance on a scale of 0 (feeling empowered) to 4 (helpless). A maximum of 15 seconds was available to the participant to convey each rating. All in all, the whole experiment (spread over two sessions) including preparation time took about three hours per subject, who was paid a participation fee of €40.

2.5 Rating Analysis

2.5.1 Self-assessments: Music vs movie clips

As mentioned earlier, one objective behind compiling the DECAF database was to examine the effectiveness of different stimuli in eliciting similar emotional responses across subjects. In this section, we compare the self-assessment (or explicit) valence-arousal ratings for music and movie clips provided by the DECAF participants. Since self-reports are a conscious reflection of the user’s emotional state upon viewing the stimulus, one can expect any differences between the ratings for music and movie clips to also impact affect recognition from physiological responses.

Fig. 2.4 presents distributions of the V-A ratings provided by the 30 DECAF participants for movie and music clips. The blue, magenta, black and red colors respectively denote high arousal-high valence (HAHV), low arousal-high valence (LAHV), low arousal-low valence (LALV) and high arousal-low valence (HALV) stimuli as per the ground-truth ratings derived from Table 2.1 for movie clips and [57] for music videos. A U-shape, attributed to the difficulty in evoking low arousal but strong valence responses [60, 57], is observed for both movie and music clips. The ‘U’ bend is particularly pronounced in the case of music clips, implying that a number of stimuli were perceived to be close-to-neutral in valence, and there is considerable overlap among the four quadrants. For movie clips, perfect agreement with the ground-truth is noted for valence, but cluster overlap is observed along the arousal dimension.

We performed two-sample t -tests to check if the arousal characteristics of movie/music stimuli influenced their valence ratings— these tests revealed that valence ratings differed very significantly for HA music ($t(18) = 9.4208, p < 0.000001$), HA movie ($t(16) = 13.5167, p < 0.000001$) clips and LA movie clips ($t(16) = 11.586, p < 0.000001$), but somewhat less significantly for LA music clips ($t(18) = 5.6999, p < 0.00005$). Conversely, similar significance

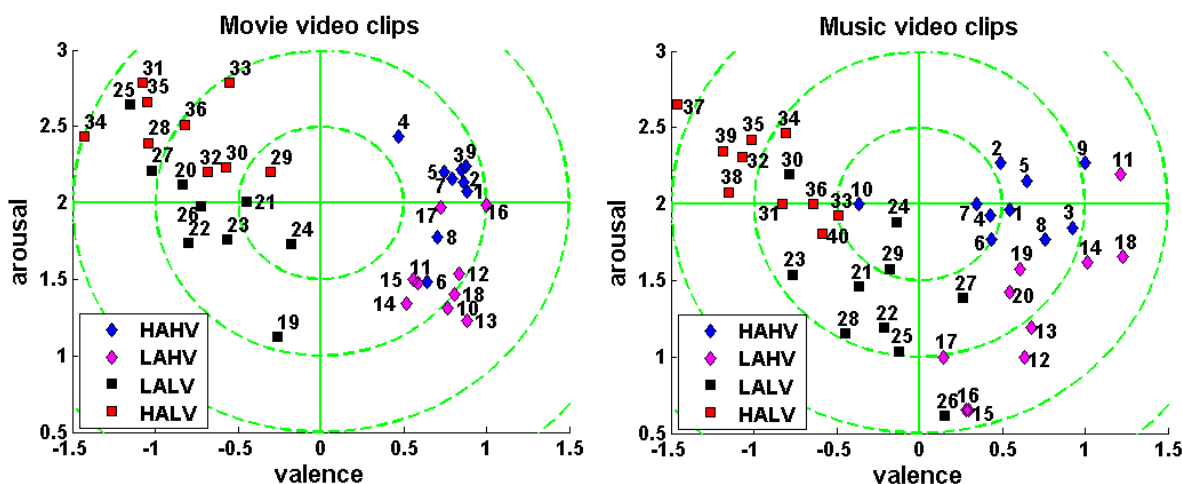


Figure 2.4: Mean V-A ratings for movie (left) and music clips (right) derived from DECAF participants.

levels were observed while comparing arousal ratings for HV music ($t(18) = 4.2467, p < 0.0005$) and movie ($t(16) = 4.2988, p < 0.0005$), as well as LV music ($t(18) = -4.8256, p < 0.005$) and movie ($t(16) = -3.3194, p < 0.005$) stimuli. Overall, the valence-arousal distinction was slightly better for movie vis-à-vis music clips.

To evaluate how consistently emotional responses were elicited across subjects, we measured agreement between the ground-truth and participant ratings using the Cohen’s Kappa measure assuming that ground-truth V-A labels were provided by an ‘ideal’ annotator. To this end, we assigned high/low V-A labels to the stimuli based on each user’s median ratings, and computed κ between the ground-truth and user judgements. The mean κ over all subjects for music-valence, movie-valence, music-arousal and movie-arousal were found to be 0.50 ± 0.17 , 0.67 ± 0.24 , 0.14 ± 0.17 and 0.19 ± 0.17 respectively. Agreement with the ground-truth was higher for movie stimuli, implying that movie stimuli evoked intended emotions more consistently across users. Also, agreement was considerably higher for valence, indicating stronger differences in arousal perception across subjects.

2.6 Data Analysis

This section describes the procedure for data preprocessing and feature extraction from (i) MEG signals, (ii) physiology signals, (iii) face videos and (iv) multimedia signals. All the cut-off frequencies and smoothing parameters employed were adopted from [56, 109, 57]. For both MEG and peripheral physiological modalities, we computed (1) time-continuous features for dynamic emotion analysis and (ii) statistical measures⁷ computed over the time-continuous features, considering only the final 50 seconds.

2.6.1 MEG preprocessing and feature extraction

MEG preprocessing involved three main steps, (i) Trial segmentation, (ii) Spectral filtering and (iii) Channel correction, that were handled using the MATLAB Fieldtrip toolbox [83]. Since magnetometer outputs are prone to environmental and physiological noise, we only used the gradiometer outputs for our analysis.

Trial Segmentation: Participant responses corresponding to each trial were extracted by segmenting the MEG signal from 4 seconds prior to stimulus presentation (pre-stimulus) to the end of stimulus. Per subject, there were 36 and 40 trials for the movie clips and music videos respectively.

Frequency domain filtering: Upon downsampling the MEG signal to 300 Hz, low-pass and high-pass filtering with cut-off frequencies of 95 Hz and 1 Hz respectively were performed. The high-pass filter removes low frequency ambient noise in the signal (*e.g.*, generated by moving vehicles). Conversely, the low-pass filter removes high frequency artifacts generated by muscle activities (between 110-150 Hz).

Channel correction: Dead and bad channels were removed from the MEG data. Dead channels output zero values, while bad channels are outliers with

⁷mean (μ), standard deviation (σ), skewness, kurtosis, percentage of values above $\mu + \sigma$, and percentage of values below $\mu - \sigma$

respect to metrics such as signal variance and signal amplitude z -score over time. To preserve the dimensional consistency of MEG data over all trials and subjects, removed channels were replaced with interpolations from neighboring channels.

Time-Frequency analysis (TFA): The spectral power in certain frequency bands has been found to contain valuable information for affect recognition in a number of EEG studies. The multitaper and wavelet transforms are typically used in order to achieve better control over frequency smoothing, and high frequency smoothing has been found to be beneficial when dealing with brain signals above 30 Hz [77]. Therefore, we used variable-width wavelets to transform the preprocessed MEG signal to the time-frequency domain for spectral power analysis.

MEG-TFA Features: We used a time-step of 1s for temporal processing of the MEG signal from each trial, and a frequency step of 1 Hz to scan through a frequency range of 1-45 Hz. We linearly varied the wavelet width with frequency, increasing from 4 for lower frequencies to 8 for higher frequencies. Upon applying a wavelet transform on the MEG data, we performed the following steps: (a) We used a standard Fieldtrip function for combining the spectral power of each planar gradiometer pair to obtain 102 combined-gradiometer (GRAD) responses. (b) In order to better elucidate the MEG response dynamics following stimulus presentation for each subject, individual trial power was divided by a *baseline* power, obtained as the mean over two seconds pre-stimulus from all trials. (c) To increase dynamic range of the spectral power, the time-frequency output was logarithm transformed.

Channel Grouping: On computing the MEG spectral power over 102 GRAD pairs, in order to reduce data dimensionality while preserving spatial information, the 102 channels were divided into nine groups according to functionality of different brain regions namely: Vertex, left temporal, right temporal, left parietal, right parietal, left occipital, right occipital, left frontal and right frontal

(Fig. 2.5). The sensors in each group encode different brain functionalities that may directly or indirectly relate to emotions, and we show that this grouping is beneficial for affect recognition in Sec. 2.8. Per subject and movie/music clip, time-frequency analysis outputs nine (one per group) 3D matrices with the following dimensions: $K \times \text{clip length time points} \times 45 \text{ frequencies}$, where K denotes the number of GRAD channels per group.

DCT features: The Discrete Cosine Transform (DCT) is often used in signal, image and speech compression applications due to its strong energy compaction ability. Also, the DCT feature space has been shown to efficiently compress spatio-temporal patterns of MEG data without impacting model precision [54]. We employed DCT to compress the MEG-TFA output on a per-second basis, as well as for single-trial classification. Per second, from each of the 9 lobes we extracted 60 DCT coefficients (4 along spatial and 15 along spectral respectively), and concatenated them to extract 540 DCT features. For single-trial classification, from each brain lobe, we used the first $n = 2$ DCT coefficients from the spatial, temporal and spectral dimensions to obtain a total of $9 \times 8 = 72$ features. We observed that classification results did not improve with $n > 2$ DCT coefficients per dimension— this could be attributed to the fact that our model training involves much fewer examples as compared to the feature dimensionality.

2.6.2 Peripheral physiological feature extraction

hEOG features

The horizontal EOG signal has information about eye movements, point-of-gaze and eye blinks. Muscular facial activities and eye blinks appear as high frequency components in the EOG signal. Eye movements, blinks and facial muscular activities have been found to be highly correlated with emotional responses [57, 109].

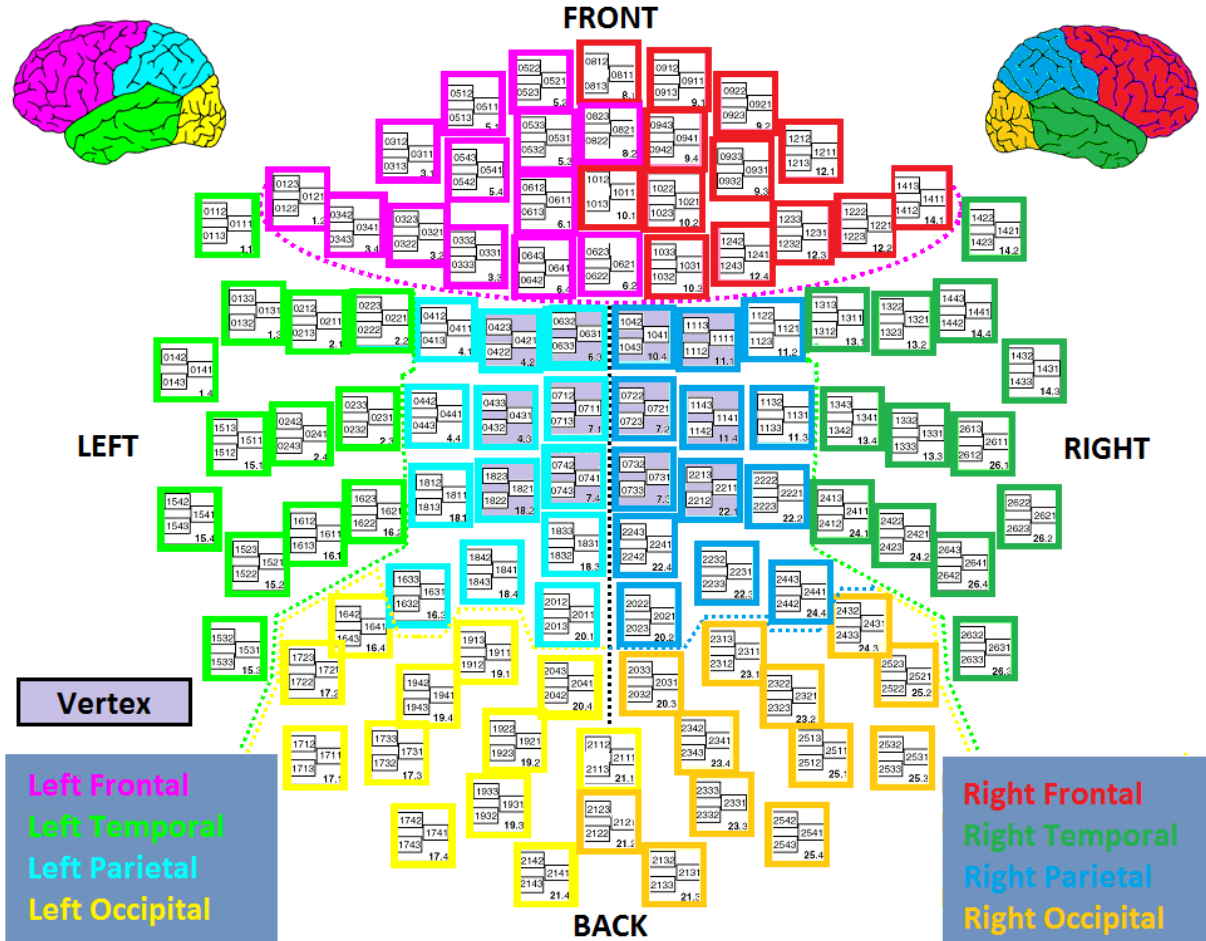


Figure 2.5: Elekta Neuromag MEG channel positions. Channels corresponding to different lobes are color-coded (figure adapted from www.megwiki.org, best viewed under zoom).

Eye movements: To extract eye movement information, we low-pass filtered the signal with 5 Hz cut off, and then used wavelet transform to extract power spectral density (PSD) in 0-2 Hz range with a frequency resolution of 0.2 Hz, and temporal resolution of 50ms. Then for each second, we averaged the PSD values over frequency ranges of $\{[0, 0.1), [0.1, 0.2), [0.2, 0.3), [0.3, 0.4), [0.4, 0.6), [0.6, 1.0), [1.0, 1.5), [1.5, 2)\}$. Therefore, we obtained 8 features per second to describe eye movements.

Facial muscle activity: Facial muscular activities mainly relate to the movement of zygomatic major muscles, which occurs when a subject exhibits a smile,

frown or other facial expressions. We limited the signal to 105-145 Hz, and then used wavelet transform to extract PSD with a frequency resolution of 1 Hz and temporal resolution of 500 ms.

Then for each second, we averaged the PSD values over $\{[105, 115), [115, 130), [130, 145)\}$ frequency ranges. Since there are many muscles controlling facial activities, we used the three bands to obtain fine-grained information regarding muscular activities. Therefore per second, we obtained three values to represent zygomatic activities. Overall, from hEOG, we obtained 11 vectors of clip-length duration.

ECG features

From the ECG signal, we extracted information from both the original signal and its PSD.

Heart beats: We detected heart beats through R-peak detection in the ECG signal. Upon removal of low frequency components, R-peaks were detected as the amplitude peaks. We then computed inter-beat-intervals (IBI), heart rate (HR) and heart rate variability (HRV) as the derivative of HR. Upon smoothing HR with a Kaiser window of temporal width 10 sec, and shape parameter $\beta = \frac{1}{6}$, we computed two features (smoothed HR and HRV) per second from which, statistical measures over IBI, smoothed HR, and HRV during the final 50 seconds of each trial were derived for affect recognition.

Power spectral density: ECG was recorded at 1 KHz sampling rate, and we used a wavelet transform over the ECG signal to extract the PSD in the frequency range of 0-5 Hz. Then, the mean PSD magnitudes over the frequency intervals $\{(0, 0.1], (0.1, 0.2], (0.2, 0.3], (0.3, 0.4], (0.4, 0.5], (0.5, 0.6], (0.6, 1], (1, 1.5], (1.5, 2], (2, 2.5], (2.5, 5.0)\}$ were used as features— this gave us 11 values per second.

For single-trial classification alone, additional low-frequency information characterizing emotions was extracted as in [57]. We downsampled the ECG signal

from 1 KHz to 256 Hz, and removed the low frequency drift. Then, we estimated the signal PSD using Welch’s method with a window length of $15 \times sr$ and the overlap of $10 \times sr$, where sr denotes signal sampling rate. We used the mean PSD over $\{[0, 0.1), [0.1, 0.2), [0.2, 0.3), [0.3, 0.4]\}$ bands, and the logarithm PSD obtained for the sub-bands obtained on dividing $[0, 2.4]$ into 10 equal intervals to obtain 14 more ECG PSD features.

Trapezius EMG

EMG effectively captures the mental stress of users [104]. As bipolar EMG electrodes are placed above the trapezius muscle, heart-related artifacts are observed in the signal and the EMG signal consists of two components: (1) Heart activities such as heart beats can be mainly inferred from the 0-45 Hz range, and (2) Trapezius EMG can be obtained from the $\{[55, 95), [105, 145)\}$ range.

Heart activities: We low-passed the signal to within 45 Hz, and used wavelet transform to extract the PSD map with frequency and temporal resolution of 0.2 Hz and 50 ms respectively. Per second and trial, we computed the mean PSD over the following frequency bands: $\{[0, 0.5), [0.5, 1.5), [1.5, 2.5), [2.5, 3.5), [3.5, 5.0), [5.0, 10), [10, 15), (15, 25), [25, 45)\}$, to describe heart activities when the ECG signal was unavailable.

Muscle activities: We band-passed the EMG signal between 55-145 Hz and employed wavelet transform to extract the PSD map with frequency resolution of 1 Hz, and temporal resolution of 500 ms. Per each second and trial, we computed two values corresponding to mean PSD over the $\{[55, 95), [105, 145)\}$ frequency bands to characterize trapezius muscle activities, and aforementioned statistical measures over the final 50 seconds were used for affect recognition.

2.6.3 Facial Expression Analysis

We used histogram equalization to enhance contrast in the recorded NIR facial videos, and then employed the facial tracker described in [41] to track 12 facial landmarks (Figure 2.6). Statistical measures over the activation of these landmarks in the final 50 seconds of each trial were used for classification.

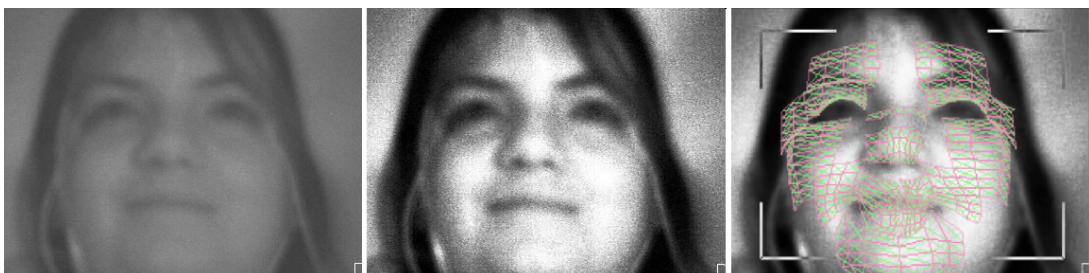


Figure 2.6: Participant’s facial video before (left) and after (middle) histogram equalization. Tracking 3D grid is shown on the right.

2.6.4 Multimedia features

We computed low-level audio visual features from the movie and music clips as described in [57] for comparing different modalities, and identifying the salient emotional information sources— extracted features are listed in Table 2.2. All in all, 49 video features and 56 audio features were extracted. For single-trial classification, we computed statistics over 1-second segments, while using statistics from features computed at the frame level for fine-grained, per-second emotion estimation described in Sec. 2.9.

2.7 MEG correlates with user ratings

We now present correlations observed between users’ self-assessments and their MEG responses. In order to directly compare our results with [57], we performed MEG feature extraction identical to [57] briefly described as follows.

2.7. MEG CORRELATES WITH USER RATINGS

Table 2.2: Extracted audio-visual features from each movie clip (feature dimension listed in parenthesis).

Audio features	Description
MFCC features (39)	MFCC coefficients [66], Derivative of MFCC, MFCC Autocorrelation (AMFCC)
Energy (1) and Pitch (1)	Average energy of audio signal [66] and first pitch frequency
Formants (4)	Formants up to 4400Hz
Time frequency (8)	mean and std of: MSpectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio [66]
Zero crossing rate (1)	Average zero crossing rate of audio signal [66]
Silence ratio (2)	Mean and std of proportion of silence in a time window [66, 19]
Video features	Description
Brightness (6)	Mean of: Lighting key, shadow proportion, visual details, grayness, median of Lightness for frames, mean of median saturation for frames
Color Features (41)	Color variance, 20-bin histograms for hue and lightness in HSV space
VisualExcitement (1)	Features as defined in [126]
Motion (1)	Mean inter-frame motion [69]

Following artefact rejection, we downsampled the MEG signal to 256Hz and then band-limited the same to within 1-48 Hz. Upon combining gradiometer outputs, the spectral power between 3 and 47 Hz over the last 30 seconds of each clip was extracted using Welch’s method with a window size of 256 samples. Mean power over the θ ([3-8] Hz), α ([8-14] Hz), β ([14-30] Hz) and γ ([30-45] Hz) for each of 102 MEG sensors were correlated with the users’ self-assessments.

We computed Spearman correlations between the above MEG-PSD outputs and participants’ self ratings. Following [57], per subject, trial, emotion dimension and frequency band, correlations were computed over the 102 combined

GRAD outputs. Upon computing correlations for each subject, and assuming independence [61], p -values obtained for each subject and condition were fused over all users using Fisher’s method. Different from [57], we also accounted for multiple comparisons by controlling false discovery rate (FDR) using the procedure proposed in [11], and the observed significant correlations are highlighted in Fig. 2.7 ($p < 0.05, 0.01, \text{ and } 0.001$ are respectively denoted in cyan, magenta, and red).

Observations: Observations similar to [57] can also be noted from Fig. 2.7. Thanks to the higher spatial resolution of MEG, a greater number of significant correlates and a wider range of correlations ($[-0.15, 0.25]$ with MEG vs $[-0.1, 0.1]$ with EEG) are observed with MEG signals as compared to EEG. For both movie and music stimuli, we observe a negative correlation between α , β and γ powers and the arousal level over the vertex, the parietal and occipital lobes, which is consistent with the findings in [57]. Over the temporal and occipital lobes, we observe a positive correlation between the θ , β and γ powers and the valence level. Note that the occipital and temporal lobes encode low-level audio-visual information which are responsible for inducing emotions [126]. The possibility of facial muscle activities, which are also prominent at high frequencies, influencing the observed correlations between valence/arousal ratings and MEG responses is minimal as facial activities are likely to occur in response to both negative and positive valence stimuli (*e.g.*, funny and disgust). Finally, a few significant negative correlates in the parietal lobe, and few positive correlates in the occipital lobe are observed between dominance ratings and the MEG β, γ powers.

Movie vs music: As evident from Fig. 2.7, larger and more significant correlations are observed for movie clips as compared to music video clips, which suggests that emotions are more strongly and consistently evoked by movie stimuli. In particular, no correlations with $p < 0.001$ are observed for music videos for the arousal and dominance dimensions. However, a larger number of

correlations are observed over all frequency bands for arousal with music clips. We mention here that some of the detectable correlates for movie stimuli may have arisen from extraneous factors— *e.g.*, correlates between θ , α powers and valence ratings may be attributed to eye movements/blinks. Likewise, positive correlation between γ power and dominance over the occipital lobes could be explained by low-level visual cues [79], while the similar but weaker correlate observed for arousal could be owing to the strong positive correlation between arousal and dominance ratings (0.57 ± 0.24) across participants. Further examination to more accurately identify the information source responsible for the above correlations would involve (1) HPI-based MEG signal compensation, (ii) Independent component analysis, and (iii) Brain source localization using MR brain scans, which is left to future work.

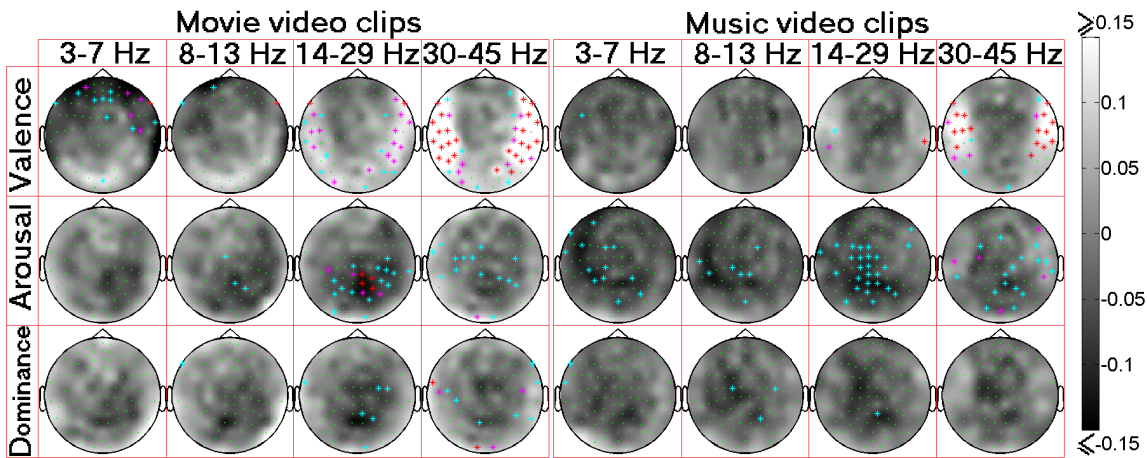


Figure 2.7: Spearman correlation analysis between the MEG responses and participants’ self-assessments. Correlation over each channel (in green) is denoted by the gray level, and significant ($p < 0.05$, $p < 0.01$, and $p < 0.001$) correlations are highlighted with * marks (in cyan, magenta, and red).

2.8 Experimental Results

We now present comparisons between MEG vs EEG, and movie vs music clips based on single-trial classification results.

Table 2.3: Mean binary classification performance for music-video clips with the schema described in [57]. F1-scores of distributions significantly over 0.5 are highlighted (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). NR denotes 'not reported'.

	Music (SS)					
	Arousal		Valence		Dominance	
	Acc	F1	Acc	F1	Acc	F1
EEG [57]	0.62	0.58**	0.58	0.56**	NR	NR
Max Baseline [57]	0.64	0.50	0.59	0.50	NR	NR
MEG	0.62	0.58***	0.59	0.55*	0.62	0.53*
Max Baseline	0.52	0.50	0.54	0.50	0.66	0.50

2.8.1 Single-trial Classification: MEG versus EEG

In order to evaluate our MEG-based approach against the EEG framework described in [57], we attempted single-trial binary (*high/low*) classification of valence and arousal employing (i) labels derived from subject-wise self-reports and (ii) extracting MEG features in a manner identical to [57]. Employing the Naive-Bayes classifier and subject-specific models, only the top 10% discriminative features based on Fisher feature selection criteria were used in each loop of a leave-one-trial-out cross-validation scheme. Very comparable results with EEG and MEG obtained with this procedure (Table 2.3) suggest that the affect encoding power of EEG and MEG are comparable. However, the increased spatial resolution of MEG allows for fine-grained affective analysis, which enables similar or superior recognition performance on music and movie clips using the features extracted in Sec. 2.6 as described later.

While the fairest comparison between EEG and MEG would entail simultaneous recording of the two modalities for identical subjects and stimuli, such a study may be impossible to implement in practice. We have compared emotion recognition performance based on the results observed on two random subject populations that are comparable in size, and this is the second best possible way of performing a comparison in our view. Designing better approaches for com-

paring the efficacy of different modalities for user-centric emotion recognition is a research problem requiring further investigation.

2.8.2 Classification procedure and results

On a per-user basis, we attempted to recognize the emotional *valence* (V), *arousal* (A) and *dominance* (D) of a test music/movie clip as *high/low* based on the MEG and peripheral physiological responses. Given the large subjectivity in user responses for music videos in [57], subject-specific labels were used for each stimulus. However, as (i) many significant correlates observed between ratings and MEG responses of the user population, and (ii) the stimulus label should reflect the perception of the population instead of individuals, we repeated the classifications with both population-based (denoted as PB in Table 2.4) and subject-based (SB in Table 2.4) labels.

Under PB labeling, each stimulus was assigned a *high/low* (V/A/D) label based on whether its rating was higher or lower than the mean rating provided by the participant population for the stimulus set. Likewise, the SB label for each stimulus denoted whether its rating was higher/lower than the mean subject rating. The proportion/distribution of positive and negative classes for movie and music V,A,D under PB/SB tagging is presented in Table 2.4. For SB labeling, the mean and standard deviation of the positive class distribution are specified. Under PB labeling, the proportion of positive and negative classes is most imbalanced for music and movie arousal, whereas the most balanced distributions under SB labeling are observed for movie valence and music arousal. Given the unbalanced positive and negative classes, we use F1-scores as the primary measure to compare classification performance with different stimulus types and information modalities.

We used a linear SVM classifier for our experiments and the mean accuracy and F1-scores obtained over the 30 participants using leave-one-trial-out cross-validation are tabulated in Table 2.4. The optimal SVM slack param-

ter was tuned by considering values in $[10^{-4}, 10^4]$ using an inner leave-one-out cross-validation loop. As baselines, we present the F1-scores of (i) a random classifier, (ii) majority-based voting⁸ and (iii) voting based on training class distribution— note that the maximum baseline F1-score is 0.50. Instances where the F1-score distribution across subjects is significantly higher than 0.5 as determined by a paired t -test are highlighted in Table 2.4.

To demonstrate how the higher spatial resolution of MEG benefits affect recognition, we present results achieved with features extracted exclusively from each brain lobe, and also the concatenation of features from all lobes (MEG Early Fusion or MEF). In addition, we present accuracies and F1-scores achieved using (i) the combination of hEOG, ECG and tEMG responses (peripheral physiology or PP), (ii) facial expressions (FE), (iii) multimedia features (MM), and (iv) late fusion of the decisions from the the MEF, PP, FE and MM classifiers following the methodology proposed in [58]. If $\{p_i\}_{i=1}^4$ denote the posterior probabilities output by the four classifiers and $t_i = \alpha_i F_i / \sum_{i=1}^4 \alpha_i F_i$, where α_i 's denote fusion weights and F_i denotes F1-score of the i^{th} classifier on training data, the optimal weights $\{\alpha_i^*\}$ are chosen as those maximizing F1-score on the training set using an inner cross-validation loop. Posterior probability of the test sample is computed as $\sum \alpha_i^* p_i t_i$, which is then used to assign the test label.

2.8.3 Discussion of classification results

In Table 2.4, the obtained F1-scores clearly demonstrate that the increased spatial resolution of MEG benefits affect analysis and recognition. For all conditions, the classification performance obtained with MEG features from at least one of the nine brain lobes is similar to or better than the performance achieved with MEF, where features of all the brain lobes are pooled together. This result

⁸With leave-one-out classification on a balanced class distribution (Table 2.4), majority-based voting would yield 0% accuracy as the test-label class is in minority in the training set.

2.8. EXPERIMENTAL RESULTS

Table 2.4: **Single trial classification for music and movie clips**– (Upper) classification results using MEG information from each of the brain lobes. (Middle) Unimodal and multimodal classification results. (Bottom) Baseline comparisons along with the distribution of positive samples are tabulated. Mean F1 scores derived from a distribution significantly above chance level (0.50) are highlighted (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$). PB, SB respectively denote use of population and subject-based labels in the classification framework.

		Movie (PB)			Music (PB)			Movie (SB)			Music (SB)		
		A	V	D	A	V	D	A	V	D	A	V	D
Vertex	Acc	0.59	0.57	0.57	0.51	0.51	0.52	0.55	0.55	0.51	0.53	0.50	0.53
	F1	0.58***	0.57***	0.57***	0.51	0.51	0.51	0.54	0.53	0.48	0.52	0.49	0.49
Left Temporal	Acc	0.60	0.60	0.58	0.51	0.51	0.52	0.59	0.58	0.51	0.54	0.50	0.54
	F1	0.60***	0.60***	0.58***	0.51	0.51	0.51	0.59***	0.57**	0.49	0.52	0.49	0.51
Right Temporal	Acc	0.62	0.56	0.57	0.55	0.53	0.53	0.59	0.55	0.54	0.60	0.54	0.54
	F1	0.62***	0.55**	0.57***	0.55*	0.53*	0.53*	0.58**	0.53	0.51	0.58***	0.53	0.51
Left Parietal	Acc	0.60	0.56	0.57	0.52	0.52	0.55	0.55	0.56	0.53	0.53	0.48	0.52
	F1	0.60***	0.55**	0.57***	0.52	0.51	0.54*	0.54*	0.54*	0.49	0.52	0.47	0.49
Right Parietal	Acc	0.58	0.57	0.57	0.51	0.51	0.52	0.55	0.55	0.58	0.51	0.53	0.54
	F1	0.57**	0.57***	0.56***	0.50	0.50	0.52	0.53	0.53	0.55**	0.50	0.52	0.51
Left Occipital	Acc	0.58	0.59	0.57	0.51	0.50	0.52	0.53	0.56	0.54	0.55	0.48	0.53
	F1	0.57**	0.58***	0.56**	0.51	0.50	0.52	0.51	0.54*	0.50	0.54*	0.47	0.50
Right Occipital	Acc	0.60	0.56	0.56	0.50	0.53	0.50	0.57	0.54	0.55	0.54	0.53	0.53
	F1	0.60***	0.55**	0.56*	0.50	0.53	0.50	0.56**	0.53	0.52	0.53	0.51	0.49
Left Frontal	Acc	0.59	0.56	0.57	0.55	0.51	0.51	0.56	0.56	0.53	0.57	0.55	0.60
	F1	0.58***	0.56***	0.57***	0.54*	0.50	0.51	0.55**	0.55**	0.50	0.55**	0.54*	0.56**
Right Frontal	Acc	0.55	0.59	0.61	0.50	0.52	0.50	0.51	0.54	0.53	0.54	0.52	0.53
	F1	0.55***	0.59***	0.61***	0.49	0.52	0.49	0.50	0.53	0.49	0.53	0.51	0.49
MEG Early Fusion	Acc	0.60	0.61	0.59	0.53	0.53	0.54	0.55	0.58	0.55	0.58	0.56	0.55
	F1	0.60***	0.61***	0.59***	0.52	0.53	0.54*	0.54*	0.58***	0.53	0.55**	0.55**	0.53*
Peripheral Physiology	Acc	0.55	0.60	0.50	0.55	0.59	0.56	0.56	0.60	0.56	0.57	0.55	0.57
	F1	0.54*	0.59***	0.50	0.54*	0.59***	0.55**	0.55**	0.59***	0.54*	0.56**	0.54*	0.54**
Facial Expressions	Acc	0.58	0.64	0.53	0.60	0.61	0.53	0.56	0.61	0.55	0.58	0.60	0.55
	F1	0.57**	0.64***	0.53	0.59**	0.60***	0.53	0.54**	0.61***	0.54	0.56**	0.58***	0.52
Multimedia Content	Acc	0.58	0.64	0.33	0.85	0.73	0.57	0.52	0.61	0.53	0.62	0.68	0.58
	F1	0.57	0.64	0.33	0.85	0.72	0.57	0.51	0.60***	0.52	0.61***	0.67***	0.55*
Late Fusion	Acc	0.70	0.79	0.66	0.85	0.82	0.66	0.66	0.73	0.72	0.73	0.76	0.74
	F1	0.68***	0.77***	0.64***	0.84***	0.81***	0.65***	0.62***	0.71***	0.66***	0.70***	0.73***	0.67***
Random	Acc	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
	F1	0.49	0.50	0.50	0.49	0.50	0.50	0.49	0.49	0.48	0.49	0.49	0.48
Majority	Acc	0.58	0.00	0.53	0.57	0.53	0.00	0.57	0.53	0.60	0.52	0.54	0.66
	F1	0.37	0.00	0.35	0.37	0.34	0.00	0.37	0.33	0.36	0.32	0.34	0.39
Class-ratio	Acc	0.51	0.50	0.50	0.51	0.50	0.50	0.54	0.52	0.56	0.52	0.53	0.57
	F1	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50
+ve Class proportion	Mean	58.3%	50.0%	52.8%	57.5%	52.5%	50.0%	48.4%	49.3%	41.9%	49.3%	46.3%	45.6%
	STD	-	-	-	-	-	-	13.6%	9.5%	14.9%	10.7%	10.9%	19.0%

is unsurprising as the various brain lobes are known to encode different types of emotional information, as also suggested by the correlation analysis in Sec. 2.7. Under PB stimulus labeling, the best F1-scores for movie and music arousal are obtained for the right temporal lobe, while the left and right temporal lobes respectively are found to encode optimal information for decoding the valence of movie and music stimuli. Best performance for dominance is obtained with right-frontal lobe features for movies, and left parietal for music.

Another salient observation is that despite the subjectivity in emotion perception and expression, reliable and above-chance emotion recognition is achieved upon associating the physiological responses of each user with stimulus labels assigned by the population. For movie clips in particular, much better classification performance achieved under PB labeling as compared to SB labeling. In practice, emotion (or genre) tags to movies or music videos are attached based on the perception of the *general audience*, and not on the basis of individual perception. Likewise, for the purpose of affect recognition and emotion elicitation, it would be desirable to work with control stimuli consistently capable of evoking the target emotion from target users. Movie clips (and corresponding user responses) compiled as part of DECAF are an important contribution in this respect.

The obtained results also point to the complementarity of different signals in encoding emotions. Consistent with the findings in [57], MEG signals are seen to effectively encode arousal and dominance, while peripheral physiology signals efficiently encode valence. Facial expressions are also seen to best encode valence, while audio-visual features achieve best arousal recognition for music clips with PB labels. This complementarity was also evident when finding the best two and three information modalities for recognizing valence and arousal under PB labeling— considering feature pairs, MEG and peripheral physiological features produced the best arousal recognition for movie clips ($F1=0.66^{***}$), while peripheral and audio-visual features best recognized valence from mu-

sis clips ($F1=0.83^{***}$). Facial activities and multimedia content provided best recognition of valence from movies ($F1=0.78^{***}$) and arousal from music clips ($F1=0.87^{***}$). Considering triplets, the combination of MEF, PP and MM consistently produced the best F1-scores for movie-arousal ($F1=0.71^{***}$), movie-valence ($F1=0.81^{***}$), music-arousal ($F1=0.87^{***}$), music-valence ($F1=0.85^{***}$). F1-scores obtained by fusing the outputs of all modalities are slightly lower than those obtained from combinations of feature triplets, suggesting that feature selection may be necessary for optimal fusion results.

Finally, comparing the emotion recognition performance with music and movie clips, superior F1-scores achieved using MEG features for population-rated movie clips again confirms that they serve as better control stimuli for affect recognition studies. For music stimuli, relatively higher recognition is achieved with subject-specific labels, and the best performance with PB labels is achieved for arousal using multimedia features.

2.9 Continuous Emotion Estimation

DECAF also contains time-continuous arousal (A) and valence (V) annotations for the 36 movie clips acquired from seven experts, who were very familiar with the movie clips, but were not part of the MEG study. While the user ratings acquired in Sec. 2.4 are useful for recognizing the *general* stimulus emotion, dynamic V-A ratings are used for estimating the *emotional highlight* in a given clip. We show how these annotations were utilized to predict V-A levels of time-contiguous snippets using (i) multimedia audio-visual (MM), and (ii) MEG features.

Experiments and Results: We asked seven experts to provide per-second V-A ratings for 36 movie clips listed in Table 2.1 using the G-Trace software [22]. The experts, who could familiarize themselves with scene dynamics by viewing the movie clips as many times as they wanted to prior to rating them, were re-

quired to annotate the *target emotion* meant to be evoked in the viewer (in terms of V-A levels) for each second of the video. Upon rescaling the annotations using z -score normalization, Kendall’s coefficient of concordance (W) was used to measure the dynamic inter-annotator agreement— overall W was found to be 0.47 ± 0.27 for arousal, and 0.64 ± 0.18 for valence, signifying good agreement. Re-computing W over the *first* and *second* half of the clips, we observed W to be 0.35 ± 0.25 , 0.43 ± 0.28 and 0.58 ± 0.24 , 0.54 ± 0.23 for V-A respectively, implying that expert assessments were more consistent for the emotionally salient second halves of the clip (all clips began with a neutral segment). Finally, the median annotation was used as the **gold standard** dynamic rating for each clip. Dynamic V-A ratings are illustrated in Fig. 2.8.

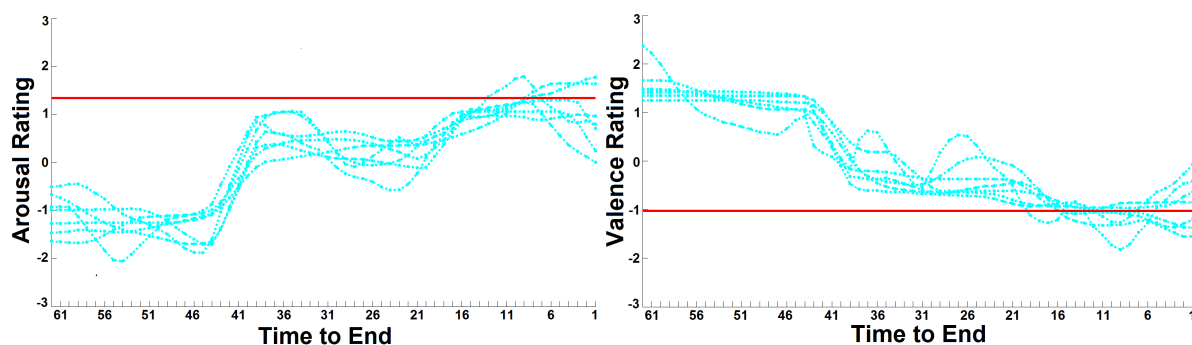


Figure 2.8: Time-continuous A (left), V (right) ratings for Clip 36 in Table 2.1 from seven experts are plotted in cyan. Both continuous and static ratings (red) are z -score normalized and are in the range $[-3, 3]$.

We then attempted prediction of dynamic V-A levels in time-contiguous snippets derived from the movie clips using (i) audio-visual and (ii) MEG features. Per-second features extracted in Sec. 2.6 were used to this end. Apart from Lasso sparse regression, we also employed Multi-task learning (MTL) based regressors— given a set of T related tasks (movie clips related in terms of V-A in this case), MTL [133] seeks to *jointly* learn a set of weights $W = \{W_t\}_{t=1}^T$, where W_t models task t . MTL enables simultaneous learning of similarities as well as differences among tasks, leading to a more efficient model

than learning each task independently. In this work, we employed three MTL variants from the MALSAR library [142]– multi-task Lasso, Dirty MTL where the weight matrix $W = P + Q$, with P and Q denoting group-common and task-specific components, and sparse graph-regularized MTL (or SR MTL), where *a priori* knowledge on task-relatedness is incorporated in the learning process so that weight similarity is only enforced among related tasks.

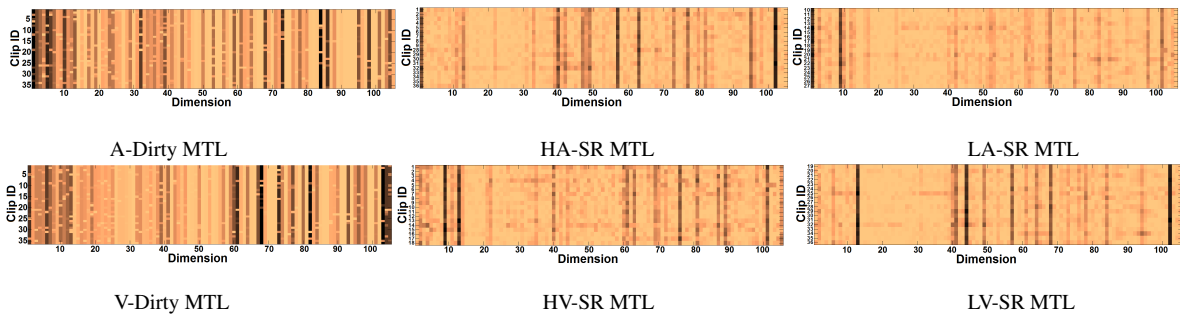


Figure 2.9: Learned weights for arousal (top) and valence (bottom) for the movie clips with Dirty MTL and SR MTL. Audio-visual features over the entire clip length were used for model training. Larger weights are denoted using darker shades. MM features (106 in total) are arranged in the order specified in Sec. 2.6. Best viewed under zoom.

V-A weights for the 36 movie clips learned from audio-visual (MM) features (concatenation of audio and video features) through the Dirty and SR MTL approaches are presented in Fig. 2.9. *A-priori* knowledge available in the form of ground truth labels (Table 2.1) were used to group related stimuli and input to the SR MTL algorithm. SR MTL weights learnt for high and low arousal clips are shown in the top row, while the bottom row presents weights learned for high and low valence clips. MFCCs are found to be the most salient audio features, while color and brightness video features are the best predictors for both valence and arousal. Concerning SR MTL outputs, visual excitement features are found to be characteristic of high arousal clips, while inter-frame motion is indicative of high-valence clips.

Finally, dynamic V-A level prediction performance using MM and MEG features (average MEG response of the 30 DECAF participants was used here) on

Table 2.5: Valence/Arousal prediction with multimedia (MM) and MEG features. RMSE mean, standard deviation over four runs are reported. Range of V-A levels is [-3, 3]. Best model is shown in bold.

		First		Second		
		5 s	15 s	5 s	15 s	
Valence	MM	Lasso	1.98±1.25	3.07±1.48	1.68±0.18	2.81±0.97
		MT-Lasso	1.00±0.05	1.66±0.54	1.18±0.14	2.03±0.71
		Dirty MTL	1.11±0.06	1.79±0.55	1.27±0.16	2.10±0.69
		SR MTL	1.09±0.09	1.55±0.39	1.89±0.13	2.80±0.74
	MEG	Lasso	1.30±0.09	1.87±0.46	2.03±0.25	2.93±0.78
		MT-Lasso	1.32±0.09	1.98±0.54	1.54±0.21	2.47±0.81
		Dirty MTL	1.42±0.10	2.44±0.82	1.51±0.19	2.44±0.82
		SR MTL	1.09±0.05	1.58±0.41	2.07±0.17	2.84±0.69
Arousal	MM	Lasso	1.54±0.47	2.11±0.77	2.18±0.58	3.28±2.17
		MT-Lasso	0.91±0.11	1.47±0.47	1.10±0.08	1.89±0.66
		Dirty MTL	1.07±0.09	1.62±0.46	1.23±0.08	1.97±0.61
		SR MTL	1.01±0.07	1.42±0.35	1.86±0.13	2.48±0.53
	MEG	Lasso	1.11±0.08	1.65±0.45	1.75±0.06	2.53±0.66
		MT-Lasso	1.12±0.09	1.71±0.51	1.41±0.11	2.27±0.73
		Dirty MTL	1.19±0.11	1.84±0.56	1.38±0.11	2.25±0.75
		SR MTL	0.99±0.08	1.42±0.36	1.73±0.06	2.44±0.60

5 and 15 second snippets randomly extracted from the first and second half from each of the movie clips is presented in Table 2.5– remainder of the movie clips was used for model training. The root mean square error (RMSE) measure is used for comparison– evidently, larger prediction errors are noted for snippets from the second half, and for 15-sec segments. MTL considerably outperforms Lasso regression, implying that jointly learning from features of multiple movie clips is beneficial as compared to clip-wise learning, while slightly better prediction performance is achieved with MM features considering the best model for each condition.

2.10 Conclusion

The DECAF database compiled with the aim of evaluating user-centered affect recognition with (i) MEG vs EEG sensing, and (ii) movie vs music clips, is

presented in this chapter. The increased spatial resolution of MEG enables fine-grained analysis of cognitive responses over brain lobes in turn aiding affect recognition, while coherence between explicit ratings and implicit responses is greater across users for movie clips, suggesting that they are better control stimuli for affect recognition studies. While classification results for valence, arousal and dominance are presented with the aim of comparing with [57], dominance may be hard to qualify in a movie-watching context even if it has been found to be relevant with regard to musical compositions. This study was limited to sensor-space analyses of MEG responses— source-space analysis was not performed, and is left to future work. Finally, dynamic emotion prediction with time-continuous emotion annotations available as part of DECAF is demonstrated, and simultaneously learning from multimedia/MEG features from all clips is found to be more beneficial than learning one model per clip. Unlike EEG, MEG is a relatively new technology, and with improvements in techniques such as HPI-based MEG signal compensation, we believe that much higher recognition performance than that achieved in this introductory work is possible.

Chapter 3

Emotion and Personality Recognition using Commercial Sensors

While the emotional state of humans can change often, their personality may change very slowly over time [4]. Commonly the personality is described in the Five Factor Model. Therein the big-five traits are traditionally defined as Extraversion, Agreeableness, Conscientiousness, Neuroticism or Emotional Stability, and Openness or Creativity [72]. Personality plays an increasingly important role in computing areas - especially technologies that need to understand and predict human behavior could benefit from Personality Computing approaches [123]. Collecting personality data from questionnaires requires effort from the user, whereas automatic data collection not.

As stated in chapter 1, section 1.3, the personality traits of individuals correlate with their emotional responses to a certain affective content. So that, the personality traits of a user has impact on how a person perceives an affective content. Hereby, the research question is whether we could assess the personality of users, given their implicit/explicit psycho-physiological responses to certain affective contents. The research question is covered in section 3.1 where a publicly available¹ dataset is developed to answer² the question using portable and

¹mhug.disi.unitn.it/wp-content/ASCERTAIN/ascertain.html

²The research is an extension to our previous work [125] and it is published [116] in IEEE Transactions on Affective Computing, Nov. 2016.

off-the-shelves market available sensors.

When dealing with such sensors, the probability of the presence of noise in input signals increases (See 3.1.2). The issues with noisy signals motivated us to conduct a research covered in section 3.2 to deal with the noise. We propose and validate a system and a method to take into the consideration the quality of input signals within a multi-modal affect recognition problem that has a high noise tolerance.

3.1 The ASCERTAIN Dataset and Research

Despite rapid advances in Human-computer Interaction (HCI) and relentless endeavors to improve user experience with computer systems, the need for agents to *recognize* and *adapt* to the **affective state** of users has been widely acknowledged. While being a critical component of human behavior, affect is nevertheless a highly subjective phenomenon influenced by a number of contextual and psychological factors including **personality**.

The personality–affect relationship has been actively studied ever since a correlation between the two was proposed in Eysenck’s personality model [28]. Eysenck posited that Extraversion, the personality dimension that describes a person as either talkative or reserved, is accompanied by low cortical arousal—*i.e.*, extraverts require more external stimulations than introverts. His model also proposed that neurotics, characterized by negative feelings such as depression and anxiety, are more sensitive to external stimulation and become easily upset or nervous due to minor stressors.

Many affective studies have attempted to validate and extend Eysenck’s findings. Some have employed explicit user feedback in the form of affective self-ratings [81, 43], while others have measured implicit user responses such as Electroencephalogram (EEG) activity [113] and heart rate [23] for their analyses. However, few works have investigated affective correlates of traits other

than Extraversion and Neuroticism. Conversely, social psychology studies have examined personality mainly via non-verbal social behavioral cues (see [123] for a review), but few works have modeled personality traits based on emotional behavior. Conducting studies to examine the personality–affect relationship is precluded by problems such as subject preparation time, invasiveness of sensing equipment and the paucity of reliable annotators for annotating emotional attributes.

This work builds on [125] and examines the influence of personality differences on users’ affective behavior via the ASCERTAIN database³. We utilize ASCERTAIN to (i) understand the relation between *emotional attributes* and *personality traits*, and (ii) characterize both via users’ physiological responses. ASCERTAIN contains personality scores and emotional self-ratings of 58 users in addition to their affective physiological responses. More specifically, ASCERTAIN is used to model users’ emotional states and *big-five* personality traits via heart rate (Electrocardiogram or ECG), galvanic skin response (GSR), EEG and facial activity patterns observed while viewing 36 affective movie clips.

We specifically designed a study with movie scenes as they effectively evoke emotions [32, 53], as typified by genres such as *thriller*, *comedy* or *horror*. Also, different from existing affective databases such as DEAP[57], MAHNOB [109] and DECAF [53], ASCERTAIN comprises data recorded exclusively using commercial sensors to ensure ecological validity and scalability of the employed framework for large-scale profiling applications.

Using the ASCERTAIN data, we first examine correlations among users’ valence (V) and arousal (A) self-ratings and their personality dimensions. We then attempt to isolate physiological correlates of emotion and personality. Our analyses suggest that the relationships among emotional attributes and personality traits are better captured by non-linear rather than linear statistics. Finally, we present single-trial (binary) recognition of A,V and the big-five traits consider-

³<http://mhug.disi.unitn.it/index.php/datasets/ascertain/>

ing physiological responses observed over (a) *all*, and (b) *emotionally homogeneous* (e.g., high A, high V) clips. Superior personality recognition is achieved for (b), implying that personality differences are better revealed by comparing responses to emotionally similar stimuli. The salient aspects of ASCERTAIN are:

1. To our knowledge, ASCERTAIN is the first physiological database that facilitates both emotion and personality recognition. In social psychology, personality traits are routinely modeled via questionnaires or social behavioral cues. Instead, this is one of the first works to assess personality traits via affective physiological responses (the only other work to this end is [47]).
2. Different from the DEAP [57], MAHNOB [109] and DECAF [53] databases, we use *wearable, off-the-shelf* sensors for physiological recordings. This enhances the ecological validity of the ASCERTAIN framework, and above-chance recognition of emotion and personality affirms its utility and promise for commercial applications.
3. We present interesting insights concerning correlations among affective and personality attributes. Our analyses suggest that the emotion–personality relationship is better captured via non-linear statistics. Also, personality differences are better revealed by comparing user responses to emotionally similar videos (or more generally, under similar affect inducement).

From here on, Section 3.1.1 reviews related literature to motivate the need for ASCERTAIN, while Section 3.1.2 details the materials and methods employed for data compilation. Section 3.1.3 presents descriptive statistics, while correlations among users’ affective ratings and personality dimensions are analyzed in Section 3.1.4. Section 3.1.5 details physiological correlates of emotion and personality, while Section 3.1.6 presents recognition experiments. Section 3.1.7 discusses the correlation and recognition results.

Table 3.1: Comparison of user-centered affective databases. ‘var’ denotes variable.

Name	No. subjects	No. stimuli	Recorded signals	Annotations		Comments
				Affect	Personality	
HUMAINE [26]	var	var	audio, visual, physiological	yes	no	includes 6 sub-collections (some non-public)
DEAP [57]	32	40	physiological	yes	no	focus on music videos
DECAF [53]	30	76	face, physiological	yes	no	compares music and movie clips
MAHNOB-HCI [109]	27	20	face, audio, eye gaze, physiological	yes	no	includes video and image stimuli
ASCERTAIN	58	36	face, physiological	yes	yes	connects emotion and personality

3.1.1 Related Work

This section reviews related work focusing on (a) multimodal affect recognition, (b) personality assessment and (c) the personality–affect relationship.

Multimodal affect recognition

As emotions are conveyed by content creators using multiple means (audio, video), and expressed by humans in a number of ways (facial expressions, speech and physiological responses), many affect recognition (AR) methods employ a multimodal framework. Common content-based modalities employed for AR include audio [10, 9, 62], visual [70, 135, 86] and audio-visual [17, 29, 97]. Recent AR methodologies have focused on monitoring user behavior via the use of physiological sensors (see [127] for a review). Emotions induced by music clips are recognized via heart rate, muscle movements, skin conductivity and respiration changes in [56]. Lisetti *et al.* [68] use GSR, heart rate and temperature signals to recognize emotional states. As part of the HUMAINE project [26], three naturalistic and six induced affective databases containing multimodal data (including physiological signals) are compiled from 8–125 participants. Tavakoli *et al.* [93] examine the utility of various eye fixation and saccade-based features for valence recognition, while Subramanian *et al.* [115] correlate user responses with eye movement patterns to discuss the impact of emotions on visual attention and memory.

Koelstra *et al.* [57] analyze blood volume pressure, respiration rate, skin temperature and Electrooculogram (EOG) patterns for recognizing emotional states induced by 40 music videos. MAHNOB-HCI [109] is a multimodal database

containing synchronized face video, speech, eye-gaze and physiological recordings from 27 users. Abadi *et al.* [53] study Magnetoencephalogram (MEG), Electromyogram (EMG), EOG and ECG responses from users for music and movie clips, and conclude that better emotion elicitation and AR are achieved with movie clips.

Personality recognition

The *big-five* or five-factor model [21] describes human personality in terms of five dimensions— Extraversion (*sociable vs reserved*), Neuroticism or the degree of emotional stability (*nervous vs confident*), Agreeableness (*compassionate vs dispassionate*), Conscientiousness (*dutiful vs easy-going*) and Openness (*curious/creative vs cautious/conservative*).

A comprehensive survey of personality computing approaches is presented in [123]. The traditional means to model personality traits are questionnaires or self-reports. Argamon *et al.* [5] use lexical cues from informal texts for recognizing Extraversion (*Ex*) and Neuroticism (*Neu*). Olguin *et al.* [82] and Alameda-Pineda *et al.* [2] show that non-verbal behavioral measures acquired using a sociometric badge such as the amount of speech and physical activity, number of face-to-face interactions and physical proximity to other objects is highly correlated with personality. Much work has since employed non-verbal behavioral cues in social settings for personality recognition including [63], where *Ex* is recognized using speech and social attention cues in round-table meetings, while [117, 138] predict *Ex* and *Neu* from proxemic and attention cues in party settings.

Among works that have attempted recognition of all five personality factors, Mairesse *et al.* [71] use acoustic and lexical features, while Staiano *et al.* [111] analyze structural features of individuals' social networks. Srivastava *et al.* [110] automatically complete personality questionnaires for 50 movie characters utilizing lexical, audio and visual behavioral cues. Brouwer *et al.* [14]

estimate personality traits via physiological measures, which are revealed subconsciously and more genuinely (less prone to manipulation) than questionnaire answers. In a gaming-based study, they observe a negative correlation between (i) heart rate and *Ex*, and (ii) skin-conductance and *Neu*.

Personality-Affect relationship

The relationship between personality and affect has been extensively examined in social psychology [130], but not in a computational setting. Eysenck's seminal personality theory [28] posits that extraverts require more external stimulation than introverts, and that neurotics are aroused more easily. Many studies have since studied the personality–affect relationship by examining explicit or implicit user responses. Personality effects on brain activation related to valence (V) and arousal (A) is investigated in [43], which concludes that *Neu* correlates negatively with positive V, and positively with A. In an EEG-based study [113], a negative correlation is observed between *Ex* and A, while a positive correlation is noted between *Neu* and A especially for negative valence stimuli.

The impact of personality traits on affective user ratings is studied using path analysis in [121]. Feedback scores from 133 students are analyzed in [81] to conclude that neurotics experience positive emotions similar to emotionally stable counterparts in pleasant situations, even though they may experience negative emotions more strongly. Event-related potentials and heart rate changes are studied in [23] to confirm a positive correlation between *Neu* and A for negative stimuli, while a signal-detection task is used in [35] to suggest that extraverts are generally less aroused than introverts. Brumbaugh *et al.* [15] examine correlations among the big-five traits, and find *Ex* and *Neu* to be associated with increased A while viewing negative videos. Abadi *et al.* [47] attempt recognition of the big-five traits from affective physiological responses, and our work is most similar to theirs in this respect. Nevertheless, we consider more users and a larger stimulus set in this work (58 users and 36 clips vs 36 users and

16 clips in [47]), and show superior personality trait recognition on comparing physiological responses to emotionally homogeneous clips.

Spotting the research gap

Examination of related literature reveals that AR methodologies are increasingly becoming *user-centric* instead of *content-centric*, suggesting that emotions better manifest via human behavioral cues rather than multimedia content-based (typically audio, visual and speech-based) cues. Nevertheless, the influence of psychological factors such as personality on emotional behavior has hardly been examined, in spite of prior work suggesting that personality affects one's a) feelings [130, 67], b) emotional perception [43, 113] and c) multimedia preferences [59, 100].

Motivated by the above findings and the lack of publicly available data sets positioned at the intersection of personality and affect, we introduce ASCERTAIN, a multimodal corpus containing physiological recordings of users viewing emotional videos. ASCERTAIN allows for inferring both personality traits and emotional states from physiological signals. We record GSR, EEG, ECG signals using wearable sensors, and facial landmark trajectories (EMO) using a web-camera. In the light of recent technological developments, these signals can be acquired and analyzed instantaneously. Also, Wang and Ji [127] advocate the need for less-intrusive sensors to elicit natural emotional behavior from users. Use of wearable sensors is critical to ensure the ecological validity, repeatability and scalability of affective computing studies, which are typically conducted in controlled lab conditions and with small user groups.

Table 3.1 presents an overview of publicly available user-centric AR datasets. Apart from being one of the largest datasets in terms of the number of participants and stimuli examined for analysis, ASCERTAIN is also the first database to facilitate study of the personality–affect relationship.

3.1.2 ASCERTAIN Overview

Fig. 3.1 presents an overview of the ASCERTAIN framework and a summary of the compiled data is provided in Table 3.2. To study the personality–affect relationship, we recorded users’ physiological responses as they viewed the affective movie clips used in [53]. Additionally, their explicit feedback, in the form of *arousal*, *valence*, *liking*, *engagement* and *familiarity* ratings, were obtained on viewing each clip. Finally, personality measures for the big-five dimensions were also compiled using a big-five marker scale (BFMS) questionnaire [89]. We now describe (1) the procedure adopted to compile users’ emotional ratings, personality measures and physiological responses, and (2) the physiological features extracted to measure users’ emotional responses.

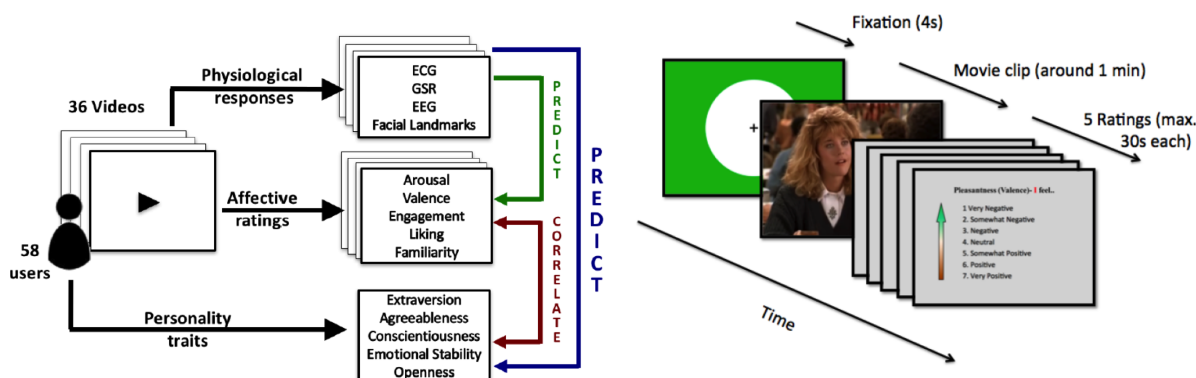


Figure 3.1: (a) ASCERTAIN study overview. (b) Timeline for each trial.

Table 3.2: Summary of the ASCERTAIN database.

Number of Participants	58
Number of Videos	36
Video Length	51–128 seconds ($\mu \pm \sigma = 80 \pm 20$)
Self-reported ratings	Arousal, Valence, Engagement Liking, Familiarity
Personality Scales	Extraversion, Agreeableness Conscientiousness, Neuroticism, Openness
Physiological signals	ECG, GSR, Frontal EEG, Facial features

Materials and Methods

Subjects: 58 university students (21 female, mean age = 30) participated in the study. All subjects were fluent in English and were habitual Hollywood movie watchers.

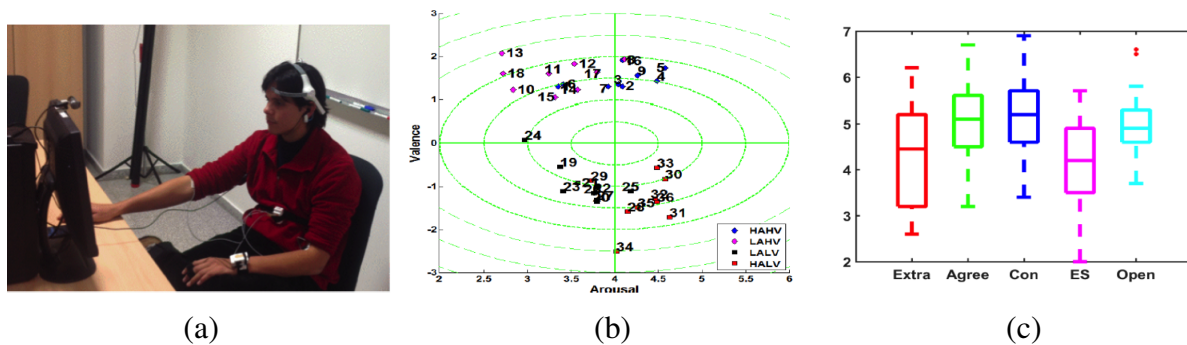


Figure 3.2: (a) Participant with sensors (EEG, ECG and GSR visible) during the experiment, (b) Mean Arousal-Valence (AV) ratings for the 36 movie clips used in our experiment and (c) Box-plots showing distribution of the big-five personality trait scores for 58 users.

Materials: One PC with two monitors was used for the experiment. One monitor was used for video clip presentation at 1024×768 pixel resolution with 60 Hz screen refresh rate, and was placed roughly one meter before the user. The other monitor allowed the experimenter to verify the recorded sensor data. Following informed consent, physiological sensors were positioned on the user's body as shown in Fig. 3.2(a). The GSR sensor was tied to the left wrist, and two electrodes were fixed to the index and middle finger phalanges. Two measuring electrodes for ECG were placed at each arm crook, with the reference electrode placed at the left foot. A single dry-electrode EEG device was placed on the head like a normal headset, with the EEG sensor touching the forehead and the reference electrode clipped to the left ear. EEG data samples were logged using the `Lucid Scribe` software, and all sensor data were recorded via bluetooth. A webcam was used to record facial activity. Synchronized data recording and

pre-processing were performed using MATLAB Psychtoolbox⁴.

Protocol: Each user performed the experiment in a session lasting about 90 minutes. Viewing of each movie clip is denoted as a *trial*. After two practice trials involving clips that were not part of the actual study, users watched movie clips randomly shown in two blocks of 18 trials, with a short break in-between to avoid fatigue. In each trial (Fig. 3.1(b)), a fixation cross was displayed for four seconds followed by clip presentation. After viewing each clip, users self-reported their emotional state in the form of affective ratings within a time limit of 30 seconds. They also completed a personality questionnaire after the experiment.

Stimuli: We adopted the 36 movie clips used in [53] for our study. These clips are between 51–127 s long ($\mu = 80$, $\sigma = 20$), and are shown to be uniformly distributed (9 clips per quadrant) over the arousal-valence (AV) plane.

Affective ratings: For each movie clip, we compiled valence (V) and arousal (A) ratings reflecting the user’s affective impression. A 7-point scale was used with a -3 (*very negative*) to 3 (*very positive*) scale for V, and a 0 (*very boring*) to 6 (*very exciting*) scale for A. Likewise, ratings concerning engagement (*Did not pay attention – Totally attentive*), liking (*I hated it – I loved it*) and familiarity (*Never seen it before – Remember it very well*) were also acquired. Mean user V,A ratings for the 36 clips are plotted in Fig. 3.2(b), and are color-coded based on the ground-truth ratings from [53]. Ratings form a ‘C’-shape in the AV plane, consistent with prior affective studies [57, 53].

Personality scores: Participants also completed the big-five marker scale (BFMS) questionnaire [89] which has been used in many personality recognition works [138,

⁴<http://psychtoolbox.org/>

Table 3.3: Extracted features for each modality (feature dimension stated in parenthesis). *Statistics* denote mean, standard deviation (std), skewness, kurtosis of the raw feature over time, and % of times the feature value is above/below $\text{mean} \pm \text{std}$.

Modality	Extracted features
ECG (32)	Ten low frequency ([0-2.4] Hz) power spectral densities (PSDs), four very slow response ([0-0.04] Hz) PSDs, IBI, HR and HRV statistics.
GSR (31)	Mean skin resistance and mean of derivative, mean differential for negative values only (mean decrease rate during decay time), proportion of negative derivative samples, number of local minima in the GSR signal, average rising time of the GSR signal, spectral power in the [0-2.4] Hz band, zero crossing rate of skin conductance slow response ([0-0.2] Hz), zero crossing rate of skin conductance very slow response ([0-0.08] Hz), mean SCSR and SCVSR peak magnitude.
Frontal EEG (88)	Average of first derivative, proportion of negative differential samples, mean number of peaks, mean derivative of the inverse channel signal, average number of peaks in the inverse signal, statistics over each of the 8 signal channels provided by the Neurosky software.
EMO (72)	Statistics concerning horizontal and vertical movement of 12 motion units (MUs) specified in [41].

63, 117]. Scale distributions for the big-five traits are shown in Fig. 3.2(c). The most and least variance in personality scores are noted for the Extraversion and Openness traits respectively.

Physiological feature extraction

We extracted physiological features corresponding to each trial over the final 50 seconds of stimulus presentation, owing to two reasons: (1) The clips used in [53] are not emotionally homogeneous, but are more emotional towards the end. (2) Some employed features (see Table 3.3) are nonlinear functions of the input signal length, and fixed time-intervals needed to be considered as the movie clips were of varying lengths. Descriptions of the physiological signals examined in this work are as follows.

Galvanic Skin Response (GSR): GSR measures transpiration rate of the skin. When two electrodes are positioned on the middle and index finger phalanges and a small current is sent through the body, resistance to current flow changes

with the skin transpiration rate. Most of the GSR information is contained in low-frequency components, and the signal is recorded at 100 Hz sampling frequency with a commercial bluetooth sensor. Following [56, 57, 109], we extracted 31 GSR features listed in Table 3.3.

Electroencephalography (EEG): EEG measures small changes in the skull’s electrical field produced by neural activity, and information is encoded in the EEG signal amplitude as well as in certain frequency components. We used a commercial, single dry-electrode EEG sensor⁵, which records eight information channels sampled at 32 Hz. The recorded information includes frontal lobe activity, level of facial activation, eye-blink rate and strength, which are relevant emotional responses.

Electrocardiogram (ECG): Heart rate characteristics have been routinely used for user-centered emotion recognition. We performed R-peak detection on the ECG signal to compute users’ inter-beat intervals (IBI), heart rate (HR), and the heart rate variability (HRV). We also extracted power spectral density (PSD) in low frequency bands as in [56, 109].

Facial landmark trajectories (EMO): A facial feature tracker [41] was used to compute displacements of 12 interest points or motion units (MU) in each video frame. We calculated 6 statistical measures for each landmark to obtain a total of 72 features (Table 3.3).

Data Quality

A unique aspect of ASCERTAIN with respect to prior affective databases is that physiological signals are recorded using commercial and minimally invasive sensors that allow body movement of participants. However, it is well

⁵www.neurosky.com

known that body movements can degrade quality of the recorded data, and such degradation may be difficult to detect using automated methods. Therefore, we plotted the recorded data for each modality and trial, and rated the data quality manually on a scale of 1 (*good data*)–5 (*missing data*). For ECG, we evaluated the raw signal from each arm as well as the R-peak amplitude. The presence/absence of facial tracks and correctness of the tracked facial locations were noted for EMO. For GSR, we examined the extent of data noise, and rated EEG (i) on the raw signal, (ii) by summarizing the quality of δ (< 4 Hz), θ (4–7 Hz), α (8–15 Hz), β (16–31 Hz) and γ (> 31 Hz) frequency bands, and (iii) on the pre-calculated *attention* and *meditation* channels available as part of the EEG data. Plots and tables with explanations on data quality are available with the dataset. Fig. 3.3 presents an overview of the data quality for the four considered modalities, with the proportion of trials for which the quality varies from 1–5 highlighted. About 70% of the recorded data is good (corresponding to levels 1-3) for all modalities except ECG, with GSR data being the cleanest. Maximum missing data is noted for EEG, reflecting the sensitivity of the EEG device to head movements.

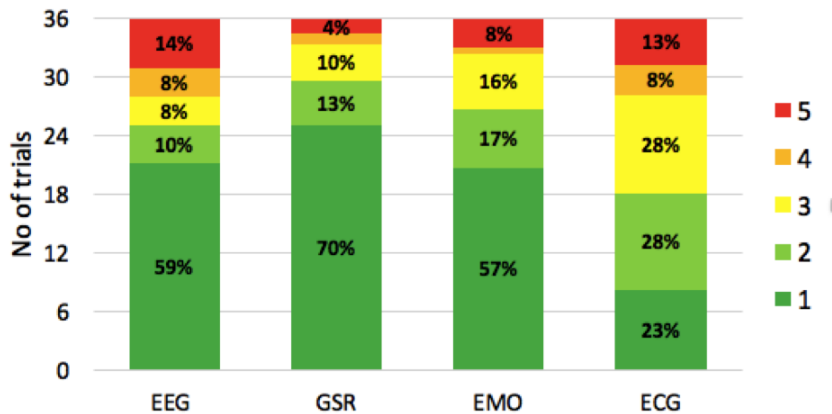


Figure 3.3: Bar plot showing proportion of trials for which data quality ranges from best (1) to worst (5).

3.1.3 Descriptive Statistics

In this section, we present statistics relating to user self-reports and personality scores.

Analysis of Self-ratings

As mentioned previously, we selected 36 movie clips such that their emotional ratings were distributed uniformly over the AV plane as per ground-truth ratings in [53], with 9 clips each corresponding to the HAHV (high arousal-high valence), LAHV (low arousal-high valence), LALV (low arousal-low valence) and HALV (high arousal-low valence) quadrants⁶. The targeted affective state was mostly reached during the ASCERTAIN study as shown in Fig. 3.2(b). A two-sample t -test revealed significantly higher mean A ratings for HA clips as compared to LA clips ($t(34) = 5.1253, p < 0.0001$). Similarly, mean V ratings for HV and LV clips were significantly different ($t(34) = 17.6613, p < 0.00005$). Overall, emotion elicitation was more consistent for valence as in prior works [53, 57].

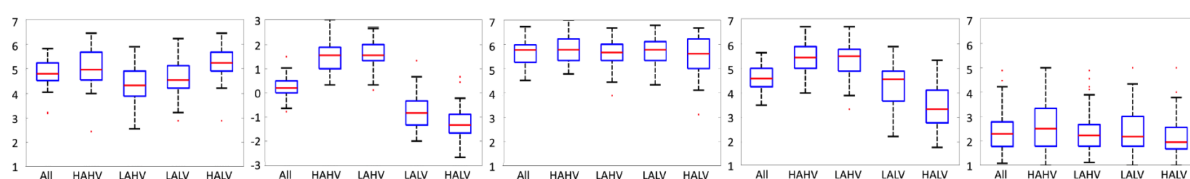


Figure 3.4: Boxplots of the mean Arousal, Valence, Engagement, Liking and Familiarity ratings for the different video sets.

We computed agreement among participants’ A,V ratings using the Krippendorff’s alpha metric— agreement for A and V were respectively found to be 0.12 and 0.58, implying more consensus for clip valence as above. We then computed the agreement between the ASCERTAIN and DECAF [53] populations using the Cohen’s Kappa (κ) measure. To this end, we computed κ be-

⁶For consistency’s sake, quadrant-wise video labels derived based on ratings from [53] are used in this work.

tween ground-truth (GT) labels from [53] and each user’s A,V labels assigned as *high/low* based on the mean rating– the mean agreement over all users for A and V was found to be 0.24 and 0.73 respectively. We also computed the κ measure between GT and the ASCERTAIN population based on the mean A,V rating of all users– here, an agreement of 0.39 was observed for A and 0.61 for V. Overall, these measures suggest that while individual-level differences exist in affective perception of the movie clips, there is moderate to substantial agreement between assessments of the ASCERTAIN and DECAF populations implying that the considered movie clips are effective for emotion elicitation.

Fig. 3.4 presents box-plots describing the distribution of the arousal (A), valence (V), engagement (E), liking (L) and familiarity (F) user ratings for (i) all, and (ii) quadrant-based videos. Clearly, low-arousal videos are perceived as more ‘neutral’ in terms of A and V, which leads to the ‘C’ shape in Fig. 3.2(b). All videos are perceived as sufficiently engaging, while HV clips are evidently more liked than LV clips. Also, the presented movie clips were not very conversant to participants, suggesting that the ASCERTAIN findings are overall unlikely to be influenced by familiarity biases.

Affective Ratings vs Personality Scales

To examine relationships between the different user ratings, we computed Pearson correlations among self-reported attributes as shown in Table 3.4. Since the analysis involves attribute ratings provided by 58 users for each of the 36 clips, we accounted for multiple comparisons by limiting the false discovery rate (FDR) to within 5% using the procedure outlined in [11]. Highlighted numbers denote correlations found to be significant over at least 15 users (25% of the population) adopting the above methodology.

Focusing on significant correlations, A is moderately correlated with E, while V is found to correlate strongly with L mirroring the observations of Koelstra *et al.* [57]. A moderate and significant correlation is noted between E and L

Table 3.4: Mean Pearson correlations between self-ratings across users. *s denote significant correlations ($p < 0.05$) upon limiting FDR to 5%.

	A	V	E	L	F
Arousal	1	0.02	0.42*	0.19	0.15
Valence		1	0.21	0.68*	0.17
Engagement			1	0.42*	0.24
Liking				1	0.34*
Familiarity					1

Table 3.5: Pearson correlations between personality dimensions (* $\Rightarrow p < 0.05$)

	E	A	Co	ES	O
Extraversion	1	0.36*	0.19	-0.12	0.30*
Agreeableness		1	0.21	0.34*	0.30*
Conscientiousness			1	0.26	0.04
Emotional Stability				1	-0.10
Openness					1

implying that engaging videos are likely to appeal to viewers’ senses, and similarly, between F and L confirming the mere exposure effect observed in [12] attributing liking to familiarity. Nevertheless, different from [57] with music videos where a moderate correlation is noted between A and V ratings, we notice that the A and V dimensions are uncorrelated for the ASCERTAIN study, which again reinforces the utility of movie clips as good control stimuli. To validate our experimental design, we tested for effects of video length on A,V ratings but did not find any.

Table 3.5 presents Pearson correlations between personality dimensions. Again focusing on significant correlations, moderate and positive correlations are noted between Extraversion (*Ex*) and Agreeableness (*Ag*), as well as between *Ex* and Openness (*O*)– prior studies have noted that *Ex* and *O* are correlated via the sensation seeking construct [3]. *Ag* is also found to moderately and

positively correlate with Emotional Stability (*ES*) and *O*. Conversely, weakly negative-but-insignificant correlations are observed between (i) *Ex* and *ES*, and (ii) *ES* and *O*.

Table 3.6: Partial correlations between personality scales and self-ratings (* $\Rightarrow p < 0.05$).

		Ex	Ag	Co	ES	O
All	Arousal	0.03	-0.10	0.05	0.07	0.06
	Valence	0.19	-0.02	0.02	-0.18	0.07
	Engage	-0.30*	0.01	0.09	0.00	-0.10
	Liking	-0.13	0.07	-0.22	0.21	-0.02
HAHV	Arousal	-0.01	0.02	-0.01	0.22	-0.11
	Valence	-0.12	-0.38*	-0.11	-0.12	-0.16
	Engage	-0.30*	0.16	0.17	0.10	-0.09
	Liking	0.20	0.22	-0.00	0.10	0.22
LAHV	Arousal	-0.06	0.07	0.06	0.11	0.04
	Valence	-0.03	0.05	-0.08	-0.10	0.23
	Engage	-0.22	0.03	-0.06	-0.10	-0.24
	Liking	0.20	-0.01	0.13	0.17	0.12
LALV	Arousal	0.03	-0.09	0.02	-0.07	0.09
	Valence	0.20	0.06	0.01	-0.22	0.15
	Engage	-0.22	-0.01	0.02	-0.05	-0.04
	Liking	-0.14	0.03	-0.19	0.12	-0.10
HALV	Arousal	0.20	-0.25	-0.00	-0.16	0.09
	Valence	0.22	0.01	0.09	-0.06	-0.05
	Engage	-0.30*	0.01	0.10	0.12	-0.10
	Liking	-0.26*	0.03	-0.35*	0.12	-0.07

Partial correlations between emotional and personality attributes are tabulated in Table 3.6. Considering all movie clips, a significant and moderately negative correlation is noted between *Ex* and E, implying that introverts were more immersed with emotional clips during the movie-watching task. A few more significant correlates are observed when mean ratings for quadrant-wise (or emotionally similar) videos are considered. Delineating, *Ag* is negatively correlated with V for HAHV videos, while the negative correlation between *Ex*

Table 3.7: R^2 and best three predictors for the five personality dimensions. Full model coefficients are shown in parentheses. * $\Rightarrow p < 0.05$.

	Ex	Ag	Co	ES	O
All	0.14* (0.14) V,E,L	0.02 (0.02) A,V,L	0.07 (0.07) V,E,L	0.06 (0.06) A,V,L	0.02 (0.02) A,V,E
HAHV	0.16* (0.16) V,E,L	0.17* (0.17) V,E,L	0.05 (0.05) A,V,E	0.12* (0.13) A,V,L	0.05 (0.06) A,V,L
LAHV	0.07 (0.08) A,E,L	0.02 (0.02) A,V,E	0.02 (0.02) A,E,L	0.04 (0.05) A,E,L	0.13* (0.13) V,E,L
LALV	0.12 (0.12) V,E,L	0.02 (0.02) A,V,L	0.05 (0.05) A,E,L	0.05 (0.05) A,V,L	0.03 (0.03) A,V,L
HALV	0.16* (0.20) V,E,L	0.09 (0.09) A,V,L	0.15 (0.16) V,E,L	0.04 (0.05) A,E,L	0.03 (0.04) A,E,L

and E manifests for high-arousal (HAHV and HALV) stimuli. Also notable is the moderately negative correlation between *Ex* and L, and also between Conscientiousness and L for HALV movie clips. Surprisingly, a negative correlation is noted between *O* and E for LAHV clips. Consistent with prior studies [20], V is positively correlated with *Ex* in general, with a significant and moderately positive correlation noted for LALV clips. Finally, a moderately negative correlation is observed between V and *ES* for LALV clips consistent with the observations made in [81].

We also performed linear regression analyses with user self ratings as predictors and personality attributes as the target variables for the different video sets, and the coefficients of determination/squared correlations (R^2) for the different video sets are presented in Table 3.7. R^2 values with the three best predictors along with the predictor names are listed outside parentheses, while squared correlations with the full model are listed within braces. Considering all movie clips, the best linear model is obtained for *Ex* with V, E and L ratings as predictors. Among the four AV quadrants, significant squared correlations are observed for the *Ex* and *Ag* traits with V,E,L predictors, and for the *ES* trait with

arousal, valence and liking ratings as predictors for HAHV clips. A significant model is also obtained for Openness with V,E,L predictors considering mildly positive HALV clips. Overall, it is easy to observe from the table that (i) there is little difference in the predictive power of the best-three-predictor and full models, and (ii) the linear models have rather limited predictive power, with the best model explaining only 17% of the personality scale variance. Cumulatively, Tables 3.6 and 3.7 cumulatively suggest that the relationship between emotional and personality variables is not well modeled using linear statistics, and it is perhaps worthwhile to explore the use of non-linear measures to this end. From here on, given the high degree of correlation between A and E and between the V and L, we will only focus on A and V dimensions in the rest of the chapter.

Mutual Information Analysis

Mutual information (MI) is a popular metric to capture non-linear relationships between two random variables, and measures how much information is known about one variable given the other. Formally, the MI between two random vectors $X = \{x\}$ and $Y = \{y\}$ is defined as:

$MI(X, Y) = \sum_{x,y} P_{XY}(x, y) \log \frac{P_{XY}(x, y)}{P_X(x) \cdot P_Y(y)}$ where $p_{XY}(x, y)$ is the joint probability distribution, while $P_X(x)$ and $P_Y(y)$ are the respective marginal probabilities. We attempted to describe the relationship between emotional ratings and personality scales via the normalized mutual information (NMI) index [114] defined as: $NMI(X, Y) = \frac{MI(X, Y)}{\sqrt{(H(X)H(Y))}}$, where $H(X)$ and $H(Y)$ denote entropies of X and Y .

NMI with personality scales for arousal and valence ratings are shown in Fig. 3.5. In contrast to linear correlations, both A and V share a high degree of mutual information with all five personality traits. Considering all movie clips, emotional ratings share slightly higher MI with A than with V. Also, a strictly higher MI measure is noted when emotionally similar clips are considered in-

stead of all clips. Among personality traits, *Ex* and Conscientiousness (*Con*) share the most MI with V,A attributes– in contrast, little correlation is observed between *Con* and A,V in Table 3.6). Conversely, lowest MI is noted for Openness (*O*). One notable difference exists between A and V though– higher MI with arousal is noted for high HV clips, while for all personality traits barring *Ag*, greater MI with valence is observed for LV clips than for HV clips.

Arousal and Valence both share high mutual information with all five personality traits. In general MI for Arousal is higher than for Valence. The highest MI is shared with Conscientiousness and Extroversion and the lowest with Creativity for both Arousal and Valence suggesting that those personality dimensions have the highest impact on emotion perception.

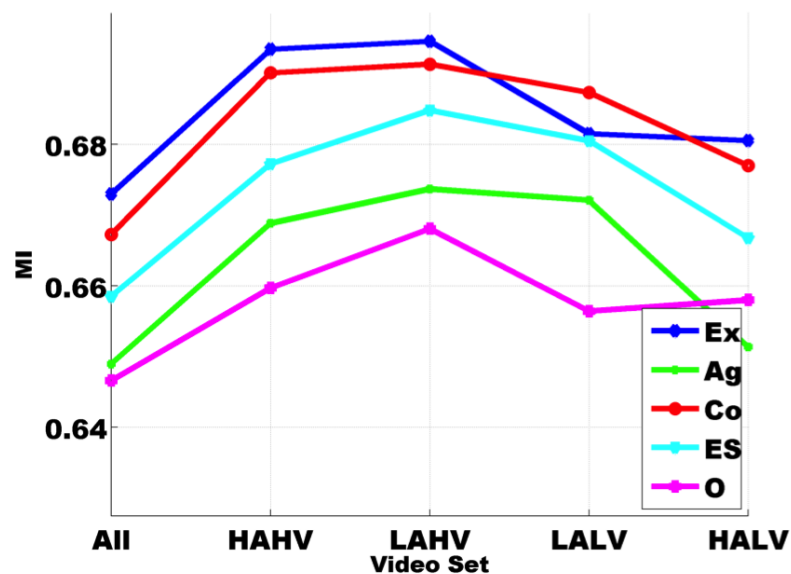


Figure 3.5: NMI between big-five trait scales and A (left), V (right) ratings.

3.1.4 Personality measures vs user ratings

We now examine the relationship between user V,A ratings and personality scales in the context of hypotheses (H1–H3) put forth in the literature. To this

end, we determined *high/low* trait groups (e.g., emotional stable vs neurotic) for each personality dimension by dichotomizing personality measures based on the median score— this generated balanced *high* and *low* sets for the *Ex* and *ES* traits, and an unbalanced split for the remaining traits, with the most imbalance (33 vs 25) noted for *Ag*. We then proceeded to analyze the affective ratings for each group.

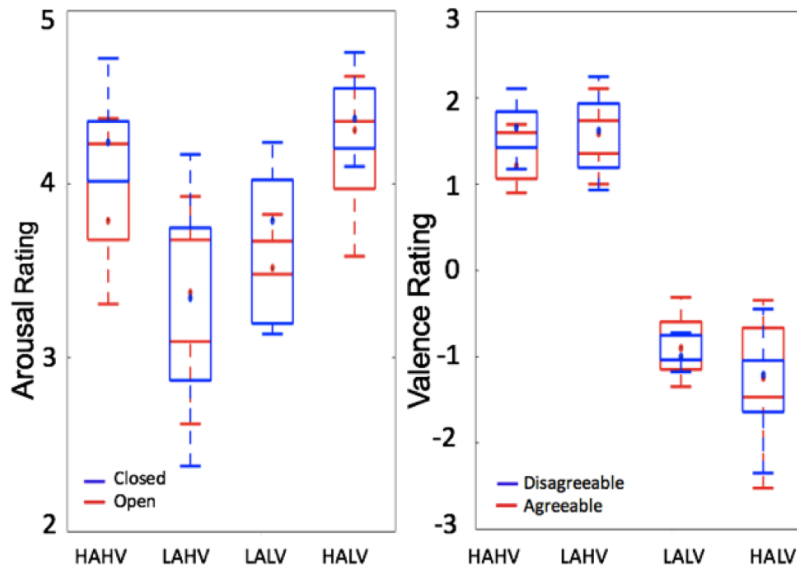


Figure 3.6: Quadrant-wise comparisons of (left) A ratings by *open* and *closed* groups, and (right) V ratings by *agreeable* and *disagreeable* groups .

H1: Extraversion vs Arousal and Valence

The correlation between Extraversion and arousal has been investigated in many studies— EEG measurements [113], signal detection analysis [35] and fMRI [43] have shown lower arousal in extraverts as compared to introverts, consistent with Eysenck’s personality theory. Also, *Ex* has been found to correlate with positive valence in a number of works [20]. Analyses presented in Table 3.6 reveal little correlation between *Ex* and A for all video categories. While two-tailed *t*-tests confirmed that extraverts and introverts rated high A and low A

videos differently ($p < 0.00001$ in both cases), no differences could be identified between their A ratings excepting that extraverts provided marginally lower ratings for HA clips ($t(56) = -1.4423, p = 0.0774$, left-tailed). Focusing on V ratings, positive correlation between *Ex* and V breaks down for HV clips in Table 3.6. Two-sample *t*-tests also failed to reveal any differences. Therefore, statistical analyses weakly support the negative correlation between *Ex* and A, but do not corroborate the positive correlation between *Ex* and V.

H2: Neuroticism vs Arousal and Valence

The relationship between *Neu* and A has also been extensively studied— a positive correlation between *Neu* and A is revealed through fMRI responses in [43], and EEG analysis [113] corroborates this observation for negative V stimuli. [81] further remarks that neurotics experience negative emotions stronger than emotionally stable persons. In contrast, differing observations have been made regarding the relationship between *Neu* and V. Negative correlation between *Neu* and positive V is noted in [43], whereas a positive relationship between the two for low A stimuli is observed in [121]. [81] remarks that the *Neu*-V relation is moderated by situation— while neurotics may feel less positive in unpleasant situations, they experience positive emotions as strongly as *ES* counterparts in pleasant conditions.

Negative correlation between Emotional Stability (*ES*) and A (or positive correlation between *Neu* and A) is noted only for HALV clips in Table 3.6. However, post-hoc *t*-tests failed to reveal differences between A ratings for the two categories. Also, Table 3.6 generally suggests a negative correlation between *ES* and V— *t*-test comparisons further revealed marginally lower V ratings provided by *ES* subjects for LALV clips ($t(16) = -1.3712, p = 0.0946$, left-tailed). Overall, our data does not support the positive relationship between *Neu* and A, and suggests a weakly positive correlation between *Neu* and V.

H3: Openness vs Valence and Arousal

Among the few works to study Openness, [121] notes a positive correlation between Openness (O) and V under low arousal conditions, which is attributed to the intelligence and sensitivity of creative individuals⁷, enabling them to better appreciate subtly emotional stimuli. Table 3.6 echoes a positive (even if insignificant) correlation between O and V for LA clips but post-hoc t -tests to compare V ratings of *open* and *closed* groups failed to reveal any differences. However, we noted that *closed* individuals felt somewhat more aroused by HA clips than *open* individuals ($t(56) = -1.5011, p = 0.0695$, left-tailed) as shown in Fig. 3.6(a). Fine-grained analysis via left-tailed t -tests to compare quadrant-wise ratings again revealed the slightly higher arousal experienced by *closed* subjects for HAHV clips ($t(16) = -1.3753, p = 0.0940$, left-tailed). In summary, our data weakly confirms a positive relationship between O and V as noted in [121], but suggests a negative correlation between O and A .

Agreeableness and Conscientiousness

Table 3.6 shows a negative but insignificant correlation between Ag and A for HALV videos. Comparison of A ratings by *agreeable* and *disagreeable* groups revealed marginally lower A for *agreeable* subjects for HA clips ($t(56) = -1.2964, p = 0.10$, left-tailed), and subsequent quadrant-wise comparisons attributed this finding to significantly lower A ratings provided by the *agreeable* group for strongly negative HALV clips ($t(16) = -2.6587, p < 0.01$, left-tailed). This trend could possibly be attributed to the association of *disagreeable* persons with negative feelings such as deceit and suspicion. Table 3.6 also shows a negative correlation between Ag and V for highly positive HAHV clips. T -test comparisons again revealed that *agreeable* subjects provided somewhat lower V ratings for HV clips ($t(56) = -1.4285, p = 0.0793$, left-tailed), and

⁷Creativity strongly correlates with Openness [74].

this was particularly true of HAHV clips for which significantly lower ratings were provided by the *agreeable* group ($t(16) = -2.0878, p < 0.05$, left-tailed). Conscientiousness scale differences did not influence VA ratings in any way.

3.1.5 Physiological correlates of emotion and personality

Linear and non-linear analyses presented in the previous sections suggest that correlations between emotional and personality attributes are better revealed while examining user responses to emotionally similar clips. If explicit ratings provided by users are a conscious reflection of their emotional perception, then the analyses employing physiological signals should also reveal similar patterns. We attempt to identify linear and non-linear physiological correlates of emotion and personality considering responses to *all* and *quadrant-specific* clips in this section.

Linear correlates of Emotion and Personality

We attempted to discover physiological correlates of emotional and the big-five personality attributes via partial Pearson correlations. Given the large number of extracted physiological features (Table 3.3) as compared to the population size for this study, we first performed a principal component analysis (PCA) on each feature modality to avoid overfitting, and retained those components that explained 99% of the variance. This gave us 8–9 predictors for each of the considered modalities. Table 3.8 presents correlations between these principal components, users' affective ratings and personality scales (R° denotes number of significant correlates). For affective dimensions, we determined significant correlates considering mean user VA ratings provided for the 36 clips. We also trained regression models with the significantly correlating components as predictors of the dependent emotion/personality variable, and the squared correlations (R^2) of these models are also tabulated.

3.1. THE ASCERTAIN DATASET AND RESEARCH

Table 3.8: Physiological correlates of emotion and personality attributes. R° denotes the number of significant feature correlates, while R^2 is the coefficient of determination for the regression model with the significant correlates as predictors. Bold values denote linear regression models with a significant R^2 statistic.

Video Set	Feature	Arousal		Valence		Extra.		Agreeable		Conscient		Em. Stab.		Open	
		R°	R^2	R°	R^2	R°	R^2	R°	R^2	R°	R^2	R°	R^2	R°	R^2
All	ECG	1	0.25			1	0.25			1	0.32	2	0.30	1	0.26
	GSR											1	0.24		
	EMO														
	EEG					1	0.08							1	0.19
HAHV	ECG	1	0.30					1	0.24			2	0.32	2	0.31
	GSR														
	EMO			1	0.19					1	0.17				
	EEG							1	0.09			1	0.12	1	0.14
LAHV	ECG	1	0.23	1	0.28	1	0.29			3	0.41	2	0.29	2	0.36
	GSR													1	0.14
	EMO													1	0.17
	EEG														
LALV	ECG	1	0.32	1	0.24					2	0.31	1	0.28		
	GSR													2	0.31
	EMO													1	0.16
	EEG					1	0.12								
HALV	ECG	1	0.33			2	0.44	1	0.23	1	0.28	2	0.33	1	0.26
	GSR														
	EMO	1	0.14			1	0.14					1	0.26	1	0.20
	EEG			1	0.10	1	0.09							1	0.15

Examining Table 3.8, the relatively few (maximum of 3) number of significant predictors can be attributed to the sparse number of principal components employed for analysis. Considering correlations with A and V, more correlates are observed for A than for V overall. At least one significant correlate is noted for all modalities except GSR. ECG is found to correlate most with A, with one correlate observed for all video types. ECG also has the most number of correlates with V (one significant correlate for LAHV and LALV clips). One EMO correlate is noted for both A and V respectively in the HAHV and HALV quadrants. A solitary EEG correlate is noted for V considering HALV clips.

A larger number of physiological correlates are observed for personality traits as compared to emotional attributes. Across all five video types, the least

number of correlates are noted for Agreeableness, while most correlates are noted for Openness. The ECG modality again corresponds the maximum number of correlates, while no correlates are observed for GSR. EEG and EMO correlates are mainly noted for the Openness trait. In general, a larger number of physiological correlates are noted for emotionally similar videos for all traits. Also, linear models with a significant R^2 statistic are mainly obtained with emotion-wise similar clips, suggesting that physiology-based linear models can better predict personality traits while examining user responses under similar affective conditions. Most number of significant models are obtained for Openness, while not even one significant model is obtained for Agreeableness. Finally, focusing on the significant quadrant-specific models, the best models are noted for Extraversion (0.44 with ECG features and HALV videos) and Conscientiousness (0.41 with ECG for LAHV clips). This implies that linear physiological models acquire sufficient power to moderately explain personality variations under such conditions.

Non-linear correlates

To examine non-linear physiological correlates of emotion and personality, we performed a mutual information analysis as previously between extracted features from the four modalities and the said attributes. Given the varying number of features for each modality, we segregated the NMI distribution over all features and the emotion/personality rating using 10-bin histograms. Fig. 3.7 presents the first moment or the mean of the NMI histogram distribution computed over the different video sets for each emotional/personality attribute.

It is easy to note from Fig. 3.7 that personality attributes share more MI with the user physiological responses than A and V, similar to the linear analyses. GSR features share maximum MI with A (highest value of 0.73 for LAHV clips), while EMO features share the most MI with V (peak of 0.75 for HALV clips). In contrast, peak MI of 0.81 is noted between ECG features and *Ex*. For

3.1. THE ASCERTAIN DATASET AND RESEARCH



Figure 3.7: (From top to bottom) Bar plots showing the means of the NMI histograms for the four modalities. Best viewed under zoom.

both emotion and personality attributes, at least one of the NMIs observed with quadrant-based videos is higher than the NMI with all movie clips, implying that a fine-grained examination of the relationship between sub-conscious physiological responses and conscious self-ratings is more informative. Focusing on affective attributes, higher MI between ratings and physiological responses is noted for A for all modalities except EMO. Among the four modalities, ECG and EMO respectively share the most and least MI with A, while EMO and EEG share the highest and least MI with V.

Focusing on the big-five personality traits, the highest NMI histogram means over all modalities are observed for *Ex* and *Con* followed by *ES*, *Agree* and *O*. This trend is strikingly similar to the pattern of MI between affective ratings and personality scores obtained in Fig. 3.1.3. Examining sensing modalities, ECG features share the highest MI with all the personality dimensions, while EEG features correspond to the lowest NMI means.

3.1.6 Recognition results

We performed binary recognition of both emotional and personality attributes to evaluate if the proposed user-centric framework can effectively achieve both. This section details the experiments and results thereof.

Emotion recognition

A salient aspect of our work is the exclusive use of commercial sensors for examining users' physiological behavior. To evaluate if our emotion recognition results are comparable to prior affective works which used laboratory-grade sensors, we followed a procedure identical to the DEAP study [57]. In particular, the most discriminative physiological features were first identified for each modality using Fisher's linear discriminant with a threshold of 0.3. Features corresponding to each user were then fed to the naive Bayes (NB) and linear SVM classifiers as shown in Table 3.9. A leave-one-out cross-validation scheme employed where one video is held out for testing, while the other videos are used for training. The best mis-classification cost parameter C for linear SVM is determined via grid search over $[10^{-3}, 10^3]$ again using leave-one-out cross-validation.

Table 3.9 presents the mean F1-scores over all users obtained using the NB and SVM classifiers with unimodal features and the decision fusion (W_{est}^t) technique described in [58]. In decision fusion, the test sample label is computed as $\sum_{i=1}^4 \alpha_i^* t_i p_i$. Here, i indexes the four modalities used in this work, p_i 's denote posterior SVM probabilities, $\{\alpha_i^*\}$ are the optimal weights maximizing the F1-score on the training set and $t_i = \alpha_i F_i / \sum_{i=1}^4 \alpha_i F_i$, where F_i denotes the F1-score obtained on the training set with the i^{th} modality. Note from Section 3.1.2 that there is an equal distribution of high/low A and V, implying a class ratio (and consequently, a baseline F1-score) of 0.5

Observing Table 3.9, above-chance emotion recognition is evidently achieved

Table 3.9: **Affective state recognition** with linear SVM and Naive Bayes (NB) classifiers. Mean F1-scores over all participants for the four modalities, peripheral Signals (ECG + GSR) and late fusion (W_{est}^t) are shown. Baseline F1-score is 0.5. Maximum unimodal F1-scores are shown in bold.

	ECG		GSR		EMO		EEG		Peripheral		W_{est}^t		Class Ratio
	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	
Valence	0.56	0.60	0.64	0.68	0.68	0.68	0.56	0.60	0.64	0.60	0.69	0.71	0.50
Arousal	0.57	0.59	0.61	0.66	0.59	0.61	0.58	0.61	0.62	0.69	0.64	0.67	0.50

with physiological features extracted using commercial sensors. The obtained F1-scores are superior to DEAP [57], which can possibly be attributed to (1) the use of movie clips, which are found to be better than music videos for emotional inducement as discussed in [53], and (2) to the considerably larger number of subjects employed in this study, which results in a larger training set. GSR features produce the best recognition performance for both A and V, while ECG features produce the worst recognition performance. Considering individual modalities, EEG features are better for recognizing A as compared to V, while the remaining three achieve better recognition of V. These results are consistent with earlier observations made in [58, 53]. Considering multimodal results, peripheral (ECG+GSR) features perform better than unimodal features for A recognition, while the best multimodal F1-score of 0.71 is obtained for V. Finally, comparing the two employed classifiers, NB achieves better recognition than linear SVM for both A and V.

Personality recognition

For binary personality trait recognition, we first dichotomized the big-five personality trait scores based on the median as in Section 3.1.4. This resulted in an even distribution of *high* and *low* trait labels for *Ex* and *ES*, while an inexact split for the other traits. As baselines, we consider majority-based voting and random voting according to class ratio. Based on majority voting, baseline F1-score for the *Ex* and *ES* traits is 0.33, and 0.34 for *Ag*, 0.35 for *Con* and 0.36 for *O*. Via

class-ratio based voting, a baseline score of 0.5 is achieved for all traits. We performed PCA on each feature modality in an identical fashion to linear correlation analyses prior to classification. A leave one-subject-out cross-validation scheme was used to compute the recognition results. Three classifiers were employed for recognition, i) naive Bayes, ii) linear (Lin) SVM and iii) Radial Basis Function (RBF) SVM. The C (linear and RBF SVM) and γ (RBF SVM) parameters were tuned via leave-one-subject-out grid search cross-validation on the training set.

Table 3.10 presents the recognition results, with the best F1-scores achieved using unimodal and multimodal features respectively denoted in bold and bold italics. For each personality trait and video set, a better-than-chance recognition F1-score (> 0.5) is achieved with at least with one of the considered modalities. Considering user physiological responses to all affective videos, the highest and lowest F1-scores are respectively achieved for *ES* (0.73) and *O* (0.53) traits—note from Fig. 3.2(c) that *ES* has the second-highest variance among the five personality dimensions, while *O* corresponds to the lowest variance in personality scores. Excepting for the *ES* trait, higher recognition scores are generally achieved considering user responses to emotionally similar videos, in line with the findings from linear and non-linear correlation analyses.

For all personality traits except *O*, an F1-score higher than 0.6 is achieved for at least some of the video quadrants. Among feature modalities, ECG features produce the best recognition performance across personality traits and video sets, followed by EEG, GSR and EMO. EEG features are found to be optimal for recognizing *Ex*, while ECG features achieve good recognition for the *Ag*, *Con* and *ES* traits. EMO and GSR modalities work best for the Openness trait. Focusing on classifiers, RBF SVM produces the best recognition performance for 13 out of 25 (5 personality traits \times 5 video sets) conditions, while linear SVM performs best only for three conditions. Linear classifiers NB and Lin SVM perform best for the *Ex* trait, while RBF SVM, performs best for the *O*

3.1. THE ASCERTAIN DATASET AND RESEARCH

trait.

Fusion-based recognition is beneficial, and higher recognition scores are generally achieved via multimodal fusion. With user responses acquired for all videos, the highest and least fusion-based F1 scores are achieved for the *ES* (0.77 with RBF SVM) and *O* (0.56 with NB) traits respectively. With quadrant-based videos, a maximum F1-score of 0.78 is noted for *Con* (with linear SVM). NB classifier works best with fusion-based recognition, and produces best performance for the *Ex* trait achieving optimal recognition for all the five video sets.

Table 3.10: Personality recognition considering affective responses to a) all, and b) emotionally homogeneous stimuli. Maximum F1-scores with unimodal classifiers are shown in bold. Maximum fusion scores are denoted in bold italics.

Videos	Method	Extravert			Agreeable			Conscient			Em. Stab			Open		
		NB	SVM (lin)	SVM (rbf)	NB	SVM (lin)	SVM (rbf)	NB	SVM (lin)	SVM (rbf)	NB	SVM (lin)	SVM (rbf)	NB	SVM (lin)	SVM (rbf)
All	ECG	0.56	0.06	0.53	0.55	0.45	0.32	0.60	0.51	0.55	0.53	0.60	0.58	0.48	0.35	0.49
	EEG	0.63	0.52	0.48	0.52	0.12	0.54	0.35	0.35	0.31	0.26	0.46	0.51	0.34	0.36	0.37
	EMO	0.35	0.31	0.45	0.40	0.35	0.42	0.39	0.36	0.34	0.44	0.36	0.47	0.50	0.36	0.26
	GSR	0.45	0.00	0.35	0.39	0.34	0.27	0.57	0.35	0.54	0.49	0.56	0.73	0.28	0.36	0.53
	W_{est}^l	0.65	0.52	0.57	0.58	0.46	0.53	0.65	0.59	0.67	0.59	0.64	0.77	0.56	0.42	0.53
HAHV	ECG	0.59	0.00	0.56	0.48	0.29	0.55	0.50	0.32	0.52	0.55	0.46	0.60	0.45	0.34	0.55
	EEG	0.63	0.43	0.63	0.54	0.10	0.10	0.34	0.35	0.33	0.19	0.32	0.57	0.41	0.35	0.45
	EMO	0.39	0.00	0.54	0.54	0.34	0.62	0.35	0.37	0.33	0.46	0.35	0.34	0.46	0.36	0.35
	GSR	0.22	0.00	0.31	0.47	0.34	0.51	0.53	0.35	0.50	0.42	0.51	0.46	0.28	0.36	0.35
	W_{est}^l	0.65	0.43	0.63	0.53	0.47	0.61	0.60	0.56	0.53	0.62	0.62	0.62	0.59	0.46	0.54
LAHV	ECG	0.55	0.02	0.53	0.58	0.45	0.60	0.70	0.78	0.74	0.55	0.46	0.41	0.56	0.42	0.49
	EEG	0.63	0.53	0.63	0.49	0.12	0.42	0.34	0.35	0.30	0.32	0.55	0.54	0.34	0.36	0.27
	EMO	0.49	0.34	0.51	0.43	0.35	0.10	0.58	0.37	0.36	0.51	0.35	0.39	0.46	0.37	0.57
	GSR	0.45	0.00	0.36	0.51	0.34	0.34	0.59	0.35	0.62	0.54	0.52	0.52	0.28	0.36	0.36
	W_{est}^l	0.67	0.52	0.66	0.62	0.49	0.60	0.74	0.78	0.76	0.62	0.60	0.61	0.67	0.49	0.59
LALV	ECG	0.58	0.10	0.49	0.43	0.29	0.36	0.55	0.55	0.74	0.53	0.58	0.50	0.55	0.36	0.43
	EEG	0.61	0.63	0.57	0.19	0.11	0.50	0.37	0.35	0.59	0.33	0.39	0.42	0.41	0.36	0.29
	EMO	0.56	0.00	0.33	0.54	0.18	0.61	0.30	0.37	0.36	0.49	0.35	0.49	0.33	0.36	0.48
	GSR	0.47	0.00	0.47	0.50	0.34	0.51	0.32	0.35	0.50	0.52	0.59	0.69	0.28	0.36	0.56
	W_{est}^l	0.64	0.61	0.58	0.57	0.34	0.59	0.58	0.56	0.76	0.69	0.69	0.75	0.57	0.46	0.63
HALV	ECG	0.50	0.00	0.51	0.51	0.32	0.62	0.57	0.57	0.62	0.59	0.56	0.66	0.45	0.33	0.50
	EEG	0.65	0.53	0.50	0.42	0.07	0.14	0.42	0.35	0.33	0.27	0.32	0.42	0.34	0.36	0.53
	EMO	0.32	0.34	0.30	0.47	0.35	0.47	0.42	0.36	0.40	0.33	0.36	0.44	0.56	0.36	0.60
	GSR	0.38	0.26	0.35	0.33	0.34	0.25	0.55	0.30	0.48	0.49	0.47	0.45	0.30	0.36	0.60
	W_{est}^l	0.67	0.57	0.59	0.55	0.46	0.55	0.62	0.58	0.59	0.59	0.65	0.67	0.57	0.42	0.66

3.1.7 Discussion

The correlation analyses and recognition results clearly convey two aspects related to personality recognition from physiological data (i) A fine-grained analysis of users' physiological responses to emotionally similar movie clips enables better characterization of personality differences— this reflects in the better linear models obtained for personality traits considering quadrant-specific videos in Table 3.8, and the generally higher NMIs for the same in Fig. 3.7. Furthermore, higher F1-scores are typically obtained when physiological responses to emotionally similar clips are used for personality trait recognition. (ii) The relationship between personality scales and physiological features is better captured via non-linear metrics— considerably high MI is noted between emotional ratings and personality scores as well as between affective physiological responses and personality traits, and this observation is reinforced with RBF-SVM producing the best recognition performance.

Interesting similarities are also evident from the correlation and recognition experiments. The NB and lin SVM classifiers work best for the Ex and ES personality traits, for which a number of linear correlates can be noted in Table 3.8. Also, minimum number of linear physiological correlates are noted for the Ag trait, for which linear classifiers do not work well (best recognition is achieved with RBF SVM for all video types except 'All' in Table 3.9). Likewise, no GSR correlate of emotion is observed in Table 3.8, which reflects in poor emotion recognition of personality traits with linear classifiers using GSR features in Table 3.9. Also, only some EMO correlates of personality traits are revealed in Table 3.8, and this modality achieves inferior personality recognition with linear classifiers.

Comparing Tables 3.7 and Table 3.9, EEG shares the least MI with all personality traits among the considered modalities, and RBF SVM performs poorly with EEG features (only one best F1 score in 25 conditions). Conversely, GSR

shares considerable MI with personality dimensions, and GSR features work best with the RBF SVM classifier in Table 3.9. Some discrepancies also arise between the correlation and recognition results. For example, among the big-five personality traits, Openness shares the least MI with all feature modalities but has a number of linear physiological correlates. However, optimal recognition for this trait is achieved with RBF SVM, even though the achieved unimodal F1-scores are the lowest for this trait.

It is pertinent to point out some limitations of this study in general. Weak linear correlations are noted between emotional and personality scores in Table 3.6, and only few physiological correlates of emotion and personality are observed in Table 3.8, which can partly be attributed to the low variance for three of the personality dimensions and particularly the Openness trait, as seen in Fig. 3.2(c). In this context, median-based dichotomization of the personality scores for binary recognition may not be the most appropriate. However, most user-centered affective studies have also demonstrated recognition in a similar manner and on data compiled from small user populations, due to the inherent difficulty in conducting large-scale affective experiments. Overall, the general consistency in the nature of results observed from the correlation and recognition experiments suggest that data artifacts may have only minimally influenced our analyses, and that reliable affect and personality recognition is achievable via the extracted physiological features. Furthermore, we will make the compiled data publicly available for facilitating related research.

Even though not analyzed in this work, the ASCERTAIN database also includes Familiarity and Liking ratings, which could be useful for other research studies. For example, studying the individual and combined influence of familiarity, liking and personality traits on affective behavior could be relevant and useful information for recommender systems. In particular, personality-aware recommender systems have become more popular and appreciated of late [39], but the fact that personality differences show up even as consumers watch affec-

tive video content can enable video recommender systems to effectively learn user profiles over time.

Familiarity and Liking ratings could be also used to replicate and extend related studies. For example, the study presented in [131] notes a connection between familiarity, liking and the amount of smiling while listening to music. Also, Hamlen and Shuell [36] find a positive correlation between liking and familiarity for classical music excerpts, which increases when an associated video is accompanied by audio. Similar effects could be tested with emotional movie clips via ASCERTAIN.

Finally, the importance of using less-intrusive sensors in affective studies has been widely acknowledged [68, 127]. Minimally invasive and wearable sensors enable naturalistic user response, alleviating stress caused by cumbersome clinical/lab-grade equipment. Choosing minimally invasive sensors is especially critical when complex behavioral phenomena such as emotions are the subject of investigation. While most available affective datasets have been compiled using lab equipment [127], ASCERTAIN represents one of the first initiatives to exclusively employ wearable sensors for data collection, which not only enhances its ecological validity, but also repeatability and suitability for large-scale user profiling.

3.2 Signal Quality Matters - QAMAF

This section covers an approach to take into the consideration the quality of input signals for enhancing the accuracy of multimodal emotion recognition results. We propose a Quality Adaptive Multimodal Affect Recognition System (QMAF) for user-centric multimedia indexing. ⁸.

⁸The work has been published in the proceedings of the ACM Conference on Multimodal Retrieval, 2016 [33] and the jointly developed dataset is publicly available to the research community:

mhug.disi.unitn.it/wp-content/QAMAF/QAMAF.html

3.2.1 Introduction

This work presents a multimodal approach on a hard implicit affective-indexing problem. We tackled cross-user and user-centric implicit affective-tagging of (weakly-affective) and short music snippets when the information sources are damaged by various noise artifacts. We achieved significantly above chance results for classification of user perceptions on affective music snippets and we also found that head movements encode *liking* perception in response to music snippets. We made the dataset publicly available for research community so that other researcher can improve our methods.

Two general approaches of generating affective tags for multimedia content are (i) using the information content of multimedia [37] and (ii) using the human affective perception (via detecting users' emotions) to tag the perceived content (implicit user-centric approach [52]). Certainly a hybrid approach [53, 57] could be successfully utilized. This work follows the second approach in a multimodal scheme.

Affective computing techniques have been successfully utilized for detection of users' emotions in response to multimedia content [140, 53, 57, 109] by leveraging the information of a plethora of modalities including facial expressions, gestures, body postures, voice, heart activities, electrodermal signals, and brain responses.

Since many of the underlying affective patterns in the above mentioned modalities are highly subjective (i.e. they significantly vary from one user to another), most of the state of the art user-centric emotion recognition studies focus on subject-based emotion recognition when the training and the test data for evaluation of a scheme comes from the same user [53, 57, 109]. However, a scalable implicit affective indexing system should be able to classify human emotions for any *unseen* user. Such framework where the test data is from unseen users is called *cross-user*.

A multimodal system for affect recognition is expected to perform better than a unimodal system, as reported in previous studies [53, 34, 57, 109], which can be attributed to the fusion of complimentary information provided by each modality.

However, the signals from the above mentioned modalities are often contaminated with various sources of noise [88], that significantly hinders the task of affect recognition. Particularly, in real-world biomedical signals, for e.g. signals obtained from wearable devices, the problem of noise contamination is more exaggerated.

This paper presents the first steps towards the validation and development of a user-centric multi-modal quality adaptive affect recognition system for cross-user implicit affective indexing of multimedia content.

Experimental Setup

We recruited a total of thirty-three participants (with age distribution of 29.7 ± 5.4 years, 21 males) for the study. The participants were asked to listen to music excerpts, originally generated by *Robin* [78], an algorithmic composer that generates western classical-like music with affective connotation in real time. The stimuli are weakly affective being evident from the fact that facial expressions in the dataset are negligible if not absent.

The participants experienced music excerpts using a AKG K512 headphone inside a silent room at the University of Trento, Italy. At the beginning of the experiment, four training excerpts were played to the participants to make the subjects familiar with the task. During the experiment, participants were presented with twenty thirty-two seconds long music excerpts in random order. While experiencing the music excerpts, participants' physiological signals, including electrocardiography (ECG), galvanic skin response (GSR), frontal electroencephalography (EEG), and facial videos were recorded. The ECG and GSR signals were recorded using the Shimmer sensors, whereas for recording EEG,

we used a NeuroSky Mindwave headset. Moreover, participants' facial videos were recorded using an A4Tech webcam at a resolution of 640 X 480 pixels and the SDM face alignment method[132] is employed to detect users' head poses and facial landmark tracks. All the recordings are available online via the dataset website⁹.

In order to measure valence, arousal and liking, participants were asked to rate them on three seven-point semantic differential scales, from 1 (negative, relaxing or unlike) to 7 (positive, exciting or like). To record the levels of arousal, valence and liking, participants were asked to type in numbers between 1-7 on a keyboard after listening to each excerpt. Moreover, to reduce emotional bias, a sequence of randomly generated notes were played, from a set of five 15 second long pre-recorded snippets, between each music excerpt. The obtained subjective scores showed substantial inter-rater agreement as measured using intra-class correlation (ICC), where ICC for arousal, valence and liking was 0.79, 0.63 and 0.81, respectively.

Signal Quality Estimator Development

Towards developing a quality adaptive affect recognition system signal quality estimators (SQEs) were developed for each modality. In order to realize the SQEs, first, the quality of the signals was assessed by two expert annotators, forming the ground truth. We employed the Cohen's kappa to measure inter-rater agreement over the annotations of the experts and measures of 0.96, 0.98, 0.94, 0.73 were observed over the quality annotations for ECG, Facial detections, ECG, GSR, respectively. The first three are indicating *almost perfect agreement* (≥ 0.9) and the agreement on GSR quality is *substantial* (≥ 0.7). The marginal disagreement on the GSR quality could be due to two issues: (i) GSR has a slow response so that the structures of GSR signals sometimes are not recognizable over short recordings, and (ii) noise artifacts that are present

⁹mhug.disi.unitn.it/wp-content/QAMAF/QAMAF.html

on the signal sometimes induce patterns that are similar to the structure of GSR responses. In cases of disagreement between the two annotators, the annotation of the second expert is employed.

Table 3.11: Features used for quality estimation and affect recognition for each modality.

Modality	Quality Estimation	Affect Recognition
NeuroSky EEG	Statistical measures (such as, mean, median, skewness, kurtosis) for EEG data, power spectral features in ranges (0-1, 0-4, 4-8, 8-12, 12-30, 30-50, 58-62 Hz)	Band powers for δ , θ , α , β and γ bands, statistical measures for cognitive measures provided by NeuroSky
ECG	Statistical measures for heart rate and heart rate variability, power spectral features from ECG for ranges (5-15, 5-40, 0-40, 1-40 Hz) and their ratios	Wavelet based power spectral features over ECG and HRV, statistical measures for the spectral features and Poincare features
GSR	Statistical measures for raw GSR signal, and GSR signal band-passed between ranges (0-0.08, 0.08-0.2, 1-2, 1-2, 10-20, 20-30 Hz)	Power spectral features, rise time, fall time and zero crossing rate for very low frequency (≤ 0.08 Hz) and low frequency (0.08 – 0.2 Hz) components of the signal
Face/Head-pose	Features encoding information regarding lips thickness, ratios, such as upper lip to lower lip thickness, eye brows width to lips width	Statistical measures, rise time, drop time and zero crossing rate for head's pitch, yaw and roll

Features listed in Table 3.11 were extracted, as they encode signal quality information for each modality. Finally, a bagging classifier with decision trees as a base estimator was implemented to differentiate between good/bad quality signals for each modality. The classification performance was validated using a leave-one-subject-out cross-validation i.e., the classifier was trained on samples from all the subjects except the one used for testing. Moreover, the performance of the SQEs was assessed using a weighted F1-score [87], as it accurately measures the classifier performance for a highly imbalanced classification problem.

3.2.2 Affect Classifier Development

As a next step, features (listed in Table 3.11, column labeled ‘Affect recognition’) that encode affect [53], were extracted. The affect encoding features were then employed with linear support vector machine (SVM, $C = 1.0$) and Naive Bayes (NB) classifiers for three binary classification problems of differentiation between high/low valence, arousal and liking, respectively. For validation of the developed classifiers, a cross-subject classification approach was adopted

using a leave-one-subject-out cross-validation [27]. Due to high intra-variability among human signal patterns, cross-subject affect recognition is a *harder*, but more *generalizable*, problem than subject specific affect recognition (as used in [53]). In this study we validated our results using cross-subject cross-validation.

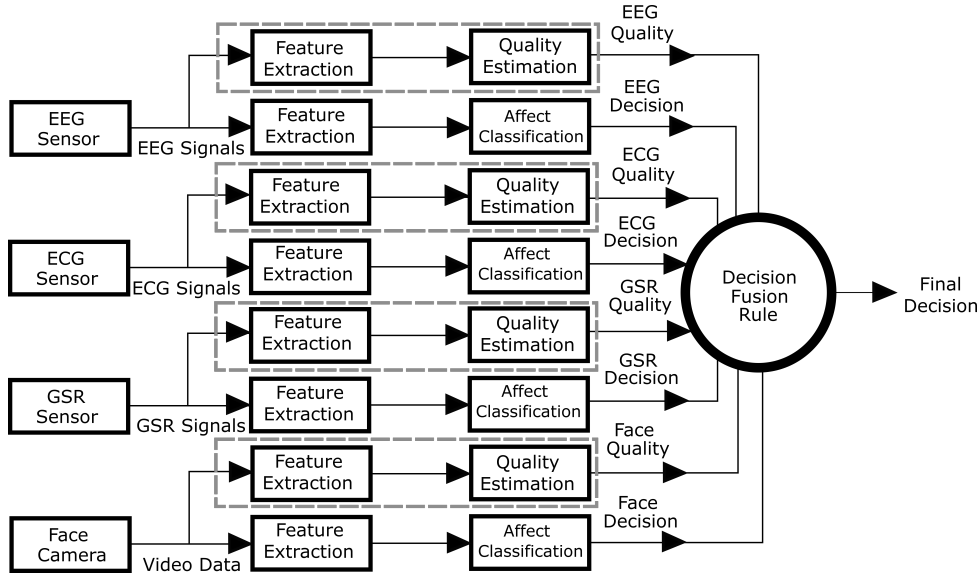


Figure 3.8: Quality adaptive multimodal decision level fusion schema.

In our proposed quality adaptive affect recognition scheme, during the development of affect binary classifiers, good quality samples form the train data and SQEs are applied to assess the quality of the test samples. The SQEs provided the reliability of the predictions thus, assisting in rejecting the bad quality samples. For accurate measurement of classifier performance for an imbalanced classification problem, we computed weighted F1-scores [87]. The performance of each classifier developed above was tested for significance against random voting using a paired *t-test*.

3.2.3 Quality Adaptive Multimodal Fusion

The quality adaptive multimodal decision fusion affect recognition system in this study is developed by modifying the decision fusion scheme presented in

[58]. In our study, we used a decision fusion classifier for multimodal fusion as depicted in Fig. 3.8. In decision level fusion, a linear combination of the individual uni-modal classifiers' outputs is calculated as the output. The decision fusion can be implemented (i) using an equal weight scheme i.e., all modalities used for affect recognition being given equal weights or (ii) using an optimal weight scheme i.e., the weight for each modality can be optimised given a set of training data [58]. The weights indeed encode the relative importance of the unimodal classifiers in calculation of the final output. The formulation and implementation of the fusion scheme that is employed in this study are described below.

Formulation

For the decision level fusion scheme, the fusion classification probability $p_0^x \in [0, 1]$ for each class $x \in \{1, 2\}$ can be denoted by

$$p_0^x = \sum_{i=1}^N \alpha_i p_i^x q_i t_i \quad (3.1)$$

where, i is the index of a particular modality used for affect recognition, N is the number of modalities used, α_i are the weights corresponding to each modality ($\sum_{i=1}^N \alpha_i = 1$), q_i ¹⁰ corresponds to the quality of the respective modality and t_i is the normalized training set performance for a particular modality, such that the fusion probabilities for all classes sum up to 1, and is given by

$$t_i = \frac{F_i}{\sum_{i=1}^N \alpha_i q_i F_i} \quad (3.2)$$

where, F_i is the F1-score obtained on the training set using a particular modality and $F_i \in [0, 1]$. Then, it can be shown that,

¹⁰ $q_i \in \{1, 0\}$: i.e. good or bad according to the output of the SQE binary classification for the i^{th} modality

$$p_0^1 + p_0^2 = \sum_{i=1}^N \alpha_i q_i t_i = \sum_{i=1}^N \left(\frac{\alpha_i q_i F_i}{\sum_{i=1}^N \alpha_i q_i F_i} \right) = 1 \quad (3.3)$$

Implementation

The quality adaptive fusion scheme described above was implemented using equal weights for EEG, ECG, GSR and head pose. Therefore, the weights used for fusion were $\alpha_i = 0.25$ and the class probabilities from each single modality is given by,

$$p_Q^x = 0.25 \times (p_{ee}^x q_{ee} t_{ee} + p_{ec}^x q_{ec} t_{ec} + p_{gs}^x q_{gs} t_{gs} + p_{hp}^x q_{hp} t_{hp}) \quad (3.4)$$

where subscripted ‘Q’ denotes quality adaptive system, abbreviations ee, ec, gs and hp denote EEG, ECG, GSR and head pose, respectively. Whereas, for non-adaptive multimodal fusion was developed using similar decision rule while excluding the quality term ‘ q_i ’ from the equation resulting in,

$$p_{nQ}^x = 0.25 \times (p_{ee}^x t_{ee} + p_{ec}^x t_{ec} + p_{gs}^x t_{gs} + p_{hp}^x t_{hp}) \quad (3.5)$$

where subscripted ‘nQ’ denotes non-adaptive system for multimodal fusion.

3.2.4 Results

The SQEs for each modality performed adequately for each modality, as the weighted F1-scores were as follows: EEG - 0.93, ECG - 0.95, Face/headpose - 0.86 and GSR - 0.78, while the weighted F1-score for random voting for each modality was 0.62. Moreover, the performance of the affect classifiers (both, quality adaptive and non-adaptive) are reported in Table 3.12, which can be compared to the baseline weighted F1-score of 0.50, obtained from random voting for valence, arousal and liking. It was observed that, using the quality adaptive affect recognition system, for arousal, all the modalities performed significantly better than chance. Moreover, for valence EEG, ECG and face/headpose

produced significant results whereas, for liking only face/headpose using a SVM classifier performed significantly better than chance. The non-adaptive affect recognition systems resulted in very few significant results. However, it should be noted that *uni-modal* quality adaptive systems had higher failure rate due to sample rejections. The non-adaptive multimodal decision fusion produced significantly better than chance classification performance for the three subjective dimensions of valence, arousal and liking, using the NB classifier whereas, no significant results were obtained using SVM classifier. Furthermore, the quality adaptive multimodal decision fusion also produced significantly better than chance classification performance for valence, arousal and liking, using the NB classifier whereas, using SVM classifier significant results were observed only for valence classification. Moreover, the sample rejection rate was brought down significantly, to 0%, using the quality adaptive multimodal fusion.

Table 3.12: Classification results for the self-assessment of valence (V), arousal (A) and liking (L) using a leave-one-subject-out cross-validation schema. The weighted F1-scores significantly higher than chance level (0.50) are highlighted (superscripted * : $p < 0.05$). The table also lists the percentage of samples rejected in quality adaptive schema for each modality.

Modality	Classifier	Quality Adaptive				Non-Adaptive		
		A	V	L	Rejections	A	V	L
NeuroSky EEG	SVM	0.57*	0.53	0.54	22.58%	0.52	0.50	0.51
	NB	0.56*	0.57*	0.53		0.52	0.50	0.51
ECG	SVM	0.59*	0.57*	0.53	18.33%	0.57*	0.52	0.50
	NB	0.54	0.57*	0.54		0.55	0.55*	0.51
GSR	SVM	0.52	0.52	0.46	13.79%	0.51	0.52	0.52
	NB	0.55*	0.48	0.54		0.54	0.52	0.53
Headpose	SVM	0.55*	0.58*	0.58*	5.76%	0.56*	0.54	0.58*
	NB	0.50	0.55*	0.53		0.53	0.52	0.55
Decision Fusion	SVM	0.60*	0.59*	0.58*	0%	0.57*	0.54	0.58*
	NB	0.57*	0.58*	0.56*		0.56*	0.56*	0.55*

3.2.5 Discussion and Conclusion

The developed SQEs performed well for each modality (weighted F1-score of about 0.90). However, for GSR the obtained weighted F1-score was relatively lower than the other modalities suggesting that better features could aid in improving the GSR signal quality estimation. Moreover, the advantage of using a quality adaptive affect recognizer is evident from Table 3.12, where the quality adaptive affect recognizer produced more number of significant results compared to a non-adaptive affect recognizer. However, a higher percentage of sample rejections for *uni-modal* quality adaptive systems resulted in their higher failure rate. The efficacy of multi-modal fusion techniques, both quality adaptive and non-adaptive, is evident as both techniques produced significant results for all three affective dimensions. The quality adaptive multimodal fusion has an added advantage of decreasing the failure rate resulting from quality adaptive uni-modal systems and achieving slightly higher performances. Moreover, the results reported in Table 3.12 validate the performance of our approach for *cross-user* affect recognition in noisy recordings (e.g. due to noise in environment) for multimedia implicit affective indexing.

Furthermore, it is worth noting that quality adaptive arousal and valence classifiers performed significantly above chance on all the modalities except on valence recognition using GSR that is in corroboration with the finding in [55]. It is worthy to mention that low unimodal performances on GSR could be due to the fact that GSR responses are slow. Therefore, GSR is an unsuitable modality for an experiment with short recordings like ours.

According to the observed results, a link between participants' head-pose (e.g. following the rhythm of music) and the likeability of a music excerpt was observed as the liking classifiers developed using head-pose features performed significantly above chance. Head-pose also significantly encodes valence and arousal perceptions.

3.3 Conclusion

We present ASCERTAIN– a new multimodal affective database comprising implicit physiological responses of 58 users collected via commercial and wearable EEG, ECG, GSR sensors, and a webcam while viewing emotional movie clips. Users’ explicit affective ratings and big-five personality trait scores are also made available to examine the impact of personality differences on AR. Among AR datasets, ASCERTAIN is the first to facilitate study of the relationships among physiological, emotional and personality attributes.

The personality–affect relationship is found to be better characterized via non-linear statistics. Consistent results are obtained when physiological features are employed for analyses in lieu of affective ratings. Finally, AR performance superior to prior works employing lab-grade sensors is achieved (possibly because of the larger sample size used in this study), and above-chance personality trait recognition is obtained with all considered modalities. Personality differences are better characterized by analyzing responses to emotionally similar clips, as noted from both correlation and recognition experiments. Finally, RBF SVM achieves best personality trait recognition, further corroborating a non-linear emotion–personality relationship.

We believe that ASCERTAIN will facilitate future AR studies, and spur further examination of the personality–affect relationship. The fact that personality differences are observable from user responses to emotion-wise similar stimuli paves the way for simultaneous emotion and personality profiling. As recent research has shown that AR is also influenced by demographics such as *age* and *gender* [118], we will investigate correlates between affective physiological responses and the aforementioned **soft-biometrics** in future, coupled with a deeper examination on the relationship between personality and affect. We will also investigate how *a-priori* knowledge of personality can impact the design of user-centered affective studies.

In the QAMAF study presented here, we developed a quality adaptive multimodal affect recognition system for cross-user and user-centric implicit multimedia indexing. The multimodal system was developed using data from four different modalities of EEG, ECG, GSR and face/headpose videos while users experienced affective music generated by an algorithmic composer, *Robin*. The signal quality for each modality was estimated first using a bagging classifier, which was followed by affect recognition. The quality adaptive uni-modal affect recognition performed better than chance however, these systems resulted in high failure rate due to bad quality sample rejection. Towards decreasing the failure rate of the uni-modal affect recognition systems, we proposed a quality adaptive multimodal decision fusion rule, giving equal weights to each modality, which performed adequately for affect recognition while lowering the failure rate. However, to improve the classification performance, quality adaptive modality weight optimisation should be explored in future studies. We release the dataset for the community, so that researchers in the community would improve our baseline results employing more innovative signal-quality adaptive methods.

Chapter 4

Crowdsourcing Continuous Affective Annotations for Video Tagging

Affective video tagging has been acknowledged as an important multimedia problem for long, given its utility for applications such as personalized media recommendation. However, most content and user-based media tagging approaches seek to recognize the *general* emotion of a stimulus (typically a movie or audio/video music clip), and only a few methods such as [37] have attempted to determine the dynamic of time-continuous emotion profile in the stimulus. This limitation is partly attributed to the fact that *interpreting* and *measuring* emotion is an inherently difficult problem—emotion is a highly subjective feeling, and the discrepancy between the emotion *envisioned* by the content creator versus the actual emotion *evoked* in consumers has been highlighted by many works. Also, learning the relationship between low-level content (typically in the form of audio-visual effects) and the high-level emotional feeling over time requires extensive training data, with annotations typically performed by multiple annotators for reliability, which is both difficult and expensive to acquire.

Recently, *crowdsourcing* (CS) has become popular for performing tedious tasks through extensive human collaboration via the Internet. When it is difficult to employ experts for analyzing large-scale data, CS is an attractive alternative, as many individuals work on smaller data chunks to provide useful informa-

tion in the form of annotations or tags. CS has been successfully employed to develop data-driven solutions for computationally difficult problems in multiple domains like natural language processing [129], and computer vision [137]. Two reasons mainly contribute to the success of CS– (1) crowd workers are paid a fraction of the wages that experts are entitled to, thereby achieving cost efficiency, and (2) the experimenter’s task becomes scalable when the original task is split into smaller and manageable micro-tasks and distributed among crowdworkers. Nevertheless, cost-effectiveness in CS is achieved at the expense of expertise– crowdworkers may lack the technical and cognitive skills or the motivation to effectively perform a given task [94]. Therefore, efficient methodologies that are robust to noisy data are crucial to the success of CS approaches.

In this study, we also propose Multi-task learning (MTL) for time-continuous valence, arousal (VA) estimation from movie scenes for which dynamic emotion annotations are acquired from crowdworkers. Given a set of *related* tasks, MTL seeks to *simultaneously* learn all tasks by modeling the similarities as well as differences among them to build task-specific classification or regression models. This joint learning procedure accounting for task relationships leads to more efficient models as compared to learning each task independently. For the purpose of learning the relationship between low-level audio-visual features and corresponding crowdworker VA annotations over time, we ask the following questions: (1) Given that emotion perception is highly subjective, and biases relating to crowdworker demographics may additionally exist, can we discover any patterns relating to their dynamic emotional perception? The exercise of seeking to acquire a *gold standard* annotation for each movie scene (or clip) via crowdsourcing is meaningful only if such patterns can be discovered. (2) If the emotional ground truth corresponding to each movie clip can be represented by a single, gold standard emotional profile, can we discover corresponding audio-visual correlates for a movie clip collection (as against a single

movie clip), which in turn, can be more effective for predicting the VA profile of a novel clip? Through extensive experiments, we demonstrate that MTL effectively answers the above questions, and is superior to single-task learning for VA prediction in novel scene segments. To summarize, this chapter makes the following contributions:

1. This is the first work to employ MTL for time-continuous emotion prediction.
2. This is also one the first work to attempt dynamic affect prediction for movie clips.

The chapter is organized as follows: Section 4.1 overviews the literature. Experimental protocol employed for recording crowdworkers' affective responses is described in Section 4.2. An application of a multi-task learning framework on the recorded dataset is provided in Section 4.5. Annotation data analysis and emotion prediction experiments are presented in Section 4.5.1, and conclusions are stated in Section 4.6.

4.1 Related work

We now examine related work on (1) Crowdsourcing, (2) Affective movie analysis, (3) CS for affective media tagging and (4) Multi-task learning.

4.1.1 Crowdsourcing

Steiner *et al.* [112] defined three types of video events and showed that these events can be detected from video sequences via crowdsourcing upon combining textual, visual and behavioral cues. Vondrick *et al.* [124] argued that frame-by-frame video annotation is essential for a variety of tasks, as in the case of time-continuous emotion measurement, even if it is difficult for human annotators. An online framework to collect valid facial responses to media content

was proposed in the work of McDuff *et al.* [73], who found significant differences between subgroups who liked/disliked or were familiar/unfamiliar with a particular commercial.

4.1.2 Affective movie analysis

A primary issue in affective multimedia analysis is the paucity of reliable annotators to generate sufficient training data and in most studies, only few annotators are used [107, 126]. Also, emotion perception varies with individual traits such as personality [43], and significant differences may be observed in affective ratings compiled from different persons over a small population. To address this problem, a number of studies have turned to crowdsourcing. In a seminal study affective movie study, Gross *et al.* [32] compiled a collection of movie clips to evoke eight emotional states such as anger, disgust, fear and neutral based on emotion ratings compiled for 250 movie clips from 954 subjects.

4.1.3 Crowdsourcing for affective media tagging

Soleymani *et al.* [108] performed crowdsourcing on a limited scale to collect 1300 affective annotations from 40 volunteers for 155 Hollywood movie clips. In another CS-based affective video annotation study, Soleymani *et al.* [105] compiled annotations for the MediaEval 2010 Affect Task Corpus on AMT, and asked workers to self-report their boredom levels. In a recent CS-based media tagging work, Soleymani *et al.* [106] presented a dataset of 1000 songs for music emotion analysis, each annotated continuously over time by at least 10 users. Nevertheless, movies denote multimedia stimuli that best approximate the real world and movie clips have been found to be more effective for eliciting emotions in viewers as compared to music video clips in [52], and that is why we believe continuous emotion prediction with movie stimuli is important in the context of affective media representation and modeling.

4.1.4 Multi-task learning

Recently, multi-task learning (MTL) has been employed in several computer vision applications such as image classification [136], image annotation [92] and visual tracking [141]. Given a set of related tasks, MTL [18] seeks to simultaneously learn a set of task-specific classification or regression models. The intuition behind MTL is simple: a joint learning procedure which accounts for task relationships is expected to lead to more accurate models as compared to learning each task separately. While MTL has been used previously for learning from noisy crowd annotations [42], we present the first work that employs MTL for time-continuous emotion prediction from movie clips.

4.2 Experimental Protocol

In this study, we asked crowd workers to continuously annotate 12 emotional movie scenes adopted from [52] via a web-based user interface— they were not allowed to access the scene content prior to the rating task. Our objective was to understand and model their emotional state over time, as they viewed the movie clips.

4.2.1 Dataset

We selected 12 video clips from [52] equally distributed among the four quadrants in the valence-arousal space. Table 4.1 presents characteristics of video clips from the different quadrants. All video clips were hosted on YouTube for access during the CS task.

4.2.2 Experimental Protocol

We posted the annotation task on Amazon Mechanical Turk (AMT) and other CS channels via the CrowdFlower (CF) platform. CF is an intermediate plat-

Table 4.1: Video clip details. HALV, LALV, HAHV and LALV respectively correspond to high-arousal low-valence, low-arousal low-valence, high-arousal high-valence and low arousal-low valence labels.

	HALV	LALV	HAHV	LAHV
No. of video clips	3	3	3	3
Min. length (sec)	79	80	86	59
Max. length (sec)	91	121	109	92
Avg. length (sec)	86.66	97.33	101	76.33

form for posting the AMT task on our behalf. Moreover, CF provides a simple gold standard qualification mechanism to discard outliers. If workers passed the qualification test, they were considered qualified to perform a given task. However, pre-designed tests are very generic and limited to simple tasks, which do not allow for trivially discarding low quality annotations. So, we performed PHP server-side scripting and redirection, collection and evaluation of all annotations real-time on our server via HTTP requests, before letting workers submit the task. The architecture of the designed CS platform is shown in Fig. 4.1.

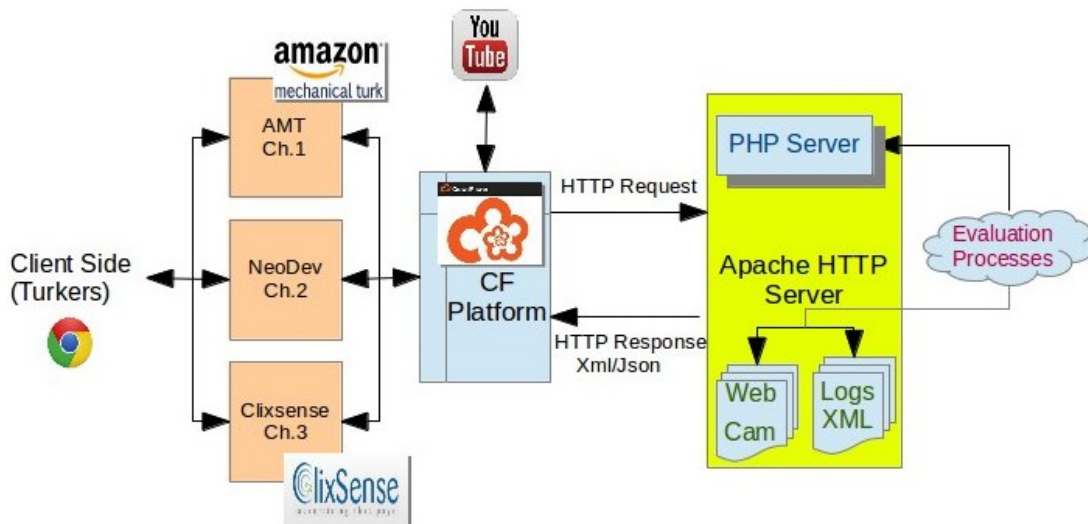


Figure 4.1: Architecture of the designed Crowdflower platform (Turkers are the AMT crowd workers).

To ensure annotation quality, each crowd worker could only annotate 5 video

clips, and at least 15 judgments were requested and collected for each video clip. We also recorded facial expressions of workers (not used in this work) as they performed the annotation. Informed consent was obtained from workers and, prior to the task, workers had to provide their demographics (age, gender and location). Time-continuous valence/arousal annotations from workers were compiled over separate sessions (a worker need not annotate for both valence and arousal for the same clip under this setting), and workers were also required to rate each clip for overall emotional valence or arousal. Each worker was paid 10 cents per video as remuneration upon successful task completion.

Workers did not get paid if their annotations and webcam facial videos were not recorded on our server. To evaluate the annotation quality, each video annotation was logged in XML format and analyzed. A continuous slider was used to record emotional rating, and if the slider had not moved for more than 80% of the clip duration, or if more than 20% of the data was lost, the annotation was automatically discarded. Also, files smaller than a pre-defined threshold were discarded. If the annotation task was left incomplete, a warning message notified the worker about the missing annotations. Workers could then re-annotate the missing videos and get paid. For motivating workers to provide good quality annotations, we rewarded them with online gift vouchers if they provided high-quality annotations. Furthermore, we introduced some constraints such as: (1) Workers could not play (or rate) multiple video clips simultaneously. (2) Workers could annotate a video as many times as they wanted to. (3) Workers were allowed to use only the Chrome browser for annotation due to unavailability of HTML5 technology support in other browsers. (4) Media player controllers were removed from the interface so that workers could not fast forward/rewind the movie clips, and finally, (5) If the annotation stopped midway, it had to be redone from scratch.

4.2.3 Annotation Mechanism

A screen shot of the user interface for recording annotations is presented in Fig. 4.2. The following components were part of the continuous annotation and facial expression recording process.

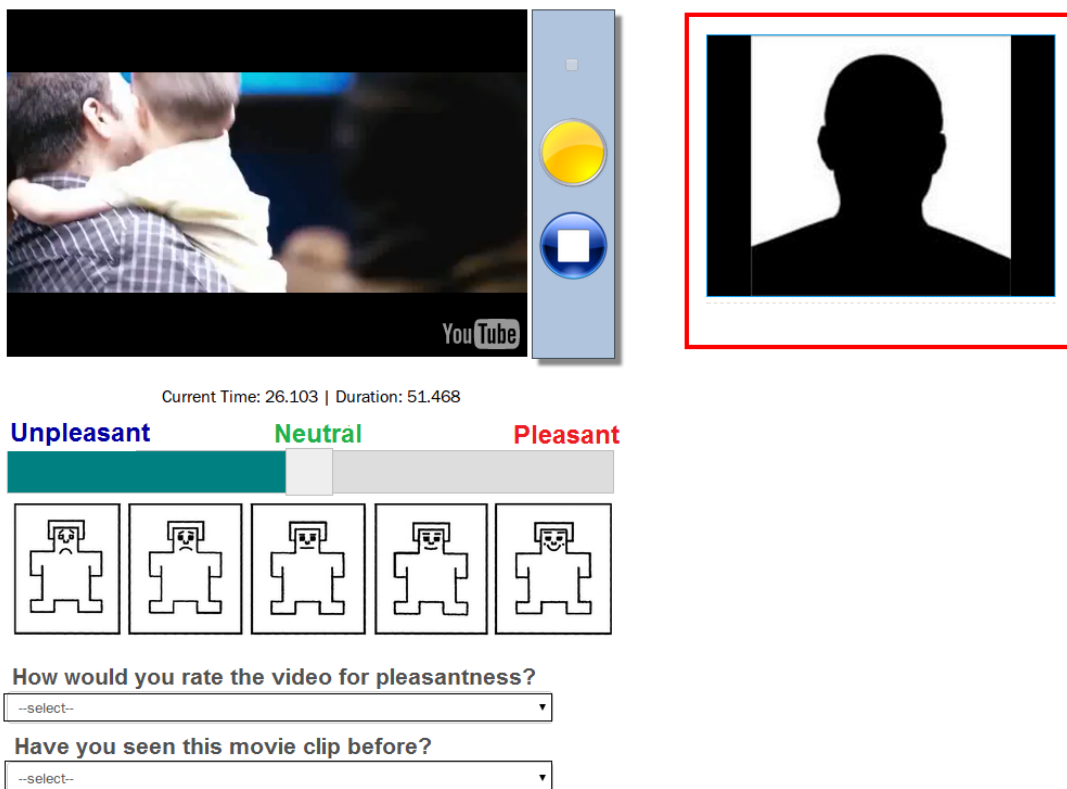


Figure 4.2: User-interface for recording workers' emotional ratings and facial expressions.

Video Player: To provide an uninterrupted video stream for workers with low bandwidth, we uploaded all movie clips onto YouTube. On the client-side, YouTube JavaScript player API was integrated and used in our web-based user interface.

Slider: The slider was used to collect the time-continuous VA ratings of workers while watching the video clips. The slider values ranged from -10 to 10 (*very unpleasant* to *very pleasant* for valence, and *calm* to *highly excited* for

arousal) for both factors. In order to facilitate workers' decision making, a standard visual scale Self Assessment Manikin (SAM) image was displayed to the workers.

Webcam Panel: To upload facial expression of workers in real-time, we used HTML5 technology to buffer the worker's webcam recording on the client-side when the play button was pressed. The buffered video was automatically uploaded in compressed, VP8 open codec format on our server when the video clip finished playing. Videos were recorded at 320x240 resolution, 30 fps.

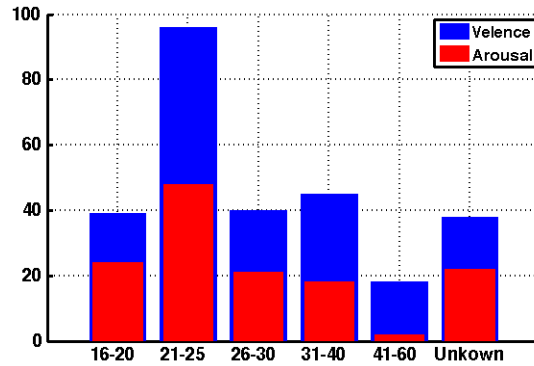
Questionnaires: Workers needed to report (1) their overall emotional (valence or arousal) rating for the movie clip on a scale of -10 to 10, and (2) their familiarity with the clip to avoid the effect of such bias on their ratings.

4.3 Multimedia Feature Extraction

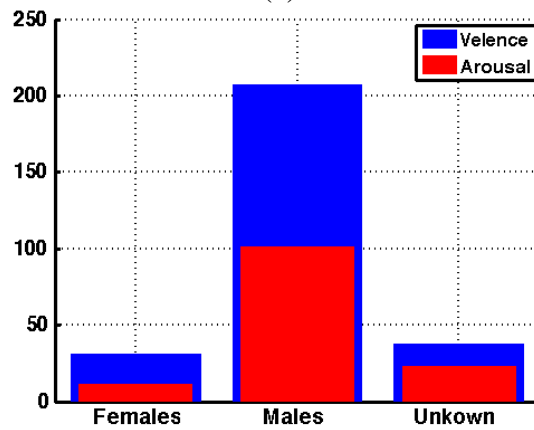
Inspired by previous affective studies [126, 91], we extracted low-level audio-visual features that have been found to correlate well with the VA dimensions. In particular, we extracted the features used in [37, 57] on a per-second basis for our regression experiments.

4.3.1 Video Features

Lighting key and color variance [126] are well-known video features known to evoke emotions. Therefore, we extracted lighting key from each frame in the HSV space by multiplying the mean by the standard deviation of V values. Color variance [57] is defined as the determinant of the covariance matrix of L, U, and V in the CIE LUV color space. Also, the amount of motion in a movie scene is indicative of its excitement level [57]. Therefore, we computed the optical flow [69] in consecutive frames of a video segment to motion magnitude for each frame. The proportions of colors are important elements for evoking emotions [122]. A 20-bin color histogram of hue and lightness values in the



(a)



(b)

Figure 4.3: Age (a) and gender (b) statistics of crowdworkers for valence and arousal annotation tasks

HSV space was computed for each frame of a segment and averaged over all frames. The mean of the bins reflect the variation in the video content. For each frame in a segment, the median of the L and S values in HSL space were computed; their average for all the frames of a segment is an indication of the segment lightness and saturation [57]. We also used the definitions in [126] to calculate shadow proportion, visual excitement, grayness and visual detail. Extracted video features are listed in Table 4.2.



Figure 4.4: Locality distributions of crowdworkers.

4.3.2 Audio Features

Sound information in the form of loudness of speech (energy of sound) is related to arousal, while rhythm and average pitch in speech relates to valence [91], while Mel-frequency cepstrum components (MFCCs) [66] are representative of the short-term sound power spectrum. Commonly used features in audio and speech processing [66] were extracted from the audio channels. To extract MFCCs, we divided the audio segment into 20 divisions and then extracted the first 13 MFCC components from each division. Using the sequence of MFCC components over a segment, we computed 13 derivatives of MFCC, DMFCC, and mean auto correlation, AMFCC proposed in [66]. Upon calculating MFCC, DMFCC and AFCC (13 values each), we used their means as features. The implementation in [80] was used to extract formants up to 4400Hz over the audio segment, and formant means were used as features. Moreover, we used the ACA toolbox [64] to calculate mean and standard deviation(std) of (i) spectral flux, (ii) spectral centroid and (iii) time-domain zero crossing rate [66] over 20 audio segment divisions. We also calculated the power spectral density and the bandwidth, band energy ratio (BER) , and density spectrum magnitude (DSM) according to [66]. Finally, we also computed the mean proportion of silence as defined in [19]. All in all, 56 audio features listed in Table 4.2 were extracted.

Table 4.2: Extracted audio-visual features from each movie clip (feature dimension listed in parenthesis).

Audio features	Description
MFCC features (39)	MFCC coefficients [66], Derivative of MFCC, MFCC Autocorrelation (AMFCC)
Energy (1) and Pitch (1)	Average energy of audio signal [66] and first pitch frequency
Formants (4)	Formants up to 4400Hz
Time frequency (8)	mean and std of: MSpectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio [66]
Zero crossing rate (1)	Average zero crossing rate of audio signal [66]
Silence ratio (2)	Mean and std of proportion of silence in a time window [66, 19]
Video features	Description
Brightness (6)	Mean of: Lighting key, shadow proportion, visual details, grayness, median of Lightness for frames, mean of median saturation for frames
Color Features (41)	Color variance, 20-bin histograms for hue and lightness in HSV space
VisualExcitement (1)	Features as defined in [126]
Motion (1)	Mean inter-frame motion [69]

4.4 A Conditioned Crowd is Better Than the Expert

1012 and 527 workers participated in the valence and arousal rating experiments respectively. Their age and gender distributions are as shown in Fig. 4.3. Their locality distribution is also presented in Fig. 4.4. The first step towards processing worker annotations involves discarding outliers and bad quality annotations. This section describes (i) the adopted procedures for filtering crowd annotations without and with prior knowledge, and (ii) employed method to aggregate experts' and workers' dynamic annotations.

4.4.1 Quality control without prior knowledge

Evaluation of workers' reliability for any CS studies is associated with some prior knowledge available in the context of the study. In the first quality approach, we assume no prior knowledge regarding the emotional content of each movie clip.

We assume the annotation (A) of an AMT worker (T) for a certain video (V) is a vector of length l seconds denoted as: $A_{TV} = (a_1, a_2, \dots, a_l)$. We perform two levels of quality control to filter out bad annotations and in an unsupervised manner.

Quality control : Level 1

(i) Due to bad network connection, browser bugs or other issues, the server may miss annotations over some durations of a video. We normally expect to receive l annotations for an l -sec long video. If assessment for the i th second of the video is received, $a_i \in A_{TV}$ vector will be a real number in $[-10, 10]$. Otherwise, the value is set to NaN (Not a Number). We discarded those annotations which have more than threshold NaN values. This threshold was empirically set to $0.15l$.

$$Filter_{NaN}(A_{TV}) = \begin{cases} True : & \frac{\#_{NaN}(A_{TV})}{l} < \tau_{NaN} \\ False : & Otherwise \end{cases}$$

where $\#_{NaN}$ denotes number of NaN entries in A_{TV} . For those annotations that passed the NaN filter, and still had some NaN values, we replaced the NaN values with the immediately preceding and valid value, or 0 otherwise.

(ii) We expected some variation in the time-continuous affect annotations, as the videos for our study were chosen from a set of control stimuli. Those annotations that do reflect the emotion dynamics are filtered using the following criteria– (a) $Filter_{STD}$ discards an annotation A_{TV} if the standard deviation ($STD(A_{TV}) < \tau_{STD}$, set empirically to 1 (0.05×20)).

$$Filter_{STD}(A_{TV}) = \begin{cases} True : & std(A_{TV}) > \tau_{STD} \\ False : & Otherwise \end{cases}$$

(b) The percentage of times the slider is moved during the annotation task is an indicator of the annotator's *participation level*, and therefore we discarded A_{TV} if it does not change over a certain duration $t > \tau_{DYN}$, set empirically to 0.8l.

$$Filter_{DYN}(A_{TV}) = \begin{cases} False : & \sum_{i=1}^l \delta(a_{i+1}, a_i) > 0.8l \\ True : & Otherwise \end{cases} \quad \text{where } \delta \text{ denotes Kro-} \\ \text{necker delta, with } \delta(i, j) = 1, i = j \text{ and } 0 \text{ otherwise.}$$

(iii) Finally, we expect some degree of consistency between the time-continuous annotations of a worker (A_{TV}) and his/her overall emotional assessment (O_{TV}) of a video. If an annotator does not (a) provide any overall assessment to a video ($O_{TV} = NaN$), or (b) enter the overall assessment as neutral ($O_{TV} = 0$), we *did not* discard the annotation. The range for time-continuous annotations was $[-10, 10]$ and for the overall assessment it was $[-2, 2]$. If $O_{TV} \neq NaN$, then we accepted the annotation if and only if the $sign(O_{TV})$ is consistent with either the sign of the mean annotation value, or with the mean value for the last 5 seconds. The last five seconds were considered since the overall assessment is likely to be impacted by the workers' most recent emotional state.

$$Filter_{CNS}(A_{TV}, O_{TV}) = \begin{cases} True : & or \begin{cases} sign(mean(A_{TV})) * sign(O_{TV}) \geq 0 \\ sign(mean(A_{TV}(l-4:l))) * sign(O_{TV}) \geq 0 \end{cases} \\ False : & Otherwise \end{cases}$$

To smooth the peaks in the time-continuous annotations, we applied a Kaiser window whose width $w = 5$ sec and whose shape parameter, β equals 1, and then rescaled the annotations by dividing them by $w = 5$ (see Fig. 4.7(c)). The discarded annotation samples upon applying level 1 of quality control procedure are highlighted in yellow in Figure 4.5(a) and Figure 4.5(b). The above process

Algorithm 1 Quality Control - Level 1

```

procedure LEVEL 1 ANNOTATIONS(Annotations, List1Top)
    List1Top  $\leftarrow \emptyset$ 
    for all  $A_{TV}$  such that  $A_{TV} \in \textit{Annotations}$  do
         $\zeta_1 \leftarrow \textit{Filter}_{NaN}(A_{TV})$ 
         $\zeta_2 \leftarrow \textit{Filter}_{STD}(A_{TV})$ 
         $\zeta_3 \leftarrow \textit{Filter}_{DYN}(A_{TV})$ 
         $\zeta_4 \leftarrow \textit{Filter}_{CNS}(A_{TV})$ 
        if ( $\zeta_1$  &  $\zeta_2$  &  $\zeta_3$  &  $\zeta_4$ ) then
            List1Top  $\leftarrow \textit{List1}_{Top} \cup \{A_{TV}\}$ 
        end if
    end for
    return List1Top
end procedure ▷ Passed Level 1 annotations (List1Top)

```

is outlined in Algorithm 1.

Quality control : Level 2

The annotations that passed through the four filters in level 1 of the quality control process are shown in pink in the left plots of Fig. 4.5(a) and 4.5(b). Level 2 of the filtering process was based on three main assumptions: we assumed that (i) there is a certain reliable annotation (RA) that best reflects the affective content of the video in terms of arousal/valence, (ii) high quality annotations correlate significantly with RA, and finally (iii) the city-block distance (ℓ_1 - norm) between high quality annotations and RA will be low. We empirically set a threshold ($\tau_{CBD} = 4$), which is 20% of the annotation range.

As illustrated in Algorithm 2, we first estimated the reliable annotation as the median of the (currently compatible) annotations. In each iteration of a convergent loop, we discarded those annotations that do not significantly correlate ($p > 0.05$) with the current RA estimate. In each iteration, we also discarded the annotations whose ℓ_1 distance from the RA estimate was more than τ_{CBD} . The remaining annotations were then *compatible* with the current estimate. After

each iteration, we updated the RA estimate using only the *compatible* annotations and looped until (i) either only one annotation remained in $List2_{Top}$ or (ii) all $List2_{Top}$ annotations were *compatible* with the current RA estimate.

During the convergence process, it may happen that some annotations *compatible* with the final RA estimate get discarded. We retrieved all such *compatible* annotations from the accepted level 1 list ($List1_{Top}$) by tracing the above loop exactly once.

Algorithm 2 Quality Control - Level 2, RangeWidth is width of the annotation range, equals 20 in our study.

```

procedure TOPANNOTATIONS( $List1_{Top}, List2_{Top}$ )
     $List2_{Top} \leftarrow List1_{Top}$ 
     $RA \leftarrow median(List2_{Top})$ 
     $Iterate \leftarrow \mathbf{True}$ 
    while  $Iterate$  do
         $Iterate \leftarrow \mathbf{False}$ 
        for all  $A_{TV}$  s.t.  $A_{TV} \in List2_{Top}$  do
             $\rho \leftarrow corr(A_{TV}, RA)$ 
             $\delta \leftarrow \ell_1(A_{TV}, RA)$ 
             $\{A_{TV}\} \leftarrow (\delta > \tau_{CBD}) \mathbf{or} p(\rho) \geq 0.05 \mathbf{or} \rho < 0$ 
             $List2_{Top} \leftarrow List2_{Top} - \{A_{TV}\}$ 
             $Iterate \leftarrow \mathbf{True}$ 
        end for
         $RA \leftarrow median(List2_{Top})$ 
    end while

    for all  $A_{TV}$  such that  $A_{TV} \in List2_{Top} \setminus List1_{Top}$  do
         $\rho \leftarrow corr(A_{TV}, Estimate)$ 
         $\delta \leftarrow \ell_1(A_{TV}, Estimate)$ 
         $\{A_{TV}\} \leftarrow (\delta < \tau_{CBD}) \mathbf{or} p(\rho) \leq 0.05 \mathbf{or} \rho > 0$ 
         $List2_{Top} \leftarrow List2_{Top} + \{A_{TV}\}$ 
    end for
    return  $List2_{Top}$ 
end procedure

```

▷ Estimate the high quality annotations in a convergent loop

▷ RA to accept discarded high quality annotations

4.4.2 Quality control with prior knowledge

Prior knowledge about a video’s affective content can help enhance the quality of obtained CS annotations. To this end, we used the ground truth ratings concerning arousal and valence provided by [52] for the 12 videos used in the second quality control approach. From the ground truth ratings, we know whether a

CHAPTER 4. CROWDSOURCING CONTINUOUS AFFECTIVE ANNOTATIONS FOR VIDEO TAGGING

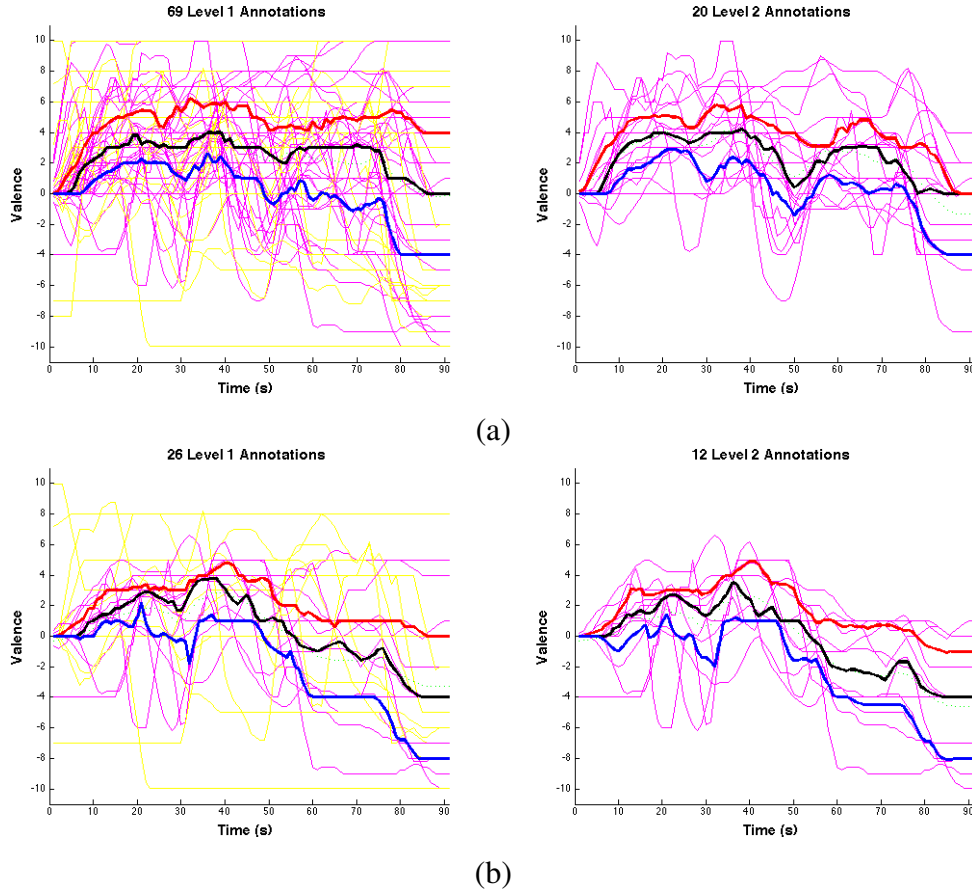


Figure 4.5: Filtering results using Quality Control without prior knowledge (a), and with prior knowledge (b). Yellow curves denote time-continuous annotations discarded upon applying Algorithm 1. In (a) and (b), magenta-colored curves in the left plot denote input to Algorithm 2, and output of Algorithm 2 in the right plot. The blue, black and red curves denote 1st , 2nd (median) and 3rd quartile of the annotation values.

certain video is perceived by the population as belonging to *positive* or *negative* valence, and *exciting* or *calm* arousal, which is employed to discard inconsistent worker annotations. Following this approach, we discarded those annotations for which (i) a worker did not provide any overall valence/arousal assessment, or (ii) the worker’s overall assessment O_{TV} is not consistent with the ground truth (OGT_V):

$$Filter_{GT}(A_{TV}, O_{TV}, OGT_V) =$$

$$\begin{cases} True : & \text{and} \begin{cases} O_{TV} = NaN \\ sign(O_{TV}) * sign(OGT_V) < 0 \end{cases} \\ False : & \text{Otherwise} \end{cases}$$

In the second quality control approach, we applied $Filter_{GT}$ before Algorithm 1 to get the *ground truth-compliant* CS annotations. Employing prior knowledge to remove annotation outliers produces fewer and cleaner annotations as seen from Fig. 4.5(b).

4.4.3 Aggregation of accepted annotations

All in all, we obtained 3 sets of annotations : (i) CS: crowdsourced annotations without employing prior knowledge, (ii) CSGT: crowdsourced annotations consistent with ground-truth ratings OGT_V , and (iii) Experts annotations (EXPT). These annotations were aggregated to obtain one representative time-continuous annotation per video that best reflects the dynamic emotions evoked in the (expert or worker) population over time. According to [75], two effective methods for aggregating time-continuous annotations involve using the mean and median of samples corresponding to a particular time point (we assume one annotation one second). Moreover, we use 4-quantiles (quartiles) to estimate (i) annotation upper-bound ($Q3$ -shown as red in Fig. 4.6), (ii) median-annotation ($Q2$ - shown with black in Fig. 4.6), and (iii) the lower bound ($Q1$ -shown with blue in Fig. 4.6). In Figure 4.6, we compare the CS, CSGT and EXPT outputs using the annotation mean (green) and median (blue) as suggested in [75]. We use $Q3$ and $Q1$ to estimate the reliability range for each annotation at a time point, and we expect more error in the estimated annotations for larger $Q3 - Q1$ values. Some statistics over $(Q3 - Q1)$ quartile differences are reported in Table 4.3. Evidently, the presence of more annotations results in a tighter estimate of the time-continuous emotion profile (CS and CSGT correspond to lower variance), which clearly demonstrates the merit of adopting a CS approach for affect analysis.

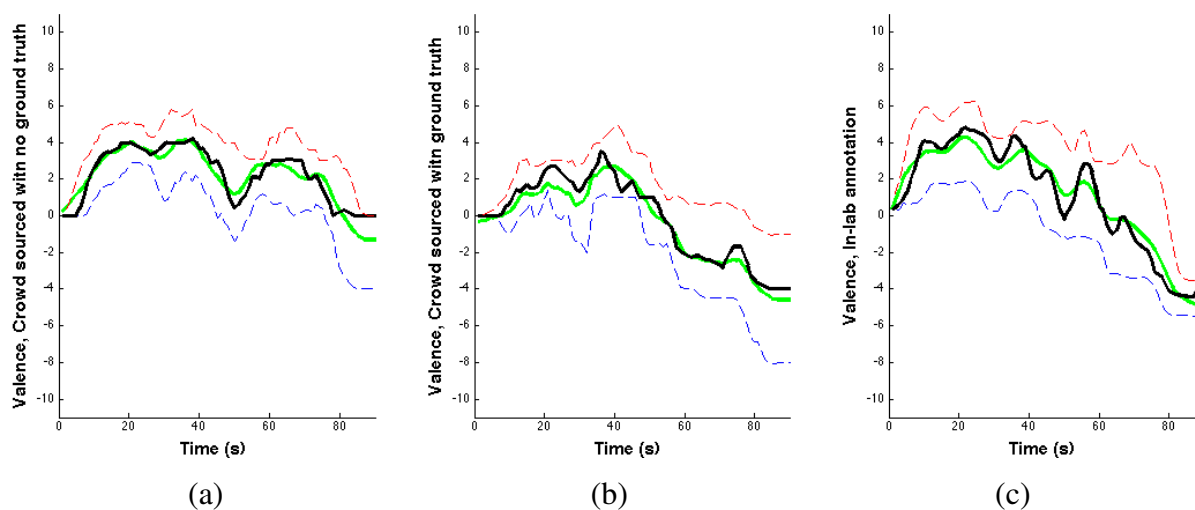


Figure 4.6: Aggregations results for (a) Crowdsourced annotations (CS) (b) Crowdsourced annotations compatible with ground truth (CSGT) and (c) Experts annotations (EXPT)

Table 4.3: Statistics over ($Q3 - Q1$) quartile differences for aggregated dynamic emotion annotations

Dimension	Type	mean	std	min	max
Valence	CS	3.232	1.129	0.000	6.490
	CSGT	3.276	1.230	0.000	7.091
	EXPT	2.954	1.430	0.097	8.404
Arousal	CS	3.226	1.138	0.000	7.491
	CSGT	3.563	1.274	0.000	8.939
	EXPT	5.241	2.190	0.431	13.621

4.4.4 Agreement Between Annotators

We also examined the level of inter-annotator agreement between the crowd and expert annotators, in order to assess the level of consistency in the clean annotations. Following [106], agreement for dynamic emotion ratings were computed using Kendal’s concordance coefficient, while Krippendor’s α was used for static emotion ratings. Results are presented in Table 4.4. Since only dynamic emotion ratings for the movie clips were collected from experts, expert agreement for static emotion results are not available.

Concerning dynamic emotion annotations, the expert agreement for both va-

4.4. A CONDITIONED CROWD IS BETTER THAN THE EXPERT

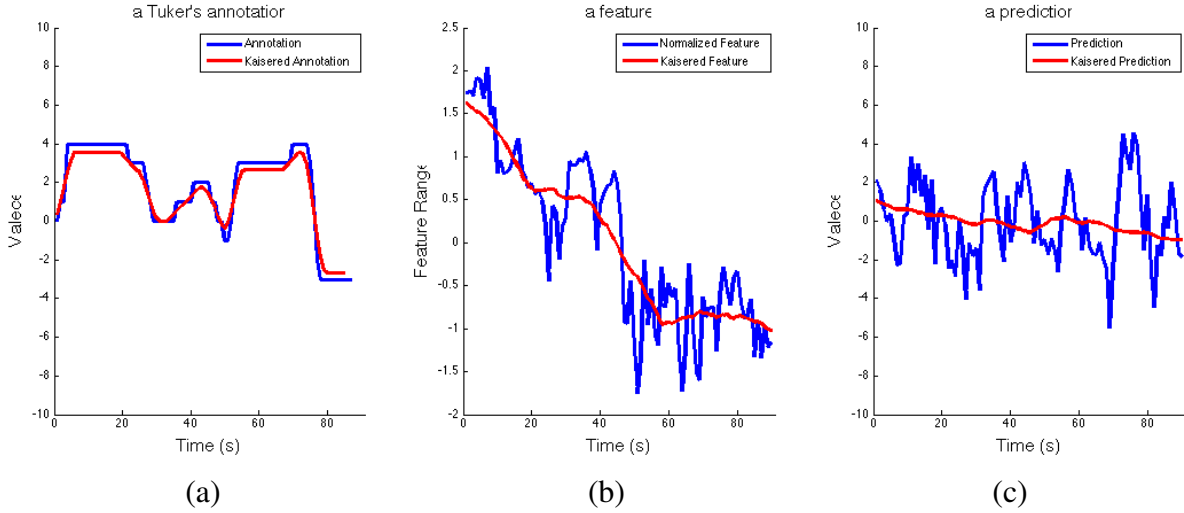


Figure 4.7: Kaiser window-based smoothing effect on annotation profile (left), feature profile (middle), and prediction profile (right)

Table 4.4: Inter-annotator agreement computed using coefficient of concordance (Kendall’s W) for dynamic emotion annotations and Krippendor’s α for static annotations.

	CS		CSGT		EXPT
	$\mu \pm \sigma$	α	$\mu \pm \sigma$	α	$\mu \pm \sigma$
Valence	0.19 ± 0.19	0.34	0.20 ± 0.19	0.37	0.45 ± 0.29
Arousal	0.22 ± 0.20	0.14	0.19 ± 0.19	0.17	0.59 ± 0.25

lence and arousal is much higher for experts as compared to crowd workers. While the variance in quartile differences over time was lesser for crowd workers as seen from Table 4.3, the general emotional agreement computed over raw annotations is still higher for experts. The inter-expert agreement for arousal is much higher than for valence, while hardly any difference is seen in the valence and arousal agreements for workers. Regarding static emotion ratings (available only for workers), a much higher agreement is noted for valence as compared to arousal.

4.4.5 Wisdom of Crowd vs. Experts

In this section, we introduce our method for comparing the quality of continuous affective annotations provided from 3 different methods namely, (i) crowdsourced annotations (CS), (ii) using the ground truth affective tags to analyze crowdsourced annotations (CSGT), (iii) and analyzing experts' annotations (acquired in-lab). We used the *glmnet* regression algorithm [102, 101] to predict continuous affective annotations of each video using a leave-one-video-out cross validation schema. State-of-the-art multimedia content analysis (MCA) features extracted from audio and video components of a movie clip [57] were input to *glmnet* trained with time-continuous emotion ratings provided by workers/experts. The algorithm proposed in [37] was used to preprocess the extracted features, and to post-process the predicted annotations. We report the correlation significance for predicted vs. actual annotations and the coefficient of determination (R^2) results for the predictions.

Experiment Schema:

We performed one-video-out cross validation and trained a regressor on the data of 11 videos ($V_{Tr} = \{v_j : j \neq i\}$ for a certain $1 \leq i \leq 12$) and predicted the continuous annotation for the test video ($V_{Ts} = \{v_i\}$). The mean dynamic emotion prediction performance is reported. In our model, we used the features and annotations from each second as training samples. $S_{Tr} = \{(f_j, t_j) : 1 \leq j \leq l, v_j \in V_{Tr}\}$, where $F_{Tr} = \{f_j : (f_j, t_j) \in S_{Tr}\}$ is the training feature set for the regression and $T_{Tr} = \{t_j : (f_j, t_j) \in S_{Tr}\}$ is the correspondent training target set. Similarly S_{Ts} , F_{Ts} , and T_{Ts} are defined respectively as test sample set, test feature set and test targets. Upon training the regression model, we tested the model over the test features (F_{Ts}) and get the *raw* predictions (E_{Ts}) for the test sample. Then we use R^2 , the coefficient of determination between the test targets (T_{Ts}) and the post-processed predictions (\acute{E}_{Ts}). In section 4.4.10, we explain the post processing method over the predictions.

4.4.6 Feature Preprocessing

Smoothing features over time

Hanjalic et al. [37] perform a preprocessing on the low-level feature fusion before training a regressor using a Kaiser window of $width = 700$ and a shape parameter $\beta = 5$. The purpose of applying the filter is to smooth the features to correlate better with continuous annotations. In their study, the predictions are at the frame level. Since in our case, the video frame rate is 30fps , we set the parameters of the Kaiser window to $width = \frac{700}{30}$ and $\beta = \frac{5}{30}$. The effect of applying such smoothing is shown in Figure 4.7(b).

Feature Normalization

We accumulated training samples and calculated the z -score of the feature sets. We used the $mean(\mu)$ and standard deviation (σ) obtained from normalized training samples: $\mu = mean(F_{Tr})$ and $\sigma = std(F_{Tr})$ to also normalize the test samples as $\frac{f_{(1,2,\dots,l)} - \mu}{\sigma} : f_{(1,2,\dots,l)} \in F_{Ts}$ and $|V_{Ts}| = l$.

Feature Selection using Linear Model ANOVA

Before feeding the features to the regressors, we fit a linear model to the training samples and then performed a one-way ANOVA on the linear model. The idea [109] was to see whether the variance in the predicted results was significantly explained by each feature or not. We discarded the features for which $p > 0.10$ (as in [109]). This feature selection method was chosen as it is consistent with the regression method used in our study.

4.4.7 Label Extraction, aggregation of continuous annotations

There are two standard ways of aggregating continuous annotations as also reported in [75]: mean and median of per-second clip annotations. In this study,

for the sake of comparison, we reported the regression results using both. Using the mean of annotations provides a more smooth annotation over time, but when the number of annotations are few, it is prone to noise. Median of the annotations is more robust to noise, but the result of median-based aggregation is usually not as smooth as for the mean. Using median we may be able to capture finer affective events from time-continuous annotations, but using the mean to aggregate the annotations, we are able to describe the emotional profile of movie scenes better.

4.4.8 Regression Algorithm and Inner-loop parameter optimization

The regression-prediction method that we applied relies on a recent implementation of GLMNET toolbox¹. As defined in Section ??, assuming F_{Tr} to be the training feature-set and T_{Tr} to be the target vector correspondent to F_{Tr} , the goal is to fit a statistical model on F_{Tr} to estimate T_{Tr} . Then we use the statistical model to predict regression responses (estimation of T_{Ts}) to test features F_{Ts} . We evaluated the performance of the regressor with the value and significance level of the $R - squared$ measure as mentioned previously.

To build the statistical model we use the Lasso optimization problem, proposed in [119]. The idea is to find regression coefficients by solving a regularized least-squares problem formulated as:

$$\hat{\beta} = \operatorname{argmin}_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

where $\lambda \geq 0$ is a tuning parameter, and the method regularizes β by trading off "goodness of fit" for a reduction in "wildness of coefficients"[101]. We optimized the λ with a method explained in [120] and implemented using the GLMNET toolbox. The method performs an inner cross-validation on the training data.

¹www.stanford.edu/hastie/glmnet_matlab/

4.4.9 Information Fusion: Early vs. Late

On top of the unimodal regression, where we trained and tested the regressor with either audio or video features, we explored two approaches to fuse the information content in the two channels to boost up performance. There are two general strategies to perform fusion of information content, (i) early fusion or feature fusion, where we simply concatenate the features from the two sources and feed in to the regressor, and (ii) late or decision fusion where we fuse the processed unimodal information at the decision level. The first method is not always very successful in enhancing the performance due to several reasons: (a) inhomogeneity of different features arising from different sources, (b) stack of redundancy in the overall feature space and therefore getting an increasingly more difficult pattern recognition problem, and (c) an increase in the number of features with the sample size remaining fixed increases error variance, and may consequently decrease overall performance. On the other hand, decision fusion has been consistently shown to outperform feature fusion in [58, 57, 109, 52]. Thus, in this study we only performed the decision fusion method

Assuming that E_A is the prediction given by regressor for audio features and E_V is the regressor for the video features, we calculated the *overall* output of the regression with E_O defined as:

$$E_O = \alpha \times E_A + (1 - \alpha) \times E_V \text{ where } 0 \leq \alpha \leq 1$$

The estimation of best parameter of α is performed through an inner one-video-out cross-validation loop on the training movie clips with the following criterion: (i) maximizing the mean of R-squared measure for the training videos (obtained over the inner loop cross validation) while the number of significant ($p < 0.05$) correlations between predictions and actual annotations must be more than 80% of the training clip size.

4.4.10 Prediction Post-processing

Hanjalic *et al.* [37] used a Kaiser window of $width = 1500$ with the shape parameter of $\beta = 5$ to post process the predictions. The purpose of applying the filter is to smooth the prediction and estimate the dynamic emotional profile of the movie. Similar to Section 4.4.6, we set the parameters of the Kaiser window to $width = \frac{1500}{30}$ and $\beta = \frac{5}{30}$. The effect of applying the Kaiser window on the predictions can be seen from Fig.4.7(b).

4.4.11 Results

Our reports include the unimodal regressions over audio/video features independently, and decision fusion of the unimodal regressors. We report regression performance for two aggregation methods reported in [75], (i) using mean of annotations and (ii) median of annotations. As in Table 4.5 and Figure 4.8, our results suggest that using the *mean* is more effective for aggregating the annotations. Obtained results suggest that all three methods yield reasonable (\bar{R}^2) measures, but the results obtained with crowd annotations are better. In particular, using affective ground truth for the video is the best way to get high quality time-continuous affect annotations.

The audio features, in our setup, provided more accurate valence predictions. However, the video features worked better than the audio features for the prediction of arousal measures. Late fusion results are generally enhanced. The number of significant predictions are more in early fusion, while in late fusion the mean coefficient of determination (\bar{R}^2) is the highest among the the employed techniques.

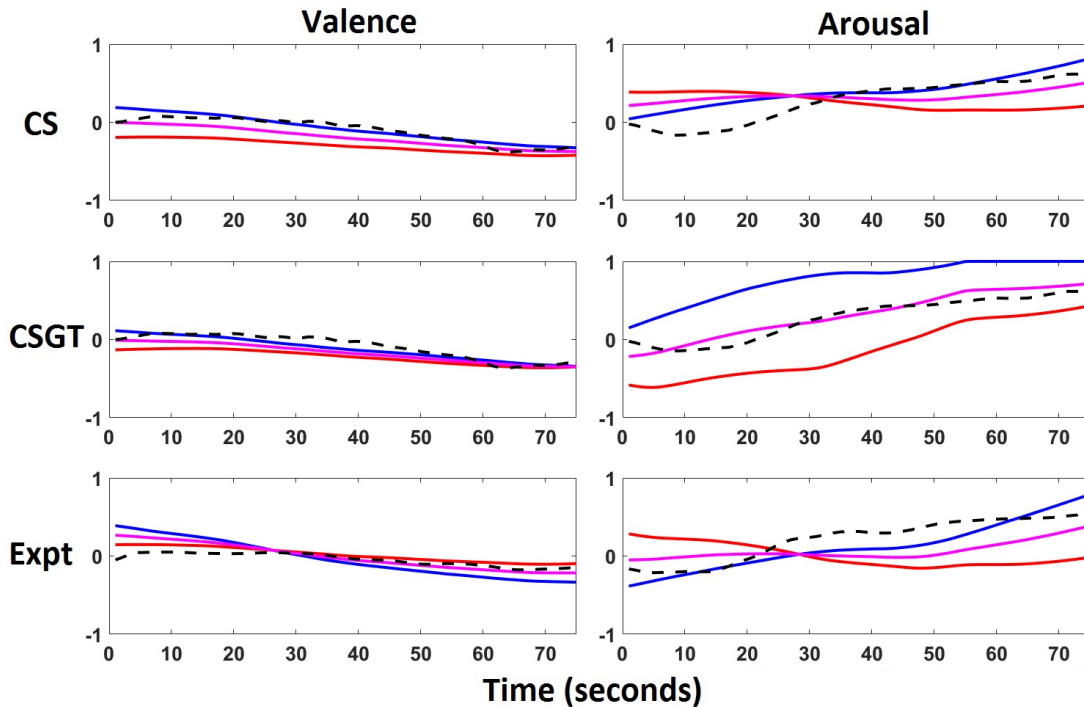


Figure 4.8: Plot of regression results using audio features (blue), video features (red), and fusion (magenta). The representative emotion profile compiled from workers and experts is shown in dashed black.

4.5 Applying a Multi-task Learning Framework

In this section², as a preliminary step towards ensuring good quality VA labels, we discarded those time-continuous annotations with (1) missing values more than threshold, (2) standard deviation less than threshold, and (3) missing overall or general VA ratings.

As mentioned previously, Multi-task learning (MTL) models both similarities as well as differences among a set of related tasks, which is more beneficial as compared to learning task-specific models. Given a set of tasks $t = 1..T$, with $X(t)$ denoting training data for the task t and $Y(t)$ their corresponding labels (ratings), MTL seeks to jointly learn a set of weights $W = [W_1..W_T]$,

²This section has been published in the proceedings of the ACM Workshop on Crowdsourcing for Multimedia, 2014 [45]

Table 4.5: Regression results over one-video-out cross-validation for 12 videos. Mean and $p = 0.99$ confidence interval measures over the R-squared (R^2) measurements are presented

Modality	Annotation Type	Valence		Arousal	
		\bar{R}^2	$CI_{p=0.99}$	\bar{R}^2	$CI_{p=0.99}$
Audio	CS	0.69	[0.41, 0.96]	0.45	[0.19, 0.72]
	CSGT	0.72	[0.47, 0.98]	0.46	[0.13, 0.80]
	Expt	0.60	[0.34, 0.86]	0.39	[0.16, 0.62]
Video	CS	0.57	[0.27, 0.87]	0.46	[0.13, 0.79]
	CSGT	0.66	[0.42, 0.89]	0.52	[0.19, 0.85]
	Expt	0.49	[0.24, 0.75]	0.52	[0.18, 0.86]
Fusion	CS	0.72	[0.50, 0.94]	0.51	[0.28, 0.73]
	CSGT	0.69	[0.41, 0.97]	0.61	[0.36, 0.87]
	Expt	0.58	[0.33, 0.83]	0.45	[0.22, 0.68]

where W_t models task t . For the problem of time-continuous VA prediction, the 12 movie clips used for crowdsourcing denote the related tasks. In this work, we used the publicly available MALSAR library [142], which contains a host of MTL algorithms for analysis. We were particularly interested in the following MTL variants:

Multi-task Lasso: which extends the Lasso algorithm [119] to MTL, and assumes that sparsity is shared among all tasks.

ℓ_{21} norm-regularized MTL [6]: which attempts to minimize the objective function $\sum_{t=1}^T \|W_t^T X_t - Y_t\|_F^2 + \alpha \|W\|_{2,1} + \beta \|W\|_F^2$, where $\|\cdot\|_F$ and $\|\cdot\|_{2,1}$ denote matrix Frobenious norm and ℓ_{21} norm respectively. The basic assumption in this model is that *all* tasks are related, which is not always true, and α, β denote the regularization parameters controlling group sparsity and norm sparsity respectively.

Dirty MTL [40]: where the weight matrix $W = P + Q$, where P and Q denote the group and task-wise sparse components.

Sparse graph Regularization (SR-MTL): where *a priori* knowledge concern-

ing task-relatedness is modeled in terms of a graph R in the objective function. This way, similarity is only enforced between W_t 's corresponding to related tasks. The minimized objective function in this case is $\sum_{t=1}^T \|W_t^T X_t - Y_t\|_F^2 + \alpha \|WR\|_F^2 + \beta \|W\|_1 + \gamma \|W\|_F^2$, where R is the graph encoding task relationships, and α, β, γ denote regularization parameters as above.

4.5.1 Data Analysis and Experiments

Upon compiling VA ratings from crowdworkers, we firstly examined if any patterns existed in the dynamic annotations. This examination was important for two reasons— (1) predicting dynamic VA levels for a stimulus instead of the overall rating is useful as it allows for determining the ‘emotional highlight’ in the scene, and comparing dynamic vs static ratings could help us understand how dynamic emotion perception influenced crowdworkers’ overall impression of a scene, and (2) given the subjectivity associated with emotions and the uncontrolled worker population, patterns in dynamic VA annotations would indicate that a reliable, *gold standard* annotation for a clip is achievable in spite of these biases.

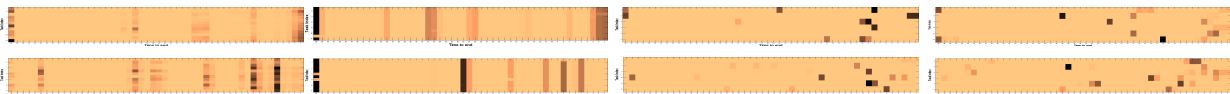


Figure 4.9: (Top) W matrix learnt for valence using (from left to right) $\ell_{2,1}$, dirty and SR MTL (HV, LV). (Bottom) W learnt for arousal using (from left to right) $\ell_{2,1}$, dirty and SR MTL (HA, LA). Larger weights are denoted using darker shades.

Given the 12 clips (tasks) related in terms of valence and arousal (from now on, clips/tasks 1-3, 4-6, 7-9 and 10-12 respectively correspond to HAHV, LAHV, LALV and HALV labels), we employed MTL to determine *if some time-points influenced the overall emotional perception for a movie clip more than others?* To this end, we used the time-continuous VA ratings to *predict* the overall VA rating for each clip. For dimensional consistency, we only used the VA

ratings for the final 50 sec of each clip for this experiment, and so, the x -axis in Fig. 4.9 denotes time to clip completion (between 50-1 sec). Weights learnt using the different MTL variants consistently suggest that the continuous VA ratings provided in the *latter* half for *all* of the movie scenes predict the overall rating better. The third and fourth columns respectively depict learnt weights for the six HV, LV/HA, LA clips, and represents the situation where *a-priori* knowledge regarding task-relatedness is fed to SR-MTL. Examining SR-MTL valence weights (cols 3,4 in row 1), one can infer that general affective impressions are created earlier in time for high-valence stimuli as compared to low-valence stimuli. Examination of SR-MTL arousal weights suggests that the most influential impressions regarding high-arousal stimuli are also created a few seconds before clip completion.

Therefore, MTL enables effective characterization of patterns concerning dynamic VA levels of crowdworkers, and this in turn implies that deriving a representative, gold standard annotation from worker annotations for each movie clip is meaningful. While MTL has been used to learn from noisy crowd data [42], we simply used the median value of the annotations at each time-point to derive the ground truth emotional profile for each movie clip. Next, we will briefly describe the audio-visual features extracted from each clip, and show how the joint learning of the relationship between audio-visual features and VA ratings allows for more effective dynamic emotion prediction.

4.5.2 Experiments and Results

In this section, we attempt to *predict* the gold standard (or ground-truth) dynamic V/A ratings for each clip from audio-visual features using MTL, and show why learning the audio visual feature-emotion relationship simultaneously for the 12 movie scenes is more effective than learning scene-specific models. Fig. 4.10 shows the model weights learnt by the various MTL approaches when they are trained with features and VA ratings over the entire clip duration for

4.5. APPLYING A MULTI-TASK LEARNING FRAMEWORK

all clips. Here again, some interesting correlates between audio-visual features and VA ratings are observed over all scenes. Considering video features, color descriptors are found to be salient for valence, while motion and visual excitement correlate with arousal better, especially for HA stimuli, as noted from SR MTL (HA) weights (column 7). Among audio features, the first few MFCC components correlate well with both V,A.

Table 4.6: RMSE-based V/A prediction performance of task-specific vs multi-task methods. RMSE mean, standard deviation over five runs are reported. Best model RMSE is shown in bold.

		Front			Back			
		5 s	10 s	15 s	5 s	10 s	15 s	
Valence	Video	Lasso	0.429±0.041	0.816±0.583	1.189±0.625	0.584±0.024	0.881±0.057	1.125±0.064
		MT-Lasso	0.191±0.028	0.319±0.064	0.549±0.042	0.206±0.014	0.443±0.067	0.593±0.108
		ℓ_{21} MTL	0.193±0.030	0.326±0.063	0.565±0.047	0.207±0.015	0.450±0.066	0.606±0.113
		Dirty MTL	0.452±0.141	0.840±0.293	1.179±0.400	0.308±0.105	0.607±0.140	0.801±0.129
		SR MTL	0.193±0.030	0.325±0.064	0.563±0.046	0.207±0.015	0.450±0.066	0.607±0.113
	Audio	Lasso	0.475±0.030	0.712±0.069	0.851±0.081	0.634±0.016	0.860±0.027	1.174±0.034
		MT-Lasso	0.241±0.023	0.348±0.014	0.487±0.038	0.237±0.024	0.400±0.039	0.527±0.033
		ℓ_{21} MTL	0.243±0.020	0.359±0.012	0.520±0.029	0.247±0.026	0.392±0.029	0.553±0.028
		Dirty MTL	0.299±0.023	0.473±0.019	0.751±0.060	0.312±0.043	0.524±0.042	0.692±0.072
		SR MTL	0.248±0.017	0.365±0.015	0.526±0.027	0.252±0.027	0.404±0.033	0.567±0.026
Arousal	Video	Lasso	0.429±0.041	0.816±0.583	1.189±0.625	0.584±0.024	0.881±0.057	1.125±0.064
		MT-Lasso	0.191±0.028	0.319±0.064	0.549±0.042	0.206±0.014	0.443±0.067	0.593±0.108
		ℓ_{21} MTL	0.193±0.030	0.326±0.063	0.565±0.047	0.207±0.015	0.450±0.066	0.606±0.113
		Dirty MTL	0.452±0.141	0.840±0.293	1.179±0.400	0.308±0.105	0.607±0.140	0.801±0.129
		SR MTL	0.193±0.030	0.325±0.064	0.563±0.046	0.207±0.015	0.450±0.066	0.607±0.113
	Audio	Lasso	0.435±0.038	0.599±0.050	0.727±0.058	0.556±0.033	0.807±0.037	1.004±0.033
		MT-Lasso	0.212±0.026	0.339±0.020	0.406±0.019	0.243±0.039	0.353±0.028	0.464±0.029
		ℓ_{21} MTL	0.212±0.028	0.345±0.022	0.437±0.027	0.238±0.032	0.358±0.022	0.475±0.027
		Dirty MTL	0.249±0.015	0.420±0.036	0.618±0.051	0.304±0.043	0.430±0.029	0.568±0.019
		SR MTL	0.214±0.027	0.350±0.031	0.431±0.026	0.242±0.031	0.363±0.026	0.487±0.030

Then, we examined if learning prediction models for all movie clips was more beneficial than training a Lasso regressor per movie clip. To this end, we held out time-contiguous data of length 5, 10 or 15 seconds from the first half (front) or second half (back) of each of the clips for testing, while the remainder of the clips were used for training. Optimal group sparsity regularization parameter for the different MTL methods, as well as optimal Lasso parameter were chosen from [0.01 0.1 1 5] employing 5-fold cross validation, and all

other parameters (where necessary) were set to 1. The root mean square error (RMSE) observed for V/A estimates over all clips (tasks) is shown in Table 4.6. MTL methods clearly outperform single-task Lasso, and consequent to our earlier finding that the latter half of all clips is emotionally salient, larger prediction errors are observed for the back portion. Also, prediction errors increase with the test clip size, and predictions are more accurate for arousal, and with audio features. Finally, sophisticated MTL methods such as dirty and SR-MTL outperform MT-Lasso and $\ell_{2,1}$ MTL. Overall, these results are demonstrative of efficient MTL-based learning utilizing relatively few training examples.

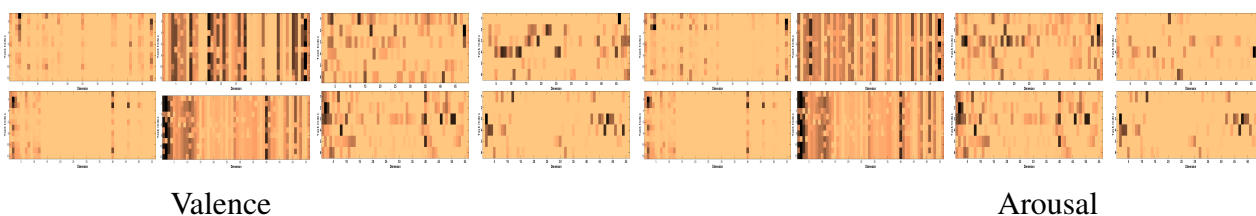


Figure 4.10: Predicting dynamic V/A ratings using (top) video and (bottom) audio features. Order of illustrated W matrices for valence (cols 1-4) and arousal (cols 5-8) is identical to Fig. 4.9. Larger weights shown using darker shades (best viewed under zoom).

4.6 Conclusion and future work

The presented study confirms that crowd annotations are useful for continuous affect analysis, and furthermore, are comparable in quality to expert annotations or even better upon careful filtering. One of the drawbacks of the regression method [37] we used is that, the method to filter out high frequency information at the feature level, and at the prediction level is only able to capture the emotion trends in the movie clips. Although the predictions are good to capture the affective trend of the videos, our method is unable to capture finer affective multimedia events as aggregation using the median criterion is worse than the mean criterion. Another limitation is a strong assumption that the samples which are fed in the regression algorithm are independent, while this is obviously not true

as consecutive samples are obviously correlated in terms of both features and annotations. We intend to employ unsupervised approaches for removing noisy annotations as anomalies[139]. We also intend to explore hidden markov models (HMM) and conditional random fields (CRF) to model this phenomenon in our future investigations.

The crowd facial responses were not analyzed in this study, and they could be valuable based on the investigations of McDuff *et al.* [73]. We also intend to make the crowd annotations publicly available in the near future.

The study also explores Multi-task learning to estimate dynamic VA levels for movie scenes. Since time-continuous VA annotations are highly difficult to acquire, we employ crowdsourcing for the same. Though emotion is a subjective feeling and the crowdworkers arose from varied demographics, MTL could effectively capture patterns concerning their dynamic emotion perception .The latter half of all clips was found to be more emotionally salient, and influenced the affective impression of the clip. We again utilized MTL to model the relationship between the representative dynamic VA profile for each clip and underlying audio-visual effects, and observed that MTL approaches considerably outperformed clip-specific Lasso models, implying that jointly learning characteristics of a collection of scenes is beneficial. Future work involves usage of (1) MTL for cleaning crowd annotations, and (2) face videos compiled in this work as an additional affective cue.

Chapter 5

Conclusion and Future Work

In this thesis, we have addressed the problem of automatic recognition of users' psychological parameters via the analysis of their spontaneous responses to affective multimedia content. We specifically tackled users' emotion and personality recognition that could be applied in (i) user profiling for multimedia recommender systems, as well as (ii) multimedia tagging.¹

In chapter 2, we presented DECAF, a multimodal dataset for **d**ecoding user physiological responses to **a**ffective multimedia content. Different from datasets such as DEAP [57] and MAHNOB-HCI [109], DECAF contains (1) brain signals acquired using the Magnetoencephalogram (MEG) sensor, which requires little physical contact with the user's scalp and consequently facilitates naturalistic affective response, and (2) explicit and implicit emotional responses of 30 participants to 40 one-minute music video segments used in [57] and 36 movie clips, thereby enabling comparisons between the EEG vs MEG modalities as well as movie vs music stimuli for affect recognition. In addition to MEG data, DECAF comprises synchronously recorded near-infra-red (NIR) facial videos, horizontal Electrooculogram (hEOG), Electrocardiogram (ECG), and trapezius-Electromyogram (tEMG) peripheral physiological responses. To demonstrate DECAF's utility, we presented (i) a detailed analysis of the

¹Our contributions to the following publications and patents are not explicitly included in this thesis:[76, 95, 30, 47, 51, 46, 44, 48]

correlations between participants' self-assessments and their physiological responses and (ii) single-trial classification results for *valence*, *arousal* and *dominance*, with performance evaluation against existing datasets. DECAF also contains *time-continuous* emotion annotations for movie clips from seven users, which we used to demonstrate dynamic emotion prediction.

In chapter 3, we presented a multimodal database for implicit Personality and Affect recognition using commercial physiological sensors. To our knowledge, ASCERTAIN is the first database to connect *personality traits* and *emotional states* via *physiological responses*. ASCERTAIN contains big-five personality scales and emotional self-ratings of 58 users along with their Electroencephalogram (EEG), Electrocardiogram (ECG), Galvanic Skin Response (GSR) and facial activity data, recorded using off-the-shelf market available sensors while viewing affective movie clips. We first examine relationships between users' affective ratings and personality scales in the context of prior observations, and then study linear and non-linear physiological correlates of emotion and personality. Our analysis suggests that the emotion–personality relationship is better captured by non-linear rather than linear statistics. We finally attempt binary emotion and personality trait recognition using physiological features. Experimental results cumulatively confirm that personality differences are better revealed while comparing user responses to emotionally homogeneous videos, and above-chance recognition is achieved for both affective and personality dimensions.

In the second section of chapter 3, we proposed and validated a system and method that take into the consideration the quality of psycho-physiological signals within a unified multi-modal emotion recognition inference system. We particularly validated the approach on a user-centric implicit affective indexing employing emotion detection based on psycho-physiological signals, such as electrocardiography (ECG), galvanic skin response (GSR), electroencephalography (EEG) and face tracking. We employed off-the-shelves market-available

commercial sensors in a natural environment. However, real world psycho-physiological signals obtained from wearable devices and facial trackers are contaminated by various noise sources that can result in spurious emotion detection, so that high signal quality is not guaranteed. Our proposed methods include the development of psycho-physiological signal quality estimators for unimodal affect recognition systems. The presented systems perform adequately in classifying users affect however, they resulted in high failure rates due to rejection of bad quality samples. Thus, to reduce the affect recognition failure rate, a quality adaptive multimodal fusion scheme is proposed. The proposed scheme yields no failure, while at the same time classify the users' arousal/valence and liking with significantly above chance weighted F1-scores in a cross-user experiment. Another finding of this study is that head movements encode liking perception of users in response to music snippets. The study also includes the release of the employed dataset including psycho-physiological signals, their quality annotations, and users' affective self-assessments.

In chapter 4, we presented the methods for the development crowdsourcing (CS) platform for capturing time-continuous emotional (valence and arousal) annotation of movie clips in order to gather adequate training data for dynamic affect analysis. Upon systematically cleaning the crowd annotations using a filtering procedure, we compared them with ratings obtained from seven experts obtained in lab conditions. Aggregating the multiple (crowd and expert) annotations using the mean/median score at each time instant, we obtained one representative dynamic emotion profile for each clip. A Generalized Linear Model (GLM) for dynamic valence/arousal estimation employing leave-one-out cross validation *predicted better with the crowd annotations* as compared to expert annotations. Overall, our results demonstrate that carefully conditioned crowd annotations are comparable (or better) in quality to those obtained from a group of experts.

In the same study, we also proposed Multi-task learning (MTL) for time-

continuous or dynamic emotion (valence and arousal) estimation in movie scenes. Since compiling annotated training data for dynamic emotion prediction is tedious, we employed crowdsourcing for the same. Even though the crowdworkers come from various demographics, we demonstrate that MTL can effectively discover (1) consistent patterns in their dynamic emotion perception, and (2) the low-level audio and video features that contribute to their valence, arousal (VA) elicitation. Finally, we show that MTL-based regression models, which simultaneously learn the relationship between low-level audio-visual features and high-level VA ratings from a collection of movie scenes, can predict VA ratings for time-contiguous snippets from each scene more effectively than scene-specific models.

Over the past decade, new technologies became a constant and important part of human life. The typical life and work style has changed, resulting in people spending more time with their “*smart*” environments including smart phones, smart wearables, smart cars and smart homes. Smart cities are being born one after each other around the globe and Internet of Things is emerging rapidly; sensors are getting everywhere in our environment, facilitating seamless users monitoring.

With the fast growth of polymer-technologies and nano-technologies, sensing technologies are getting much more accurate. Therefore, getting access to high quality users’ psycho-physiological signals gets easier over time.

Emotions are among the most complex factors of human beings, from perception an emotion to presentation of an affect. How humans perceive emotions from the stimuli in their environment could be under the impact of users’ subjective parameters including their memories and background, personality, temper, mood, age, gender, and culture. Such subjective parameter maybe inter correlated with each other and they not only have impact on how users perceive emotions within an interaction with a certain environment but also they have influence on how users respond to certain stimuli[16, 51].

A set of responses of a user could be interpreted in different ways given different contexts[16, 51]; and hence the context of an interaction have impact on determining a user's emotion.

Being given the context and/or the subjective parameters, accuracy of an emotion recognition system could be significantly improved[16, 51], and we suggest this direction as the continuation of our research.

Some of the applications of knowing users's emotion are listed below:

- Elderly monitoring, as elderly humans need care and attention. Elderly monitoring is one of the most important applications in the context of well-being.
- Monitoring patients who might suffer from bi-polarity, stress, depression is another interesting application in the context of wellbeing.
- Work environment quality assurance, where monitoring employees' affective states helps maintaining employees engaged better with their work that results in building efficient and healthy work environments; in the context of cooperate wellness.
- Monitoring users in the context of neuro-marketing, product testing and of-course game-testing.
- Tracking the affective state of trainees in the context of athlete training, education, soldier training (army), and aviation.
- Monitoring people with sensitive roles; first responders, police officers, security officers, soldiers, fire fighters, construction workers, and truck drivers.
- Drowsiness detection, drunkenness detection, engagement analysis syatems that mainly focus on end users of various markets.

-
- Building affective profile of users in various end-user applications including multimedia recommender systems.
 - There are many more applications that relate to understanding the cognitive load and affective state of a user.

Our research is among the first steps towards bringing lab-environment affective computing technologies to real world environments where coarse emotion tags could update to a transformative technology of 24/7 time-continuous emotion monitoring².

We believe many patterns on time-series of psycho-physiological that relate to affect and cognitive states are complex non-linear patterns; and hence we encourage researchers to develop large scale psycho-physiological corpora to make artificial intelligence systems able to capture such patterns. Such large scale psycho-physiological corpora could lead in building systems that (i) implicitly infer subjective and contextual parameters from users' spontaneous reactions to stimuli within their environment before (ii) such systems use the subjective and contextual information as a prior to accurately recognize one's emotions in the *wild*.

² Sensaura Inc. has shown that it is possible to perform 24/7 recognition of emotions via pattern analysis over the user's heart activity signal. Their innovative deep technologies has been validated by their industrial partners.

Bibliography

- [1] Emotion and multimedia content. In Borko Furht, editor, *Encyclopedia of Multimedia*. Springer, 2006.
- [2] Xavier Alameda-Pineda, Jacopo Staiano, Ramanathan Subramanian, Ligia Batrinca, Elisa Ricci, Bruno Lepri, Oswald Lanz, and Nicu Sebe. SALSA: A novel dataset for multimodal group behavior analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2016.
- [3] Anton Aluja, Oscar Garca, and Lus F. Garca. Relationships among Extraversion, Openness to experience, and sensation seeking. *Personality and Individual Differences*, 35(3):671–680, 2003.
- [4] Monika Ardelt. Still stable after all these years? personality stability theory revisited. *Social Psychology Quarterly*, 63(4):pp. 392–405, 2000.
- [5] S. Argamon, S. Dhawle, M. Koppel, and Pennbaker. Lexical predictors of personality type. In *Interface and the Classification Society of North America*, 2005.
- [6] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Multi-task feature learning. *Advances in neural information processing systems*, 19:41, 2007.
- [7] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. *Current Directions in Psychological Science*, 20(5):286–290, 2011.

- [8] Ellen Elizabeth Bartolini. Eliciting emotion with film: Development of a stimulus set. Master's thesis, Wesleyan University, 2001.
- [9] Anton Batliner, Kerstin Fischer, Richard Huber, Jörg Spilker, and Elmar Nöth. How to find trouble in communication. *Speech communication*, 40(1):117–143, 2003.
- [10] Anton Batliner, Christian Hacker, Stefan Steidl, Elmar Nöth, Shona D'Arcy, Martin J Russell, and Michael Wong. You stupid tin box - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In *Language Resources and Evaluation*, 2004.
- [11] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)*, 57(1):289–300, 1995.
- [12] Robert F Bornstein. Exposure and affect: Overview and meta-analysis of research, 1968–1987. *Psychological bulletin*, 106(2):265, 1989.
- [13] M. Bradley. Emotional memory: A dimensional analysis. In *Emotions: Essays on Emotion Theory*. Lawrence Erlbaum, 1994.
- [14] Anne-Marie Brouwer, Martin G Van Schaik, J E Hans Korteling, Jan B F Van Erp, and Alexander Toet. Neuroticism, Extraversion, Conscientiousness and Stress: Physiological Correlates. *IEEE Trans. Affective Computing*, 6(2):109–117, 2015.
- [15] C. C. Brumbaugh, R. Kothuri, C. Marci, C. Siefert, and D. D. Pfaff. Physiological correlates of the Big 5: Autonomic responses to video presentations. *Applied Psychophysiology and Biofeedback*, 38(4):293–301, 2013.

BIBLIOGRAPHY

- [16] Rafael A Calvo, Sidney D’Mello, Jonathan Gratch, and Arvid Kappas. *The Oxford handbook of affective computing*. Oxford University Press, USA, 2014.
- [17] George Caridakis, Lori Malatesta, Loic Kessous, Noam Amir, Amaryllis Raouzaïou, and Kostas Karpouzis. Modeling naturalistic affective states via facial and vocal expressions recognition. In *Int’l Conference on Multimodal Interaction*, pages 146–154, 2006.
- [18] Rich Caruana. Multitask learning. In *Learning to learn*, pages 95–133. Springer, 1998.
- [19] Lei Chen, Sule Gunduz, and M Tamer Ozsü. Mixed type audio classification with support vector machine. In *IEEE International Conference on Multimedia and Expo*, pages 781–784, 2006.
- [20] Paul T Costa and Robert R McCrae. Influence of Extraversion and Neuroticism on Subjective Well-Being: Happy and Unhappy People. *Journal of Personality and Social Psychology*, 38(4):668, 1980.
- [21] Paul T. Jr. Costa and Robert R. McCrae. *NEO-PI-R professional manual: Revised NEO personality and NEO Five-Factor Inventory (NEO-FFI)*, volume 4. Psychological Assessment Resources, Odessa, Florida, 1992.
- [22] Roddy Cowie, Gary McKeown, and Ellen Douglas-Cowie. Tracing emotion: an overview. *Int’l Journal of Synthetic Emotions*, 3(1):1–17, 2012.
- [23] Vilfredo De Pascalis, Enrica Strippoli, Patrizia Riccardi, and Fabiola Vergari. Personality, event-related potential (ERP) and heart rate (HR) in emotional word processing. *Personality and Individual Differences*, 36:873–891, 2004.

- [24] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [25] R. J. Dolan. Emotion, cognition, and behavior. *Science*, 298(5596):1191–1194, 2002.
- [26] Ellen Douglas-Cowie, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean-Claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis. The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. In *Affective Computing and Intelligent Interaction*, pages 488–500. 2007.
- [27] Michael Esterman, Benjamin J Tamber-Rosenau, Yu-Chin Chiu, and Steven Yantis. Avoiding non-independence in fmri data analysis: leave one subject out. *Neuroimage*, 50(2):572–576, 2010.
- [28] H. J. Eysenck. *Dimensions of Personality*. The Int’l library of psychology. Transaction Publishers, 1947.
- [29] N Fragopanagos and John G Taylor. Emotion recognition in human–computer interaction. *Neural Networks*, 18(4):389–405, 2005.
- [30] Pouya Ghaemmaghami, Mojtaba Khomami Abadi, Seyed Mostafa Kia, Paolo Avesani, and Nicu Sebe. Movie genre classification by exploiting meg brain signals. In *International Conference on Image Analysis and Processing*, pages 683–693. Springer, 2015.
- [31] M. K. Greenwald, E. W. Cook, and P. J. Lang. Affective judgement and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli. *Journal of Psychophysiology*, 3:51–64, 1989.

BIBLIOGRAPHY

- [32] James J Gross and Robert W Levenson. Emotion elicitation using films. *Cognition & Emotion*, 9(1):87–108, 1995.
- [33] Rishabh Gupta, Mojtaba Khomami Abadi, Jesús Alejandro Cárdenes Cabré, Fabio Morreale, Tiago H Falk, and Nicu Sebe. A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 317–320. ACM, 2016.
- [34] Rishabh Gupta, Khalil ur Rehman Laghari, and Tiago H. Falk. Relevance vector classifier decision fusion and eeg graph-theoretic features for automatic affective state characterization. *Neurocomputing*, 174(PB):875–884, January 2016.
- [35] Sunjai Gupta and John Nicholson. Simple visual reaction time , personality strength of the nervous system : theory approach. *Personality and Individual Differences*, 6(4):461–469, 1985.
- [36] Karla R Hamlen and Thomas J Shuell. The effects of familiarity and audiovisual stimuli on preference for classical music. *Bulletin of the Council for Research in Music Education*, pages 21–34, 2006.
- [37] A. Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.
- [38] Uri Hasson, Rafael Malach, and David J. Heeger. Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, 14(1):40 – 48, 2010.
- [39] Rong Hu and Pearl Pu. A Comparative User Study on Rating vs. Personality Quiz based Preference Elicitation Methods. In *Intelligent User Interfaces*, pages 367–371, 2009.

- [40] Ali Jalali, Sujay Sanghavi, Chao Ruan, and Pradeep K Ravikumar. A dirty model for multi-task learning. In *Advances in Neural Information Processing Systems*, pages 964–972, 2010.
- [41] Hideo Joho, Jacopo Staiano, Nicu Sebe, and Joemon 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.
- [42] Hiroshi Kajino, Yuta Tsuboi, and Hisashi Kashima. A convex formulation for learning from crowds. *Transactions of the Japanese Society for Artificial Intelligence*, 27(3):133–142, 2012.
- [43] Elizabeth G Kehoe, John M Toomey, Joshua H Balsters, and Arun L W Bokde. Personality modulates the effects of emotional arousal and valence on brain activation. *Social Cognitive & Affective Neuroscience*, 7:858–70, 2012.
- [44] Mojtaba Khomami Abadi. A quality adaptive multimodal affect recognition system for user-centric multimedia indexing. US Provisional Patent.
- [45] Mojtaba Khomami Abadi, Azad Abad, Ramanathan Subramanian, Negar Rostamzadeh, Elisa Ricci, Jagannadan Varadarajan, and Nicu Sebe. A multi-task learning framework for time-continuous emotion estimation from crowd annotations. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 17–23. ACM, 2014.
- [46] Mojtaba Khomami Abadi and Fahd Bencheekroun. System and method for controlling data transmissions using human state-based data. US Patent 20,170,041,264.

BIBLIOGRAPHY

- [47] Mojtaba Khomami Abadi, Juan Abdón Miranda Correa, Julia Wache, Heng Yang, Ioannis Patras, and Nicu Sebe. Inference of personality traits and affect schedule by analysis of spontaneous reactions to affective videos. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1–8. IEEE, 2015.
- [48] Mojtaba Khomami Abadi, Jesus A. Crdenes Cabr, and Michel Allegue. System and method for multimodal signal quality assessment, noise removal, and signal recovery on structural signals. US Provisional Patent.
- [49] Mojtaba Khomami Abadi, Seyed Mostafa Kia, Ramanathan Subramanian, Paolo Avesani, and Nicu Sebe. Decoding affect in videos employing the meg brain signal. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.
- [50] Mojtaba Khomami Abadi, Seyed Mostafa Kia, Ramanathan Subramanian, Paolo Avesani, and Nicu Sebe. User-centric Affective Video Tagging from MEG and Peripheral Physiological Responses. In *Affective Computing and Intelligent Interaction*, pages 582–587, 2013.
- [51] Mojtaba Khomami Abadi and Jean-philip Rancourt Poulin. System and method for multimodal human state recognition. US Patent 20,160,358,085.
- [52] Mojtaba Khomami Abadi, Jacopo Staiano, Alessandro Cappelletti, Massimo Zancanaro, and Nicu Sebe. Multimodal engagement classification for affective cinema. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*, pages 411–416. IEEE, 2013.

- [53] Mojtaba Khomami Abadi, Ramanathan Subramanian, Seyed Mostafa Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe. Decaf: Meg-based multimodal database for decoding affective physiological responses. *IEEE Transactions on Affective Computing*, 6(3):209–222, 2015.
- [54] Seyed Mostafa Kia, Emanuele Olivetti, and Paolo Avesani. Discrete cosine transform for MEG signal decoding. In *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 132–135, 2013.
- [55] J. Kim and E. Andr. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, Dec 2008.
- [56] Jonghwa Kim and E. Andre. Emotion recognition based on physiological changes in music listening. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008.
- [57] Sander Koelstra, Christian Mühl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.
- [58] Sander Koelstra and Ioannis Patras. Fusion of facial expressions and eeg for implicit affective tagging. *Image and Vision Computing*, 31(2):164–174, 2013.
- [59] Gerbert Kraaykamp and Koen van Eijck. Personality, media preferences, and cultural participation. *Personality and Individual Differences*, 38(7):1675–1688, 2005.

BIBLIOGRAPHY

- [60] P.J. Lang, M.M. Bradley, and B.N. Cuthbert. IAPS: Affective ratings of pictures and instruction manual. Technical report, University of Florida, 2008.
- [61] Nicole A Lazar, Beatriz Luna, John A Sweeney, and William F Eddy. Combining brains: a survey of methods for statistical pooling of information. *Neuroimage*, 16(2):538–550, 2002.
- [62] Chul Min Lee and Shrikanth S Narayanan. Toward detecting emotions in spoken dialogs. *Speech and Audio Processing*, 13(2):293–303, 2005.
- [63] Bruno Lepri, Ramanathan Subramanian, Kyriaki Kalimeri, Jacopo Staiano, Fabio Pianesi, and Nicu Sebe. Connecting meeting behavior with Extraversion - A systematic study. *IEEE Trans. Affective Computing*, 3(4):443–455, 2012.
- [64] Alexander Lerch. *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons, 2012.
- [65] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.
- [66] Dongge Li, Ishwar K Sethi, Nevenka Dimitrova, and Tom McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition Letters*, 22(5):533–544, 2001.
- [67] Tanja Lischetzke and Michael Eid. Why extraverts are happier than introverts: The role of mood regulation. *Journal of personality*, 74(4):1127–1162, 2006.

- [68] Christine Lætitia Lisetti and Fatma Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Advances in Signal Processing*, 2004(11):1672–1687, 2004.
- [69] Bruce D Lucas, Takeo Kanade, et al. An iterative image registration technique with an application to stereo vision. In *Int’l Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
- [70] Patrick Lucey, Jeffrey F Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended Cohn-Kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression. In *CVPR Workshops*, pages 94–101, 2010.
- [71] F. Mairesse, M. A. Walker, M. R. Mehl, and Moore R. K. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.
- [72] R.R. McCrae and O.P. John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, June 1992.
- [73] Daniel McDuff, RE Kaliouby, and Rosalind W Picard. Crowdsourcing facial responses to online videos. *IEEE Transactions on Affective Computing*, 3(4):456–468, 2012.
- [74] Ivan Mervielde, Filip De Fruyt, and Slawomir Jarmuz. Linking openness and intellect in childhood and adulthood. *Parental descriptions of child personality: Developmental antecedents of the Big Five*, pages 105–126, 1998.
- [75] Angeliki Metallinou and Shrikanth Narayanan. Annotation and processing of continuous emotional attributes: Challenges and opportunities. In *Automatic Face and Gesture Recognition*, pages 1–8, 2013.

BIBLIOGRAPHY

- [76] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. Amigos: A dataset for mood, personality and affect research on individuals and groups. *arXiv preprint arXiv:1702.02510*, 2017.
- [77] P.P. Mitra and B. Pesaran. Analysis of dynamic brain imaging data. *Biophysical Journal*, 76(2):691 – 708, 1999.
- [78] Fabio Morreale, Raul Masu, and Antonella De Angeli. Robin: an algorithmic composer for interactive scenarios. *Proceedings of 10th Sound and Music Comp.*, 2013.
- [79] Emily Mower, Maja J Mataric, and Shrikanth Narayanan. Human perception of audio-visual synthetic character emotion expression in the presence of ambiguous and conflicting information. *IEEE Transactions on Multimedia*, 11(5):843–855, 2009.
- [80] Kamran Mustafa and Ian C Bruce. Robust formant tracking for continuous speech with speaker variability. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(2):435–444, 2006.
- [81] Weiting Ng. Clarifying the relation between Neuroticism and positive emotions. *Personality and Individual Differences*, 47(1):69–72, 2009.
- [82] D. Olguin, B. Waber, T. Kim, A. Mohan, K. Ara, and A. Pentland. Sensible organizations: Technology and methodology for automatically measuring organizational behavior. *IEEE Trans. Systems, Man, and Cybernetics-Part B: Cybernetics*, 39(1):43–55, 2009.
- [83] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan-Mathijs Schoffelen. Fieldtrip: Open source software for advanced analysis of meg, eeg, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 2011.

- [84] P-Y. Oudeyer. The production and recognition of emotions in speech: features and algorithms. *International Journal of Human Computer Interaction*, 59(1-2):157–183, 2003.
- [85] M. Pantic and A. Vinciarelli. Implicit human-centered tagging [social sciences]. *IEEE Signal Processing Magazine*, 26(6):173–180, November 2009.
- [86] Maja Pantic, Michel Valstar, Ron Rademaker, and Ludo Maat. Web-based database for facial expression analysis. In *Int'l Conference on Multimedia and Expo*, 2005.
- [87] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, November 2011.
- [88] T. Penzel, B. Kemp, G. Klosch, A. Schlogl, J. Hasan, A. Varri, and I. Korhonen. Acquisition of biomedical signals databases. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):25–32, May 2001.
- [89] M. Perugini and L. Di Blas. Analyzing personality-related adjectives from an eticemic perspective: the big five marker scale (BFMS) and the Italian AB5C taxonomy. *Big Five Assessment*, pages 281–304, 2002.
- [90] Peter Peyk, Harald T Schupp, Thomas Elbert, and Markus Junghöfer. Emotion processing in the visual brain: a meg analysis. *Brain Topography*, 20(4):205–215, 2008.
- [91] Rosalind W Picard. *Affective computing*. MIT press, 2000.

- [92] Ariadna Quattoni, Xavier Carreras, Michael Collins, and Trevor Darrell. An efficient projection for l_1 regularization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 857–864. ACM, 2009.
- [93] Hamed R.-Tavakoli, Adham Atyabi, Antti Rantanen, Seppo J. Laukka, Samia Nefti-Meziani, and Janne Heikkil. Predicting the valence of a scene from observers eye movements. *PLoS ONE*, 10(9):1–19, 2015.
- [94] Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the crowdworkers?: shifting demographics in mechanical turk. In *Human Factors in Computing Systems*, pages 2863–2872, 2010.
- [95] Negar Rostamzadeh, Jasper Uijlings, Ionuj Mironică, Mojtaba Khomami Abadi, Bogdan Ionescu, and Nicu Sebe. Cluster encoding for modelling temporal variation in video. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 3640–3644. IEEE, 2015.
- [96] J. Schwarzbach. A simple framework (asf) for behavioral and neuroimaging experiments based on the psychophysics toolbox for matlab. *Behavior Research Methods*, pages 1–8, 2011.
- [97] Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S Huang. Emotion recognition based on joint visual and audio cues. In *Int’l Conference on Pattern Recognition*, volume 1, pages 1136–1139, 2006.
- [98] Man-Kwan Shan, Fang-Fei Kuo, Meng-Fen Chiang, and Suh-Yin Lee. Emotion-based music recommendation by affinity discovery from film music. *Expert Systems with Applications*, 2009.
- [99] Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy la-

- belers. In *ACM International Conference on Knowledge discovery and Data Mining*, pages 614–622, 2008.
- [100] Jae Woong Shim and Bryant Paul. Effects of personality types on the use of television genre. *Journal of Broadcasting & Electronic Media*, 51(2):287–304, 2007.
- [101] Noah Simon, Jerome Friedman, and Trevor Hastie. A blockwise descent algorithm for group-penalized multiresponse and multinomial regression. *arXiv preprint arXiv:1311.6529*, 2013.
- [102] Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for coxs proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011.
- [103] R. Sinha and O. A. Parsons. Multivariate response patterning of fear and anger. *Cognition and Emotion*, 10(2):173–198, 1996.
- [104] Denise M. Sloan. Emotion regulation in action: emotional reactivity in experiential avoidance. *Behaviour Research and Therapy*, 42(11):1257–1270, 2004.
- [105] M. Soleymani and M. Larson. Crowdsourcing for affective annotation of video: development of a viewer-reported boredom corpus. In *Workshop on Crowdsourcing for Search Evaluation*, 2010.
- [106] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *ACM international workshop on Crowdsourcing for multimedia*, pages 1–6, 2013.
- [107] Mohammad Soleymani, Guillaume Chanel, Joep JM Kierkels, and Thierry Pun. Affective characterization of movie scenes based on multi-

BIBLIOGRAPHY

- media content analysis and user's physiological emotional responses. In *IEEE International Symposium on Multimedia*, pages 228–235, 2008.
- [108] Mohammad Soleymani, Jeremy Davis, and Thierry Pun. A collaborative personalized affective video retrieval system. In *Affective Computing and Intelligent Interaction*, pages 1–2, 2009.
- [109] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affective Computing*, 3:42–55, 2012.
- [110] Ruchir Srivastava, Jiashi Feng, Sujoy Roy, Shuicheng Yan, and Terence Sim. Don't ask me what i'm like, just watch and listen. In *ACM Int'l Conference on Multimedia*, pages 329–338, 2012.
- [111] Jacopo Staiano, Bruno Lepri, Nadav Aharony, Fabio Pianesi, Nicu Sebe, and Alex Pentland. Friends don't lie: Inferring personality traits from social network structure. In *ACM Conference on Ubiquitous Computing*, pages 321–330, 2012.
- [112] Thomas Steiner, Ruben Verborgh, Rik Van de Walle, Michael Hausenblas, and Joaquim Gabarró Vallés. Crowdsourcing event detection in youtube video. In Marieke Van Erp, Willem Robert Van Hage, Laura Hollink, Anthony Jameson, and Raphaël Troncy, editors, *Workshop on detection, representation, and exploitation of events in the semantic web*, pages 58–67, 2011.
- [113] Georg Stenberg. Personality and the EEG: Arousal and emotional arousability. *Personality and Individual Differences*, 13:1097–1113, 1992.
- [114] C. Studholme, D.L.G. Hill, and D.J. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71 – 86, 1999.

- [115] Ramanathan Subramanian, Divya Shankar, Nicu Sebe, and David Melcher. Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes. *Journal of Vision*, 14(3), 2014.
- [116] Ramanathan Subramanian, Julia Wache, Mojtaba Khomami Abadi, Radu L. Vieriu, Stefan Winkler, and Nicu Sebe. Ascertain: Emotion and personality recognition using commercial sensors. *IEEE Transactions on Affective Computing*, PP(99):1–1, 2016.
- [117] Ramanathan Subramanian, Yan Yan, Jacopo Staiano, Lanz, Oswald, and Nicu Sebe. On the relationship between head pose, social attention and personality prediction for unstructured and dynamic group interactions. In *Int’l Conference on Multimodal Interaction*, pages 3–10, 2013.
- [118] Susan Sullivan, Anna Campbell, Sam B Hutton, and Ted 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.
- [119] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.
- [120] Robert Tibshirani, Jacob Bien, Jerome Friedman, Trevor Hastie, Noah Simon, Jonathan Taylor, and Ryan J Tibshirani. Strong rules for discarding predictors in lasso-type problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74(2):245–266, 2012.
- [121] Serdar Tok, Mehmet Koyuncu, Seda Dural, and Fatih Catikkas. Evaluation of International Affective Picture System (IAPS) ratings in an athlete population and its relations to personality. *Personality and Individual Differences*, 49(5):461–466, 2010.

BIBLIOGRAPHY

- [122] Patricia Valdez and Albert Mehrabian. Effects of color on emotions. *Journal of Experimental Psychology: General*, 123(4):394–409, 1994.
- [123] Alessandro Vinciarelli and Gelareh Mohammadi. A survey of personality computing. *IEEE Trans. Affective Computing*, 2014.
- [124] Carl Vondrick, Donald Patterson, and Deva Ramanan. Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1):184–204, 2013.
- [125] Julia Wache, Ramanathan Subramanian, Mojtaba Khomami Abadi, Radu L. Vieri, Stefan Winkler, and Nicu Sebe. Implicit Personality Profiling Based on Psycho–Physiological Responses to Emotional Videos. In *Int’l Conference on Multimodal Interaction*, 2015.
- [126] Hee Lin Wang and Loong-Fah Cheong. Affective understanding in film. *IEEE Trans. Circ. Syst. V. Tech.*, 16(6):689–704, 2006.
- [127] Shangfei Wang and Qiang Ji. Video affective content analysis: a survey of state-of-the-art methods. *IEEE Trans. on Affective Computing*, 6(4), 2015.
- [128] Jacqueline Wijsman, Bernard Grundlehner, Julien Penders, and Hermie Hermens. Trapezius muscle emg as predictor of mental stress. In *Wireless Health 2010*, pages 155–163, 2010.
- [129] Jason D Williams, I Dan Melamed, Tirso Alonso, Barbara Hollister, and Jay Wilpon. Crowd-sourcing for difficult transcription of speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 535–540, 2011.
- [130] Kathy A Winter and Nicholas A Kuiper. Individual differences in the experience of emotions. *Clinical Psychology Review*, 17(7):791–821, 1997.

- [131] Charlotte VO Witvliet and Scott R Vrana. Play it again Sam: Repeated exposure to emotionally evocative music polarises liking and smiling responses, and influences other affective reports, facial emg, and heart rate. *Cognition and Emotion*, 21(1):3–25, 2007.
- [132] Xuehan Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.
- [133] Y Yan, E Ricci, R Subramanian, G Liu, and N Sebe. Multi-task linear discriminant analysis for view invariant action recognition. *IEEE Transactions on Image Processing*, 23(12):5599–5611, 2014.
- [134] Yan Yan, Elisa Ricci, Negar Rostamzadeh, and Nicu Sebe. It’s all about habits: Exploiting multi-task clustering for activities of daily living analysis. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 1071–1075. IEEE, 2014.
- [135] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. A 3D facial expression database for facial behavior research. In *Face and Gesture*, pages 211–216, 2006.
- [136] Xiao-Tong Yuan, Xiaobai Liu, and Shuicheng Yan. Visual classification with multitask joint sparse representation. *IEEE Transactions on Image Processing*, 21(10):4349–4360, 2012.
- [137] Jenny Yuen, Bryan C. Russell, Ce Liu, and Antonio Torralba. Labelme video: Building a video database with human annotations. In *International Conference on Computer Vision*, pages 1451–1458, 2009.
- [138] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. Space speaks: towards socially and personality aware visual surveillance. In *ACM Int’l Workshop on Multimodal Pervasive Video Analysis*, pages 37–42, 2010.

BIBLIOGRAPHY

- [139] Gloria Zen, Negar Rostamzadeh, Jacopo Staiano, Elisa Ricci, and Nicu Sebe. Enhanced semantic descriptors for functional scene categorization. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 1985–1988. IEEE, 2012.
- [140] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, Jan 2009.
- [141] Tianzhu Zhang, Bernard Ghanem, Si Liu, and Narendra Ahuja. Robust visual tracking via structured multi-task sparse learning. *International journal of computer vision*, 101(2):367–383, 2013.
- [142] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.

