Implicit and automated emotional tagging of videos

Soleymani, Mohammad

UNIVERSITÉ DE GENÈVE FACULTÉ DES SCIENCES

Département d'Informatique Professeur Thierry Pun

Implicit and Automated Emotional Tagging of Videos

THÈSE

présenté à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention informatique

par

Mohammad SOLEYMANI

de

Téhéran (IRAN)

Thèse N^{o} 4368

GENÈVE
Repro-Mail - Université de Genève
2011

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès sciences
Mention informatique

Thèse de *Monsieur Mohammad SOLEYMANI*

intitulée :

"Implicit and Automated Emotional Tagging of Videos"

La Faculté des sciences, sur le préavis de Messieurs Th. PUN, professeur ordinaire et directeur de thèse (Département d'informatique), S. MARCHAND-MAILLET, docteur (Département d'informatique), de Madame M. PANTIC, professeure (Department of Computing, Imperial College, London, United Kingdom - University of Twente, Enschede, The Netherlands) et de Messieurs D. GRANDJEAN, professeur assistant (Faculté de psychologie et des sciences de l'éducation, Section de psychologie) et G. CHANEL, docteur (Département d'informatique et Faculté de psychologie et des sciences de l'éducation, Pôle de recherche national en sciences affectives), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 4 novembre 2011

Thèse - 4368 -

Le Doyen, Jean-Marc TRISCONE

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

UNIVERSITÉ DE GENÈVE
Département d'Informatique

FACULTÉ DES SCIENCES
Professeur Thierry Pun

Implicit and Automated Emotional Tagging of Videos

THÈSE

présenté à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur ès sciences, mention informatique

par

Mohammad SOLEYMANI
de
Téhéran (IRAN)

Thèse N^{o} 4368

GENÈVE
Repro-Mail - Université de Genève
2011

UNIVERSITÉ DE GENÈVE

FACULTÉ DES SCIENCES

Doctorat ès sciences
Mention informatique

Thèse de *Monsieur Mohammad SOLEYMANI*

intitulée :

"Implicit and Automated Emotional Tagging of Videos"

La Faculté des sciences, sur le préavis de Messieurs Th. PUN, professeur ordinaire et directeur de thèse (Département d'informatique), S. MARCHAND-MAILLET, docteur (Département d'informatique), de Madame M. PANTIC, professeure (Department of Computing, Imperial College, London, United Kingdom - University of Twente, Enschede, The Netherlands) et de Messieurs D. GRANDJEAN, professeur assistant (Faculté de psychologie et des sciences de l'éducation, Section de psychologie) et G. CHANEL, docteur (Département d'informatique et Faculté de psychologie et des sciences de l'éducation, Pôle de recherche national en sciences affectives), autorise l'impression de la présente thèse, sans exprimer d'opinion sur les propositions qui y sont énoncées.

Genève, le 4 novembre 2011

Thèse - 4368 -

Le Doyen, Jean-Marc TRISCONE

N.B. - La thèse doit porter la déclaration précédente et remplir les conditions énumérées dans les "Informations relatives aux thèses de doctorat à l'Université de Genève".

Acknowledgements

First and foremost, I would like to thank Thierry to give me the opportunity to work on this thesis. He has been always a supportive advisor who provided me with both freedom and trust. Thanks to Guillaume for many things that I cannot count, from being a great office mate and colleague to a true friend helping me at hard moments and sharing lots of memorable experiences. Thanks to Joep who thought me how to think critically. Thanks to Didier who was always generous in helping me with his rich neuro-psychological knowledge. Thank to Donn, for helping me with his impressive knowledge on systems and computer science and for the outdoorsy and fun experiences we shared. Thanks to Eniko with her lovely smile who has been and hopefully will be a good friend. Thanks to Maurits for being a supportive office mate and friend. I would like to give special thanks to Fokko for creating the first version of the LaTeXtemplate that I used for this thesis and all his support and kindness. CVML folks were all very friendly and helpful, the lunch time chats and coffee time discussions helped me a lot to get new ideas and feel comfortable. Stéphane, Eric, Benoît, Sophie, Jana, Farzad, Sun and Slava, thank you all. Thanks to Nicolas, Daniel, Elie and Germaine for their administrative and technical supports. I really feel lucky to have colleagues and friends like you.

I would also like to thank the researchers with whom I collaborated outside of the UniGe. I thank Maja and Jeroen at Imperial College London for giving me the opportunity to work with their group. I have certainly learned a lot and took advantage of my visits to London. Your contributions were certainly essential to the success of this thesis. Thanks to Jozef, for helping me with the experiments in London and being a welcoming host. I would like to express my exceptional thanks to Martha at Delft University of Technology who showed me an excellent example of a confident, hard working and ambitious researcher. Thanks to Sander and Christian for sharing their knowledge and the friendly discussions. I certainly owe a debt of gratitude to my Masters advisor in Iran, Prof. Hossein-Zadeh. I appreciate his patience with me when I was a beginner and un-experienced in research.

Not to forget, my parents, my brother, Majid and Saeid, and my sisters, Mehri, Zohreh and Nahid, who always supported me throughout my studies and encouraged me to be ambitious. Finally, I would like to thank my wife, Gretchen, for everything.

Résumé

Les émotions jouent un rôle important dans la sélection du contenu multimédia et la consommation de video des spectateurs. Lrd objectifs principals de cette thèse, sont de détecter et d'estimer les caractéristiques affectives de vidéos en se basant sur l'analyse automatique des contenus, ainsi que de reconnaitre les émotions ressenties par les spectateurs en réponse aux vidéos. Ces caractérisations émotionnelles peuvent être utilisées pour étiqueter/marquer le contenu. L'étiquetage implicite ou automatique des vidéos utilisant des informations affectives aide à améliorer la performance des systèmes de recommandation et de recherche.

Dans cette thèse, une base sur la théorie des émotions et le procédé pour les approches utiliseés pour conduire des expériences dans le domaine affective. Une étude de la littérature présente les études existants en matière de compréhension affective des vidéos utilisant l'analyse du contenu, et sur les techniques qui existent pour l'évaluation des émotions en réponse à des vidéos. Quatre collections de vidéos émotionnelles ont été développées, c'est à dire qu'elles contiennent des vidéos dont le contenu suscite des émotions ; ces vidéos sont utilisées comme stimuli pour les expériences et pour la compréhension émotionnelle des vidéos par analyse du contenu. Trois corpus émotionnels incluant des réponses émotionnelles individuelles aux vidéos de la part de plusieurs participants ont aussi été enregistrés.

Les axes d'analyses et d'évaluations dans cette thèse sont de deux ordres : premièrement, les méthodologies et les résultats des méthodes de reconnaissance utilisées pour détecter l'émotion en réponse aux vidéos sont présentés. Deuxièmement, les méthodologies et les résultats de la compréhension émotionnelle des medias utilisant l'analyse du contenu sont fournis. Dans le premier axe, une méthode de détection d'émotion dans un espace continu basée sur la régression est présentée et les corrélations entre les auto-évaluations des émotions et les réponses physiologiques sont montrées. Par ailleurs, une étude sur la reconnaissance d'émotions indépendamment du participant est présentée. La deuxième étude montre la performance d'une approche de reconnaissance d'émotions utilisant les signaux EEG, la distance du regard, et la réponse pupillaire des participants comme rétroactions affectives. La faisabilité d'une approche de reconnaissance d'émotions en réponse à des vidéos utilisant un tel système est montrée. Les meilleures précisions de classification de 68.5% pour trois niveau de valence et de 76.4% pour trois niveau d'arousal sont obtenues en utilisant une stratégie de fusion des modalités et une machine à support de vecteurs. Après avoir étudié les réponses à des scènes de film, les résultats et les méthodes pour l'évaluation des émotions en réponses à des clips musicaux sont donnés.

Par ailleurs, des méthodes d'analyse de contenu permettant de détecter les émotions, qui sont les plus susceptibles d'être induites par un contenu multimédia donné, sont présentées. Des caractéristiques de bas niveau du contenu qui sont utilisées pour la compréhension affective sont introduites. Encore une fois, la méthode de régression est utilisée pour la compréhension affective des vidéos, et la corrélation entre les caractéristiques du contenu, les réponses physiologiques et l'auto-évaluation des émotions ont été étudiées. Il est montré que les corrélations des caractéristiques multimédia avec les caractéristiques physiologiques et l'auto-évaluation de par l'utilisateur sont significatives. Ceci demontre l'utilité des réponses physiologiques et des caractéristiques du contenu pour l'étiquetage émotionnel des vidéos. Ensuite, un système de représentation affective pour estimer les émotions ressenties au niveau de la scène a été proposé en utilisant un système de classification bayésien. La précision de classification de 56%, obtenue sur trois classes d'émotions avec un classifieur bayésien naïf, a été améliorée à 64% après avoir utilisé des information a-priori sur le genre et l'etat emotioniel precedent.

En conclusion, des résultats prometteurs ont été obtenus dans le marquage émotionnel des vidéos. Cependant, la compréhension émotionnelle de contenu multimédia est une tâche difficile et avec l'état de l'art actuel, une solution universelle pour détecter et étiqueter tous les contenus qui conviennent à tous les utilisateurs n'est pas possible. Les systèmes basés sur l'informatique affective fonctionnent seulement s'ils sont capables de prendre en compte les profils contextuels et personnels.

Abstract

Emotions play a pivotal role in viewers' content selection and use. The main aim of this study is to detect and estimate affective characteristics of videos based on automated content analysis as well as to recognize the felt emotions in viewers in response to videos. These emotional characterizations can be used to tag the content. Implicit or automated tagging of videos using affective information help recommendation and retrieval systems to improve their performance.

In this thesis, a background on emotion theories and the process of emotional experience in response to videos are given. The literature review sheds light on existing studies on affective understanding of videos using content analysis and the existing techniques in emotion assessment in response to videos. Four emotional video datasets are developed. They consist of emotional videos to be used as stimuli for experiments and emotional understanding of videos by content analysis. Three emotional corpora including emotional, bodily responses to videos from multiple participants have been also recorded.

The analysis and evaluations directions in this thesis are twofold: first, methodology and results of emotion recognition methods employed to detect emotion in response to videos are presented. Second, methodology and results of emotional understanding of multimedia using content analysis are provided. In the first direction, a regression based method to detect emotion in continuous space is presented, and the correlates of emotional self assessments and physiological responses are shown. Moreover, a multi-modal participant independent emotion recognition study is presented. The second study shows the performance of an inter-participant emotion recognition for tagging using participants' ElectroEncephaloGram (EEG) signals, gaze distance and pupillary response as affective feedbacks. The feasibility of an approach to recognize emotion in response to videos using such a system is shown. The best classification accuracy of 68.5% for three labels of valence and 76.4% for three labels of arousal are obtained using a modality fusion strategy and a support vector machine. After studying the responses to movie scenes, the results and the methods for emotion assessment in response to music clips are given.

Moreover, content analysis methods to detect emotions that are more likely to be elicited by a given multimedia content are presented. Low level content features, which are used for affective understanding, are introduced. Again the regression method is used for affective understanding of videos and the correlation between content features, physiological responses and emotional self reports have been studied. Content based multimedia features' correlations with both physiological features and users' self-assessment of valence-arousal are shown to be significant. This implies the usefulness of physiological responses and content features for emotional tagging of videos. Next, an affective representation system for estimating felt emotions at the scene level has been proposed using a Bayesian classification framework. The classification accuracy of 56%

that was obtained on three emotional classes with a naïve Bayesian classifier was improved to 64% after utilizing temporal and genre priors.

In conclusion, promising results have been obtained in emotional tagging of videos. However, emotional understanding of multimedia is a challenging task and with the current state of the art methods a universal solution to detect and tag all different content which suits all the users is not possible. Systems based on affective computing can only work if they are able to take context and personal profiles into account.

Contents

[1] This study was done in collaboration with, Sander Koelstra, Christian Mühl, Ashkan Yazdani, and Jong-Seok Lee in the context of Petamedia European network of excellence.

List of Figures

[2] http://youtube-global.blogspot.com/2010/11/great-scott-over-35-hours-of-video.html

Acronyms and Abbreviations

ANOVA	ANalysis Of VAriance
ANS	Autonomous Nervous System
BCI	Brain Computer Interface
BP	Blood Pressure
BVP	Blood Volume Pulse
CNS	Central Nervous System
CMRR	Common-Mode Rejection Ratio
CMS	Common Mode Sense
DBN	Dynamic Bayesian Network
DFA	Discriminant Function Analysis
DLF	Decision Level Fusion
ECG	ElectroCardioGram
EEG	ElectroEncephaloGram
EMG	ElectroMyoGram
EOG	ElectroOculoGram
FLF	Feature Level Fusion
fNIRS	functional Near InfraRed Spectroscopy
GSR	Galvanic Skin Response
HR	Heart Rate
HSL	Hue, Saturation, Lightness
HSV	Hue, Saturation, Value
HIT	Human Intelligent Task
IADS	International Affective Digitized Sound system
IAPS	International Affective Picture System
KNN	K-Nearest-Neighbor
LDA	Linear Discriminant Analysis
MBP	Marquardt BackPropagation
MFCC	Mel Frequency Cepstral Coefficients

MTurk	Amazon Mechanical Turk
NES	Neuro-Endocrine System
GMM	Gaussian Mixture Models
MLP	MultiLayer Perceptron
MPEG	Moving Picture Experts Group
OCC	Orthoney, Clore and Collins
PAD	Pleasure-Arousal-Dominance
PANAS	Positive And NegAtive Schedule
PC	Personal Computer
PCA	Principal Component Analysis
pLDA	pseudoinverse Linear Discriminant Analysis
Plet	Plethysmograph
QDA	Quadratic Discriminant Analysis
RBFN	Radial Basis Function Networks
RSP	ReSPiration amplitude
RVM	Relevance Vector Machine
SAM	Self Assessment Manikin
SD	Standard deviation
SEC	Stimulus Evaluation Checks
SNS	Somatic Nervous System
SVM	Support Vector Machine
Temp	Temperature
USB	Universal Serial Bus

Chapter 1

Introduction

The digital age changed the way multimedia content is generated. Multimedia content used to be only generated by a handful of big companies, record producers and television and radio stations. Today, everybody can easily record, edit and publish multimedia content using handheld devices. According to YouTube[1] weblog [4], YouTube users, in 2010, were uploading 35 hours of videos in every minute. Although these videos can include redundant content, this number is very large considering that YouTube is only one of the video repositories with user generated content.

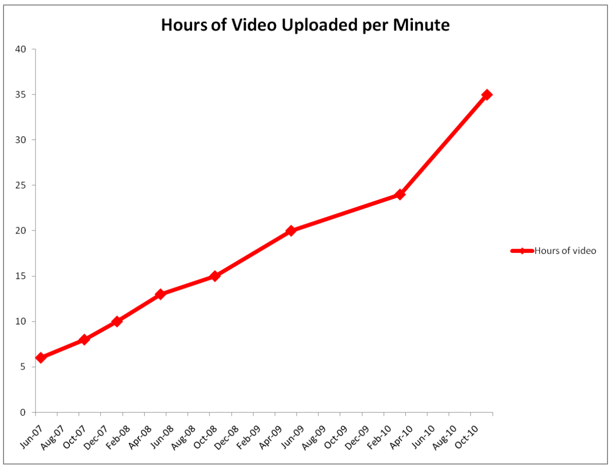


Figure 1.1: The video upload rate on youtube from June 2007 in hours per minute. This figure is taken from YouTube's official blog[3].

This number is almost seven times bigger comparing to June 2007 (see Fig. 1.1). The videos uploaded on youtube in 2010 for one week is approximately equal to 176000 full length Hollywood movies. This massive amount of digital content needs indexing to be searchable and accessible by users. Most of the current multimedia retrieval systems use user generated tags to index videos. Social media websites therefore encourage users to tag their content. However, the users' intent when tagging multimedia content does not always match the information retrieval goals which is to index the content with meaningful describing terms. A large portion of user defined tags are either motivated by the goal of increasing the popularity and reputation of a user in an online

[1] www.youtube.com

community or based on individual and egoistic judgements [5].

Multimedia indexing needs relevant terms, which can describe its content, genre and category. Researchers in multimedia therefore focused on generating characterizations of videos involving cognitive concepts depicted in videos to facilitate their indexing [6]. The cognitive concepts are selected in a way that is easy to detect by human. An alternative to the cognitive approach to indexing is a paradigm for video indexing based on the emotions which are felt by viewers watching the content [7]. This is motivated by the fact that emotions play an important role in viewers' content selection and consumption.

When a user watches video clips or listens to music, he/she may experience certain feelings and emotions [8, 9, 10] which manifest through bodily and physiological cues, e.g., pupil dilation and contraction, facial expressions, e.g., frowning, and changes in vocal features, e.g., laughter. Affective video indexing aims at deriving representations of video that characterize the emotions that they elicit. Affective experience in response to a video is personal which means it depends on the experience of a single viewer. However, there exist a more popular response which is the basis of movie genres we know, e.g., drama, comedy, horror. The main aim of the study which was done and reported in this thesis is to detect and estimate these affective characteristics of videos based on automated content analysis emotion prediction or to recognize emotions in response to videos with the aim of affective indexing. Implicit tagging refers to the effortless generation of subjective tags based on users non-verbal behavioral responses to a content. Implicit tagging of videos using affective information can help recommendation and retrieval systems to improve their performance [11, 12, 13]. Automated affective tagging tries to estimate the emotional response which is more likely to be elicited by content, e.g., loud sound or faster motions are more likely to elicit excitement, whereas implicit tagging affective tagging uses the responses of users to generate tags. Implicit or automated tagging thus does not interrupt users while listening or watching a video. Moreover, in the presence of reliable implicit or automated tagging methods, determined tags carry less irrelevant and inaccurate information. The set of relevant and refined indexing terms will improve the multimedia information retrieval results.

In automated affective tagging, a set of content features, e.g., motion component, color variance, audio energy, extracted from the content are used to estimate the emotions which are likely to be elicited after showing the content. This approach is an open loop because it does not rely on the response of the viewers and can be only used to provide an approximation for possible affective reactions to the content. The automated affective tagging can be used as a starting point to filter the data in a cold start scenario.

On the other hand, the implicit tagging results can enrich the user generated tags and automated tags to improve their quality. In order to translate a user's bodily and behavioral reactions to emotions, reliable emotion assessment techniques are required. Emotion assessment is a challenging task; even users are not always able to express their emotion by words and the emotion self-reporting error is not negligible. This is therefore difficult to define a ground truth. Affective self-reports might be held in doubt because users cannot always remember all the different emotions they had felt during watching a video, and/or might misrepresent their feelings due to self presentation, e.g., a user wants to show he is courageous whereas, in reality, he was scared [14]. The emotion recognition system provides us with an alternative that reduces the effort of deciding on the right label and on defining the right questions or methods to assess emotions explicitly.

1.1 Application Scenario

In a real system, every item usually comes with user generated tags. The automated or content based affective tagging can provide additional tags, which can semantically describe the content, e.g., exciting video, slow or sad video. Implicit tagging can further enhance these tags by first providing reliable tags for a given content and then by providing a more genuine ground truth for further processing of the content. These two tagging strategies together can reinforce each other by providing feedback and initial tags.

What I mean by affective content of videos is not the emotion expressed in the content but the emotion felt by the users while watching the videos. Users do not evaluate media content on the same criteria. Some might tag multimedia content with words to express their emotion while others might use tags to describe the content. For example, a picture receive different tags based on the objects in the image, the camera by which the picture was taken or the emotion a user felt looking at the picture. Scherer defines this by intrinsic and extrinsic appraisal [15]. Intrinsic appraisal is independent of the current goals and values of the viewer while extrinsic or transactional appraisal leads into feeling emotions in response to the stimuli. For example, the content's intrinsic emotion of a picture with a smiling face is happiness whereas this person might be a hatred figure to the viewer and the extrinsic appraisal leads into unpleasant emotions. What I want to detect is the later one that is the emotion felt by the viewer.

In the proposed application scenario, when a user watches a video his/her emotional expressions will be detectable by sensors and facial and body tracking methods. These responses can be reliably used to reliably generate affective tags. A scheme of implicit tagging scenario versus explicit tagging is shown in Fig. 1.2.

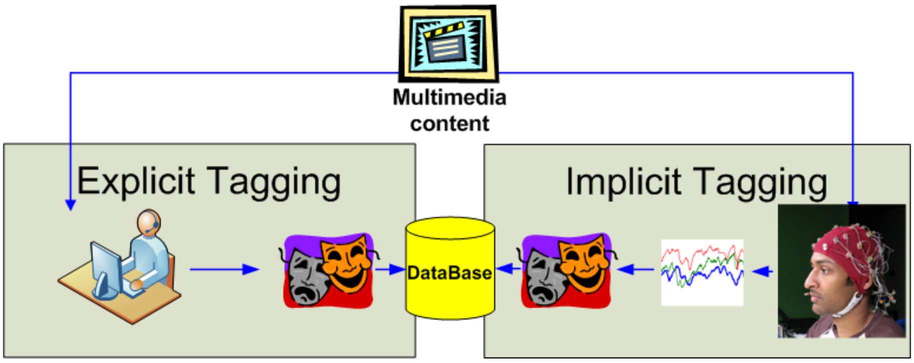


Figure 1.2: Implicit affective tagging vs. explicit tagging scenarios. The analysis of the bodily responses replaces the direct interaction between user and the computer. Therefore, users do not have to be distracted for tagging the content.

In the proposed implicit tagging scenario, multimedia content will be tagged based on the bodily reactions of users recorded by a physiological acquisition device and an eye gaze tracker. The reactions can be used both to find tags common to a population and to develop a personal profile possibly in fusion with user preferences and browsing history. With the recently marketed physiological devices such as Neurosky [4], Emotiv helmet [5], and Q-sensor [6] physiological interfaces are likely going to be the emerging human computer interfaces of the future.

1.2 Contributions

The contributions and achievements of the current thesis can be summarized in the following items:

- Study of the correlation between low level video and audio content features, such as lighting, motion component, audio energy, and emotional responses, i.e., arousal and valence self reports, and physiological responses, e.g. facial muscle activities, Galvanic Skin Response (GSR).
- Development and evaluation of a participant-dependent regression based method to detect emotion in continuous space using peripheral nervous system physiological responses.
- Development and evaluation of a multi-modal participant independent emotion recognition method with pupil dilation, gaze distance and ElectroEncephaloGram (EEG) signals. This emotion recognition method was developed to satisfy a video implicit tagging system needs.
- Development and evaluation of a regression based content analysis method to estimate the affect of viewers from the content of the video.
- Proposition of temporal and genre priors to be used in a Bayesian framework for affective classification of videos based on their content.
- Proposition of a text analysis approach for affective characterizations of movie scenes based on their subtitles.
- Development of two databases with emotional responses of 27 and 32 participants to movie scenes, online user generated videos, and music videos.
- Development of video benchmarks for affective characterization and boredom ranking.

These contributions have been reflected in the publications by the author during his thesis work [13, 3, 16, 10, 17, 18, 19, 2, 20, 21, 22]. These publications are also listed in Appendix 6.3.

1.3 Thesis Overview

In Chapter 2, I give a background on emotion theories, affective content analysis and emotion recognition in response to videos. Next, in Chapter 3, existing emotional corpora are listed and presented and then the developed video corpora for stimuli and affective analysis as well as databases of emotional responses are presented. In Chapter 4, methodology and results of emotion recognition methods employed to detect emotion in response to videos are presented. In Chapter 5, content analysis methods to detect emotions that are more likely to be elicited by a given multimedia content and their evaluations are given. Finally, Chapter 6 concludes the thesis and gives the perspectives.

[4] http://www.neurosky.com/

[5] http://www.emotiv.com/

[6] http://www.affectiva.com/q-sensor/

Chapter 2

Background

2.1 Why Do We Feel Emotions?

Emotions are complex phenomena with affective, cognitive, conative and physiological components [23]. The affective component is the subjective experience in connection with feelings. The perception and evaluation of the emotional situation is the cognitive component. The conative component is about affective expression. The conative component includes facial expressions, body gesture, and any other action which has a preparatory function for action in response to the emotional situation. The physiological components regulate physiological responses in reaction to the emotional situation, for example, increasing perspiration during a fearful experience.

Darwin suggested that emotions only exist due to their survival value and therefore, should be observable in both human and animals [24]. James [25] pioneered the idea that emotions have certain peripheral physiological responses. He suggested that emotions are felt due to changes in physiological responses. Social constructionists support another theory for the origin of emotions. They claim emotions are only products of social interaction and cultural rules. Darwinian theory of emotion emphasizes the evolution history of species and the effect of emotions on their survival whereas social constructionists emphasize the history of individuals in generating similar bodily responses to emotions [24].

Cognitive theories of emotion are the most recent theories developed to define emotion. Amongst the cognitive theories, one of the most well-accepted ones which explains the emotional process is the appraisal theory. According to this theory, cognitive judgment or appraisal of the situation is a key factor in the emergence of emotions [26, 27, 1]. According to Orthoney, Clore and Collins (OCC) [27] emotions are experienced with the following scenario. First, there is a perception of an event, object or an action. Then, there will be an evaluation of events, objects or action according to personal wishes and norms. Finally, the perception and evaluation result in a specific emotion. Considering the same scenario for an emotional experience in response to multimedia content, emotions arise first through sympathy with the presented emotions in the content [23]. During the appraisal process for an emotional experience in response to multimedia content, the viewer examines events, situations and objects with respect to their novelty, pleasantness, goal, attainability, copability, and compatibility with his/her norms. Then, the viewer's perception induces specific emotions, which changes the viewer's physiological responses, motor actions, and feelings. The component process model, proposed by Scherer [26, 28], will be discussed in more detail, in the following Section.

There are three categories of emotional processes in response to multimedia which are *emotion induction*, *emotional contagion* and *empathic sympathy* [23]. An example of *emotion induction* is when in a TV show, a politician's comment makes the viewers angry while he is not angry himself. The angry response from the viewers is due to their perception of the situation according to their goals and values. The *emotional contagion* happens when the viewer only perceives the expressed emotion from a video. For example, the induced joy as a result of sitcom laughters can be categorized in this category. In the *empathic* category, the situation or event does not affect the viewer directly but the viewer follows the appraisal steps of the present characters in the multimedia content. The *empathic* reaction can be symmetric co-emotion when the viewer has positive feelings about the character or asymmetric co-emotions in case of negative feeling about the character [29].

Empathy is a complex phenomenon which has cognitive and affective components. The affective empathy is the primitive response for sympathizing with another individual. On the contrary, cognitive empathy is understanding another person and the rational reconstruction of his/her feelings [30, 23]. Zillman developed the affective disposition theory for narrative plot [31, 29]. According to this theory, *empathic* emotions are originated from the observation of the actors by viewers. First, the character's actions are morally judged by the viewer and the judgment results in a positive or negative perception of the character. Then, depending on the character's approval or disapproval from the viewer, the viewer sympathizes emphatically or the counter empathy might happen. The intensity of the perceived emotion in response to a film depends on how much the viewer identifies himself/herself with the heroes and to what extent the one's own identity is given up while watching the video [31]. The most important emotion inducing components of movies are narrative structures and music [23].

One should be careful not to mix moods and emotions. Mood is an affective diffused state that is long, slow moving and not tied to a specific object or stimulus whereas emotions can occur in short moments with higher intensities [15].

2.2 Emotional Representations

2.2.1 Component processes

The component process model of emotions is based on the cognitive theory of emotion in which emotion is considered a cognitive process. Scherer defines emotion as "an episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism" [15]. In the absence of any of the components and factors, an affective phenomenon cannot be called emotion. For example while "feeling" is the subjective experience and an affective component of emotion, it is not the synonym of emotion [15]. Mood can be also distinguished from emotion due to its lack of specific trigger.

According to the component process model proposed by Scherer, five components corresponding to five different functions are engaged in an emotion episode. These five components are cognition, peripheral efference, motivation, motor expression and subjective feeling [1]. The cognitive component evaluates objects and events in a process called Stimulus Evaluation Checks (SEC). In SEC process, an organism examines the relevance of an action or event considering the following questions:

1. Is it relevant to the organism or does it affect the organism(relevance)?
2. What are the implications of the event or action regarding organism's goals and well-being (implications)?
3. Can the organism cope with it (copping potential)?
4. Does the event or action have any conflict with social normative significance or values (normative significance)?

According to [1] the first question can be answered by checking the novelty of the stimulus, its intrinsic pleasantness and its goal or need relevance. Implications of an event or action for an organism can be checked by causal attribution check or understanding who and why caused the action or event. The implications can be further evaluated by checking the outcome probability, goal/need conduciveness or positive effect on the organism's goals and needs and the urgency of the action or event. An action or event can be urgent when it endangers organism's goals and needs and requires fast response, e.g., fight or flight actions. Coping potential can be determined by evaluating the control of agents or organism on the event or action, the power of the organism over the stimulus and its potential for adjustment with the event or action and its consequences. Normative significance is evaluated for socially living species by checking the organism's internal and external standards.

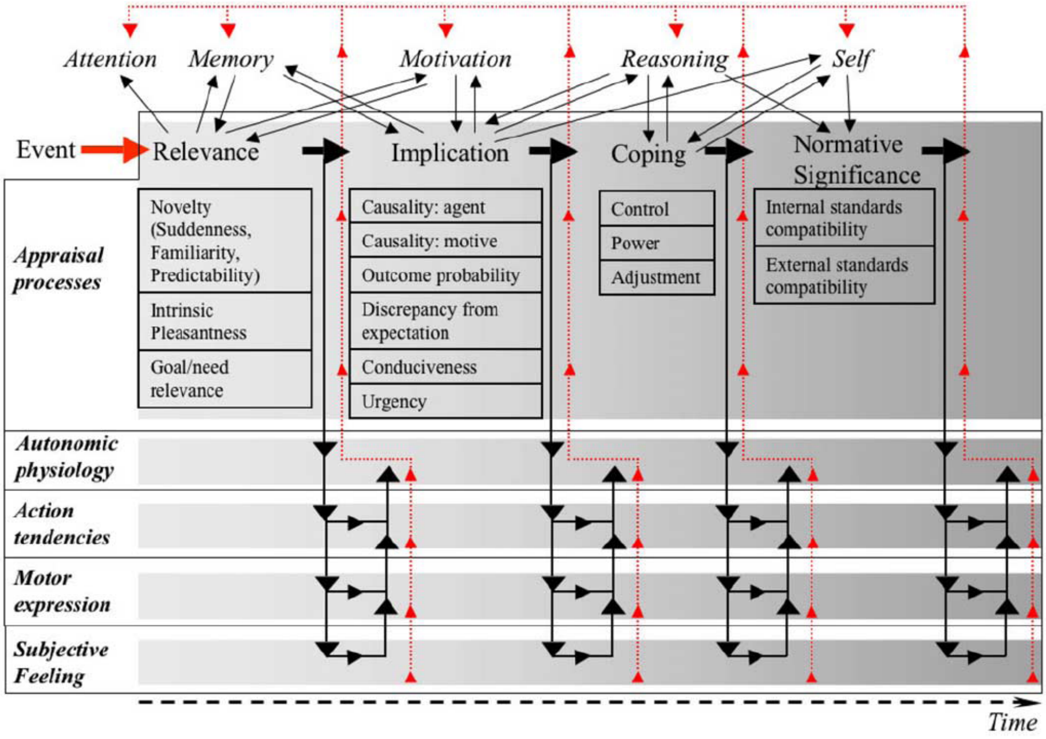


Figure 2.1: The component process model of emotion (This figure is taken from [1]).

The component process model evaluates an event or action at different layers (see Fig. 2.1). Appraisal and other cognitive functions can have a bi-directional effect on each other. A weak stimulus can create a large effect after being evaluated and giving a positive feedback to the system. For example, a relevant stimulus can increase the importance of a previously irrelevant stimulus after the relevance check in the appraisal mechanism.

The SEC evaluation will make effect on the autonomic physiology, action tendencies, motor expressions and eventually subjective feeling. The Autonomous Nervous System (ANS) along with Central Nervous System (CNS) and Neuro-Endocrine System (NES) regulate the neurophysiological component or body changes to prepare the individual for the emotional reaction. This reaction can be more intense in case of fight or flight activations, e.g., in case of fear or anger. The Somatic Nervous System (SNS) plays the role of communication of reaction and behavioral intention by driving the motor expression component, e.g., facial and vocal expression. Eventually, the central nervous system monitors the internal state and surrounding environment reactions and causes subjective feeling [15].

The emotional experience is a result of effect of an action or event on all subsystems which is triggered by SEC evaluation. The subjective feeling component communicates with other subcomponents and makes the organism to communicate the felt emotions with others [28].

2.2.2 Discrete models

Discrete emotions theories are inspired by Darwin and support the idea of the existence of the certain number of basic and universal emotions [15, 24]. Darwin suggested that emotions exist due to their importance for species' survival. Different psychologists proposed different lists of basic emotions. The so called basic emotions are mostly utilitarian emotions, and their number is usually from 2 to 14.

Plutchik proposed eight primary emotions; namely, anger, fear, sadness, disgust, surprise, anticipation, trust, and joy. He suggested that basic emotions are biologically primitive and have evolved in order to increase the reproductive success of animals [32]. Ekman studied the universality of emotions based on facial expressions and his list of basic emotion included fear, anger, surprise, disgust, joy, and sadness [24]. Scherer suggests using the term “modal” instead of “basic” emotions. Scherer also proposed a list of emotional keywords to code discrete and free choice emotional reports [15].

2.2.3 Dimensional models

Wundt [33] was the first to propose a dimensional representation for emotions. Dimensional theories of emotion suggest that emotions can be represented as points in a continuous space, and discrete emotions are folk-psychological concepts [34]. Discrete emotions also have problems in representing emotions. The main one being that keywords are not cross-lingual: emotions do not have exact translations in different languages, e.g., “disgust” does not have an exact translation in Polish [35]. Psychologists often represent emotions in an n-dimensional space (generally 2- or 3-dimensional). The most famous such space, which is used in the present study and originates from cognitive theory, is the 3D valence-arousal-dominance or Pleasure-Arousal-Dominance (PAD) space [36]. The valence scale ranges from unpleasant to pleasant. The arousal scale ranges from passive to active or excited. The dominance scale ranges from submissive (or “without control”) to dominant (or “in control, empowered”). Fontaine et al. [37] proposed adding predictability dimension to PAD dimensions. Predictability level describes to what extent the sequence of events is predictable or surprising for a person.

2.3 Emotional Self-Reporting Methods

Understanding the true emotion which was felt by a subject during an experiment has been always a challenge for psychologists. Multiple, emotional self-reporting methods have been created and used so far [38, 39, 15, 40, 41]. However, none of them gives a generalized, simple and accurate mean for emotional self-reporting. Emotional self-reporting can be done either by free-response or forced-choice formats. In the free-response format, the participants are free to express their emotions by words. In the forced-choice, participants are asked to answer specific questions and indicate their emotion. Forced-choice self-reports on affective experiments use either discrete or dimensional approaches. Discrete emotional approaches are based on the keywords selected considering linguistic and psychological studies which give us limited number of emotions [15]. The vocal, facial expressions and physiological changes were used to define discrete emotional categories. A set of these discrete emotions were found to be universal using facial expressions analysis and therefore, defined as basic emotions [42]. Based on these discrete emotions, self-reporting tools were developed in which users are asked to report their emotions with emotional words on nominal, and ordinal scales. Dimensional approaches of emotional self-reporting are based on bipolar dimensions of emotions. Two to four dimensional models have been proposed to describe emotions [36, 37]. Emotions can be reported on every dimension using ordinal or continuous scales [40]. I here list some popular self-reporting methods which have been used in human computer interaction.

Russell [43] introduced the circumplex model of affects for emotion representation. In his model, 8 emotions; namely, arousal, excitement, pleasure, contentment, sleepiness, depression, misery and distress are positioned on a circle surrounding a two dimensional activation, pleasure-displeasure space. Starting form these 8 categories, 28 emotional keywords were positioned on this circumplex as a result of a user study. The advantage of this circumplex over either discrete or dimensional models is that all the emotions can be mapped on the circumplex only with the angle. Therefore, all emotions are presented on a circular and one dimensional model.

SAM is one of the most famous emotional self-reporting tools. It consists of manikins expressing emotions. The emotions are varying on three different dimensions; namely, arousal, valence, and dominance [40]. The SAM Manikins are shown in Fig. 2.2. Users can choose a Manikin that best portrays their emotion. This method does not need verbalizing emotions and the manikins are understandable without any words. Hence, this tool is language independent. The second advantage of SAM Manikins is that it can be directly used in measuring the dimensional emotions. However, subjects are unable to express co-occurring emotions with this tool.

The Positive And NegAtive Schedule (PANAS) [44] permits self-reporting 10 positive and 10 negative affects on five points scale. An expanded version of PANAS, the positive and negative schedule - expanded form (PANAS-X), was developed later in which the possibility of reporting 11 discrete emotion groups on five points scale was added [45]. PANAS is made to report affective states and can be used to report both moods and emotions. PANAS-X includes 60

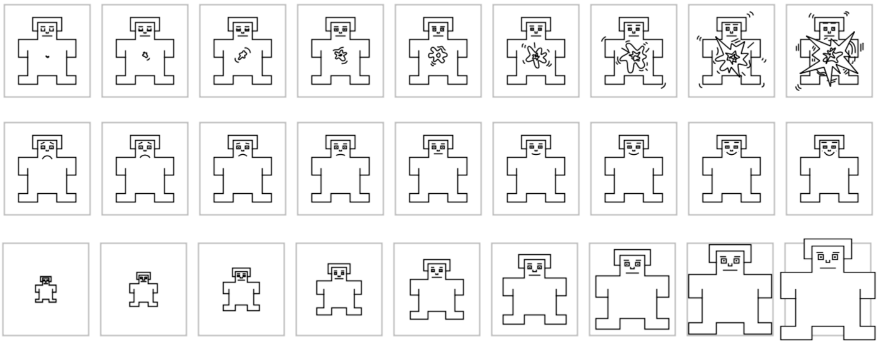


Figure 2.2: Self Assessment Manikins. From top to bottom the manikins express different levels of arousal, valence, and dominance.

emotional words and takes on average 10 minutes to be completed [45]. The time needed to fill this questionnaire make it too difficult to use in the experiments with limited time and several stimuli.

Scherer [15] positioned 16 emotions around a circle to combine both dimensional and discrete emotional approaches to create the Geneva emotion wheel. For each emotion around the wheel, five circles with increasing size from the center to the sides are displayed. The size of the circle is an indicator of the intensity of felt emotion (see Fig. 2.3). In an experiment, a participant can pick up to two emotions, which were the closest to his/her experience from 20 emotions, and then report their intensities with the size of the marked circles. In case, no emotion is felt, a user can mark the upper half circle in the hub of the wheel. If a different emotion is felt by a user, it can be written in the lower half circle. The emotions are sorted on the circle in a way to have, high control emotions on the top and low control emotions in the bottom whereas the horizontal axis which is not visible on the wheel represent valence or pleasantness.

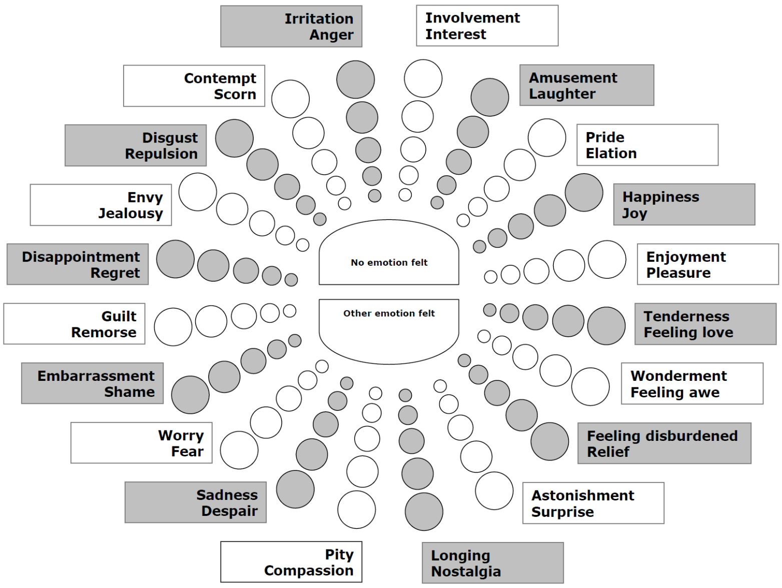
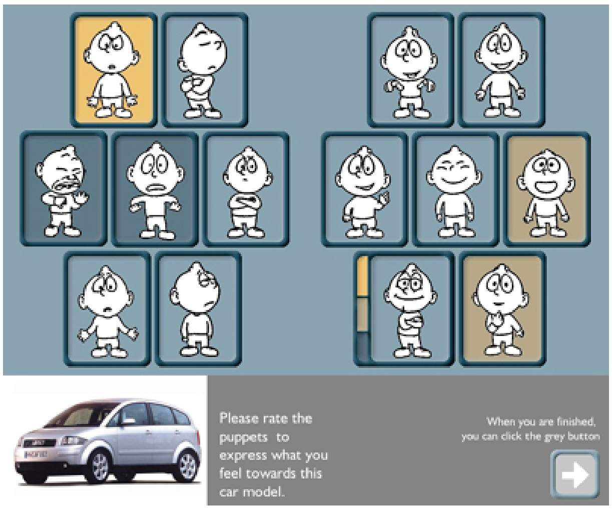


Figure 2.3: A participant can indicate his emotion on Geneva emotion wheel by clicking or choosing the circles.

PrEmo is an alternative non-verbal emotion reporting tool to report emotions in response to product design. Desmet proposed PrEmo to overcome the problem of reporting co-occurring emotions with animated characters expressing emotions [38]. PrEmo consists of 14 animated characters expressing different emotions, and it is hence language independent. Users can rate in three levels each character that they identify as the ones relevant to their felt emotions (see Fig. 2.4.



Please rate the puppets to express what you feel towards this car model.

When you are finished, you can click the grey button

Figure 2.4: Users can identify emotions they are feeling with 14 animated characters are expressing emotions. Retrieved from http://studiolab.io.tudelft.nl/desmet/PrEmo.

Zoghbi et al. proposed a joystick for emotional communication in human robot interactions [46]. Users can report the pleasantness or valence by moving the joystick forward and backward and report the level of arousal by squeezing the joystick handle. Although this tool gives the possibility to report the emotions in real time, it causes distraction for a participant in front of a stimulus and it is not yet available for other researchers to use.

2.3.1 Video affective annotation tools

Among the self-reporting tools which have been developed, a few of them were designed specifically for video affective annotation. Villon developed an annotation tool with which a user can drag and drop videos on the valence-arousal plane. This method gives the possibility of comparison between the ratings given to different videos and enable a user to give ratings relative to other videos [47]. This tool enables users to keep their previous reports into account while annotating a new video.

Feeltrace was developed to annotate the intrinsic emotion in videos [48]. This tool is originally designed to annotate the emotions expressed in videos, e.g., talk shows, acted facial expressions or gestures [49]. Although this tool gives the possibility of continuous annotation, it is not an appropriate tool for emotional self-reporting, because it is difficult for a user to both concentrate on the video and reporting changes in his emotions.

An online video affective annotation tool has been developed by Soleymani et al. [18]. In their annotation tool, a user can self-report emotions after watching a given video clip by means of SAM manikins and emotional keywords from a selected list in a drop down menu (see Fig. 3.3).

2.4 Affective Characterization of Videos

Wang and Cheong [50] used content audio and video features to classify basic emotions elicited by movie scenes. In [50], audio was classified into music, speech and environment signals and these were treated separately to shape an audio affective feature vector. The audio affective vector of each scene was fused with video-based features such as key lighting and visual excitement to form a scene feature vector. Finally, using the scene feature vectors, movie scenes were classified and labeled with emotions.

Irie et al. [51] proposed a latent topic model by defining affective audio-visual words in the content of movies to detect emotions in movie scenes. They extracted emotion-category-specific audio-visual features named affective audio-visual words. These higher level features were used to classify movie scenes using a latent topic driving model. This model takes into account temporal information which is the effect of the emotion from precedent scene to improve affect classification.

A hierarchical movie content analysis method based on arousal and valence related features was presented by M. Xu et al. [52]. In this method, the affect of each shot was first classified in the arousal domain using the arousal correlated features and fuzzy clustering. The audio short time energy and the first four Mel Frequency Cepstral Coefficients (MFCC) (as a representation of energy features), shot length, and the motion component of consecutive frames were used to classify shots in three arousal classes. Next, they used color energy, lighting and brightness as valence related features to be used for a HMM-based valence classification of the previously arousal-categorized shots. A drawback of the proposed approach is that a shot can last less than few seconds; it is thus not realistic to form a ground-truth with assigning an emotion label to each shot.

A regression based arousal and valence representation of MTV (Music-TV) clips using content features was presented in [53]. The arousal and valence values were separated into 8 clusters by an affinity propagation method. Two different feature sets were used for arousal and valence estimation which was evaluated using a ground truth. The ground truth was based on the average assessments of 11 users.

Table 2.1: The summary of video affective representation literature.

Study	**Emotion repres.**	**Dimensions or Classes**	**# Annotators**	**Modalities**	**Results**
Kang [54]	disc.	fear/anger, joy, sadness and netural	10	V	classification rate, fear: 81.3%, sadness: 76.5%, joy: 78.4%
Hanjalic & Xu [7]	cont.	valence and arousal	N/A	AV	no evaluation
Wang & Cheong [50]	disc.	fear, anger, surprise, sadness, joy, disgust and neutral	3	AV	74.7%
Arifin & Cheung [55]	cont.	pleasure, arousal, and dominance	14	AV	-
Xu et al. [52]	disc.	fear, anger, happiness, sadness and neutral	?	AV	80.7%
Irie et al. [51]	disc.	acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise and netural	16	AV	subject agreement rate 0.56

disc.: discrete, cont.: continuous, repres.: representation, V: visual, AV, audio-visual, N/A: not available

Emotional characteristics of videos have also improved music and image recommendation. Shan et al. [11] used affective characterization using content analysis to improve film music recommendation. Tkalčič et al. showed how affective information can improve image recommendation [12]. In their image recommendation scenario, affective scores of images from the International Affective Picture System (IAPS) [56] were used as features for an image recommender. They conducted an experiment with 52 participants to study the effect of using affective scores. The image recommender using affective scores showed a significant improvement in the performance of their image recommendation system.

A summary of few existing literature in emotional understanding of videos using content analysis is given in Table 2.1. There has been long standing research on emotion assessment from physiological signals [8, 57, 58, 59, 60, 61]. These studies can be divided into different categories according to the modalities recorded and the stimuli. I divide stimuli into two categories, active and passive paradigm. In an active stimulus paradigm, the participant is active in the process of eliciting an emotion whereas in a passive stimulus paradigm, a participant is passively observing or listening an emotional stimulus. The work presented in this thesis does not involve active stimuli paradigm and can be categorized as passive stimuli paradigm studies. In the following, I introduce some relevant existing passive stimuli studies.

Lisetti and Nasoz used physiological response to recognize emotion in response to movie scenes [60]. The movie scenes elicited six emotions; namely sadness, amusement, fear, anger, frustration and surprise. They achieved a high recognition rate of 84% for the recognition of these six emotions. However, the classification was based on the analysis of the signals in response to pre-selected segments, in the shown video, known to be related to highly emotional events.

Takahashi [62] recorded EEG and peripheral physiological signals from 12 participants. He then classified the responses to emotional videos into five classes; namely, joy, sadness, disgust, fear, and relax. He achieved the accuracy of 41.7% using EEG signals. However, the feature level fusion of EEG signals and peripheral physiological signals failed to improve the classification accuracy.

Kim and André [8] used music as stimuli to elicit emotions in four participants. The emotions were corresponding to the four quadrants on the valence-arousal plane; namely, positive high/low and negative high/low classes. Peripheral physiological signals were recorded while the songs corresponding to different emotional states selected by participants according to their personal preferences were playing. In a participant independent approach, their system could recognize four emotional classes with 70% accuracy.

Table 2.2: The summary of emotion recognition literature using passive stimuli paradigm and physiological signals. (Part.:participant, Clas.: classes, dep.:dependent, ind.:independent, K-Nearest-Neighbor (KNN), Linear Discriminant Analysis (LDA), Marquardt BackPropagation (MBP), pseudoinverse Linear Discriminant Analysis (pLDA), EMDC: emotion-specific multilevel dichotomous classification, Radial Basis Function Networks (RBFN), Quadratic Discriminant Analysis (QDA), MultiLayer Perceptron (MLP), ReSPiration amplitude (RSP), Temperature (Temp), Blood Pressure (BP), ElectroCardioGram (ECG), Plethysmograph (Plet), ElectroOculoGram (EOG), ElectroMyoGram (EMG), Induc.: Induction, Imp. Card.: Impedance Cardiography)

Study	**Stimuli**	**# Part.**	**Sensors**	**Classifiers**	**# Clas.**	**Classes**	**Best average result and part. dependency**
Lisetti and Nasoz [60]	video	29	GSR, Temp, accelerometer, heat flow, Heart Rate (HR)	KNN, Discriminant Function Analysis (DFA), MBP	6	sadness, anger, frustration, surprise, fear, amusement	84%, Part. ind.
Takahashi [62]	video	12	EMG, EOG, Plet, GSR, EEG	Support Vector Machine (SVM)	5	joy, sadness, disgust, fear, and relax	41.7%, Part. ind.
Bailenson et al. [63]	video	41	systolic BP, diastolic BP, mean arterial BP, GSR, ECG, Temp, Plet, Piezo-electric sensor	SVM	2	sadness/neutral, amusement/neutral	95%, Part. dep.
Kim and André [8]	music	3	ECG, RSP, GSR	EMDC, pLDA	4	positive high/low arousal, negative, high/low arousal	70%, Part. ind.
Koelstra et al. [16]	music video	5	EMG, GSR, RSP, Temp, EEG	LDA	2	low/high arousal and valence	58%, Part. dep.
Kolodyazhniy et al. [64]	video	34	ECG, EMG, Imp. Card., Temp, Capnography, RSP., Induc., Plet., BP, Piezo-electric sensor	LDA, QDA, MLP, RBFN, KNN	3	sadness, fear, neutral	77.5%, Part. ind.

Table 2.3: The summary of emotion recognition literature using active stimuli paradigm and physiological signals.(Part.:participant, Clas.: classes, Part.: participant, dep.:dependent, ind.:independent, Gaussian Mixture Models (GMM), Relevance Vector Machine (RVM), Dynamic Bayesian Network (DBN), Blood Volume Pulse (BVP))

Study	**Stimuli**	**# Part.**	**Sensors**	**Classifiers**	**# Clas.**	**Classes**	**Best average result and part. dependency**
Healey and Picard[65]	driving	24	ECG, EMG, RSP, GSR	LDA	3	three stress levels	97% Part. ind.
Picard et al. [66]	self induction	1	ECG, RSP, GSR, BVP, EMG	LDA	8	neutral, anger, hate, grief, platonic love, romantic love, joy, reverence	81% Part. dep.
Wang and Gong [58]	simulated driving	13	RSP, Temp, GSR, BVP	LDA, GMM, SVM, DBN	5	happy, angry, sad, fatigue, neutral	80% Part. dep.
Katsis et al. [67]	simulated driving	10	ECG, EMG, GSR, RSP	SVM	4	low/high stress, disappointment, euphoria	79.3% Part. dep.
Chanel et al. [59]	gaming	20	RSP, Temp, GSR, BVP, EEG	LDA, QDA, SVM	3	boredom, engagement, anxiety	63% Part. ind.
Chanel et al. [61]	recall	10	RSP, Temp, GSR, BVP, EEG	LDA, QDA, SVM, RVM	3	calm, positive excited, negative excited	70% Part. dep.

Koelstra et al. [16] recorded EEG and peripheral physiological signals of six participants in response to music videos. Participants rated their felt emotions by means of arousal, valence and like/dislike rating rating. The emotional responses of each participant were classified into two classes of low/high arousal, low/high like/dislike, and low/high valence. The average classification rates varied between 55% and 58% which is slightly above random level.

In a more recent study, Kolodyazhniy et al. [64] used peripheral physiological signals to recognize neutral, fear and sadness responses to movie excerpts [68]. During the presentation of videos to the participants, they introduced startle stimuli using randomly generated white noise sounds to boost physiological responses. Their system was able to recognize sadness, fear and neutral, emotional states with the recognition rate of 77.5% in a participant-independent approach.

Active stimuli have been also employed for emotion recognition studies. Examples are driving or simulated driving [65, 58, 67], emotion recall [59, 66] and gaming [61].

Tables 2.3 and 2.2 summarize the characteristics of different emotion recognition studies using passive and active stimuli and physiological signals.

2.4.1 Emotion recognition from Eye gaze and pupil dilation

Eye gaze and pupillary responses has been used extensively to measure attention. However, we are not aware of research on how emotions affect eye gaze while watching videos; therefore, the eye gaze itself has not been used for emotion recognition. The pupillary response is the measurement of pupil diameter over time. Pupil can dilate or constrict in response to illuminary, cognitive, attentional and emotional stimuli [69, 70].

Partala and Surakka [70] studied pupil size variation in three different emotional states; namely, positive activated, negative activated and neutral. In both positive and negative emotional conditions, pupil diameter was significantly larger than the neutral state. Bardley et al. [69] also recorded pupil diameter with heart rate and skin conductance in response to IAPS emotional pictures [56] and found significant pupil dilation in the presence of high arousal as well as unpleasant emotions.

Gao et al. [71] showed the significance of using pupillary reflex for stress assessment after reducing the light effect using a real-time feedback. They displayed words in different colors and asked participants to report displayed words' font color. According to "Stroop Color-Word Interference Test“, words with certain colors induce stress in the viewers.

2.4.2 Emotion recognition from audio-visual modalities

Most of research aiming at detecting emotions is based on the analysis of facial expressions, speech, behavioral attitudes such as posture which can be all measured by audio-visual modalities [72]. From these modalities, one of the most well studied emotional expression channels after audio is visual channel capturing facial expressions. A human being uses facial expressions as a natural mean of emotional communication. Emotional expressions are also used in human-human communication to clarify and stress what is said, to signal comprehension, disagreement, and intentions, in brief, to regulate interactions with the environment and other persons in the vicinity [73, 74]. Emotional facial expressions are often classified into six Ekman basic emotion using his action unit coding system [24]. Automatic facial expression methods are usually based on tracking multiple points, e.g., corners of lips, on a registered image of face [75].

Posture and gesture have been shown to carry valuable information related to affect [76, 77]. It has been shown how low level posture features such as orientation and distances between joints can discriminate between different emotional states [78, 76].

Audio and speech have been analyzed in two different levels [79]. The first level is explicit or linguistic level, which deals with the semantics and the words spoken. An affective dictionary was developed by Whissell which specifies each word dimensional affective scores [80]. The second level, acoustic and prosodic features, have been used to detect the emotion of speakers. Low level speech features such as Mel Frequency Cepstral Coefficients (MFCC) are often used to detect speakers' emotions in continuous as well as discrete emotional representations [81, 77, 82].

For further reading on audio-visual affect recognition methods, I refer the reader to the following surveys [83, 79, 72]

2.5 Implicit Tagging

Pantic and Vinciarelli define implicit tagging as using non-verbal behavior to find relevant keyword or tags for multimedia content [5]. Implicit tagging research has recently attracted researchers' attention, and number of studies have been published [10, 84, 85, 86]. To the best of my knowledge the following studies have been conducted in this direction.

Kierkels et al. [13] proposed a method for personalized affective tagging of multimedia using peripheral physiological signals. Valence and arousal levels of participants' emotion when watching videos were computed from physiological responses using linear regression [10]. Quantized arousal and valence levels for a clip were then mapped to emotion labels. This mapping enabled the retrieval of video clips based on keyword queries. So far this novel method achieved low precision.

Joho et al. [84, 85] developed a video summarization tool using facial expressions. A probabilistic emotion recognition based on facial expressions was employed to detect emotions of 10 participants watching eight video clips. The participants were asked to mark the highlights of the video with an annotation tool after the experiments. The expression change rate between different emotional expressions and the pronounce level of expressed emotions were used as features to detect personal highlights in the videos. The pronounce levels they used was ranging from highly expressive emotions, surprise and happiness, to no expression or neutral. They have also extracted two affective content-based features which were audio energy and visual change rate from videos to create an affective curve in the same way as the affective highlighting method proposed by Hanjalic [87].

Arapakis et al. [88] introduced a method to assess the topical relevance of videos in accordance to a given query using facial expressions showing users' satisfaction or dissatisfaction. Based on facial expressions recognition techniques, basic emotions were detected and compared with the ground truth. They were able to predict with 89% accuracy whether a video was indeed relevant to the query. In a more recent study, the feasibility of using affective responses derived from both facial expressions and physiological signals as implicit indicators of topical relevance was investigated. Although the results are above random level and support the feasibility of the approach, there is still room for improvement from the best obtained classification accuracy, 66%, on relevant versus non-relevant classification [89].

Yazdani et al. [86] proposed using a Brain Computer Interface (BCI) based on P300 evoked potentials to emotionally tag videos with one of the six Ekman basic emotions [42]. Their system was trained with 8 participants and then tested on 4 others. They achieved a high accuracy on selecting tags. However, in their proposed system, a BCI only replaces the interface for explicit expression of emotional tags, i.e. the method does not implicitly tag a multimedia item using the participant's behavioral and psycho-physiological responses.

2.6 Summary

This Chapter serves as a background review of the related research. I first gave a brief introduction to the definition of emotion and emotional experience scenarios in response to multimedia. Then, studies on multimedia content analysis with the goal of estimating the emotion which is more likely to be elicited to the viewers were given. The literature on emotion recognition using physiological signals was also briefly reviewed. Finally, the existing work on implicit tagging was provided.

Chapter 3

Affective Corpus Development

3.1 Video Corpora Development

Affective video corpora are developed with three different goals: first, emotion elicitation or mood regulation in psychological experiments; second, emotional characterization of videos using content for video indexing or highlighting and third, recognition of the intrinsic emotions in the videos, e.g., detecting the emotions which were expressed by people in the videos. One should avoid mixing these three different research tracks and the goals behind them. For example, movie excerpt which are more likely to elicit strong emotions are chosen in the first group for emotion elicitation whereas only using the strongly emotional excerpts is not appropriate for emotional characterization in the second group. Emotional characterization should be able to deal with the full spectrum of emotions in videos, from neutral videos to mixed and strong emotions. In this Section, the first two types of affective corpora will be addressed. The third type of affective video corpora which deals with intrinsic emotions or affective expression will be explained in the following Section, Section 3.2.

3.1.1 Existing emotional video corpora

Rottenberg et al. [9] created an emotional video dataset for psychological emotion elicitation studies. The excerpts, which were about 1 to 10 minutes long, were either extracted from famous commercial movies or from non-commercial videos which were used in emotional research, e.g., an amputation surgery video. First, they formed a set of excerpts with different targeted emotions; namely, amusement, anger, disgust, fear, neutral, sadness and surprise. They evaluated the excerpts based on “intensity” and “discreteness”. The “intensity” of an excerpt means whether a video received high mean report on the target emotion in comparison to other videos. The “discreteness” refers to what extent the target emotion was felt more intensely in comparison to all non-targeted emotions. The “discreteness” was measured using the ratings a video received on the target emotion in comparison to the other emotions. They ultimately formed a dataset consisting of 13 videos, from under one minute to 8 minutes long, for emotion elicitation studies.

In a more recent study, Schaefer et al. [44] created a larger dataset from movie excerpts to induce emotions. In their study, they went beyond discrete basic emotions and developed a corpus including 15 mixed feelings, in addition to six discrete emotions; namely, anger, disgust, sadness, fear, amusement, tenderness. 364 participants annotated their database using three questionnaires. After watching each video, participants answered emotional arousal on a seven points scale. Then using a modified version of Differential Emotions Scale (DES) questionnaire, they reported how much they felt each of the 16 listed emotions on a seven point scale. The third questionnaire was PANAS with 10 positive and 10 negative emotions on five points scale. 64 collected excerpts with French audio tracks are available online with their averaged assessed scores [1].

Almost every published work in the field of multimedia content analysis and emotions used a differently developed dataset. A brief description of different dataset used for emotional characterization using content features is given in the following.

Wang and Cheong [50] created and annotated a dataset consisting of 36 full length Hollywood movies which have 2040 scenes. Three annotators watched the movies and reported their emotions by Ekman basic emotions [42] to every scene. Only 14% of movie scenes received double labels, and the rest only received single emotional labels from their three annotators.

Hanjalic and Xu [7] used excerpts from “Saving private Ryan” and “Jurassic park 3” and two soccer matches in their study without annotations. Irie et al. [51] only used 206 selected emotional scenes out of 24 movies. 16 students annotated these scenes by eight Plutchik basic emotions; namely, joy, acceptance, fear, surprise, sadness, disgust, anger, and anticipation [32]. The annotators first watched the videos and then reported how much they felt each of these emotions on seven points scale. The emotional labels were assigned to the selected scenes only if more than 75% of annotators agreed on them, otherwise the neutral label was assigned to the movie scene. Xu et al. [52] used selected scenes from eight movies containing 6201 shots, which are in total 720 minutes long. The videos were manually labeled by five emotions: fear, anger, happiness, sadness and neutral spanning the arousal dimension in three levels and valence in two levels.

3.1.2 Developed emotional video corpora

Four emotional video corpora has been developed by us using four different settings. The annotations of the first and the fourth dataset were gathered a laboratory setting. The second and the fourth dataset have emotional reports from a web-based online platform, and the third dataset includes emotional reports using an online crowdsourcing platform.

3.1.2.1 Movie scenes annotated in a laboratory environment

A dataset consisting of emotional movie scenes suitable for emotion elicitation and characterization was developed. I chose 64 movie scenes from eight movies to be shown in two sessions, Due to the limited time a participant can spend in each session.

To create this video dataset, I extracted video scenes from movies selected either according to similar studies (e.g., [50, 9, 7]), or from recent famous movies. The movies included four major genres: drama, horror, action, and comedy. Video clips used for this study are extracted from the list given in Table 3.1. The extracted scenes, eight for each movie, had durations of approximately one to two minutes each and contained an emotional event (judged by the author).

[1] http://nemo.psp.ucl.ac.be/FilmStimuli

Table 3.1: The movie scenes were extracted from the listed movies

Drama movies	**Comedy movies**
The pianist (6), Hotel Rwanda (2)	Mr. Bean's holiday (5), Love actually (4)
Horror movies	**Action movies**
The ring (Japanese version) (7), 28 days later (1)	Kill Bill Vol. I (3), Saving private Ryan (8)

The complete list of the scenes with editing instructions and descriptions is available in Appendix 6.3.

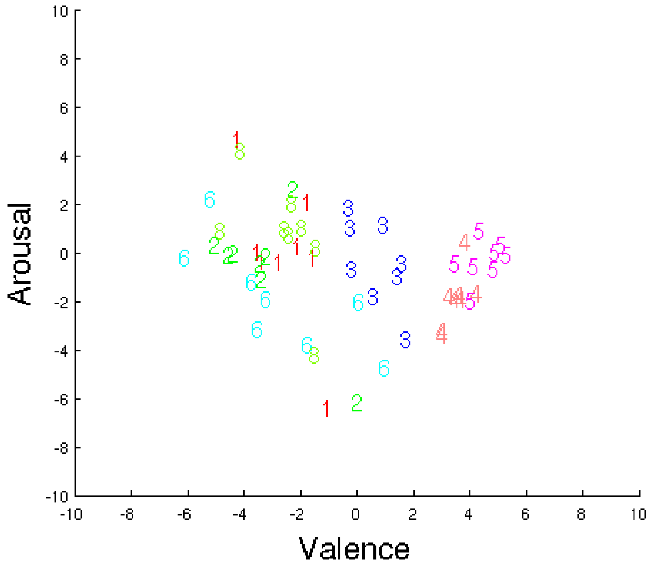


Figure 3.1: The distribution of different movie scenes on arousal and valence plane. Average arousal and valence are shown. Different numbers represent different movies (see Table 3.1).

The distribution of average arousal and valence scores are shown in Fig. 3.1. The numbers,which represent the movie scenes, are the codes assigned to movies in Table 3.1. The variance of valence dimension increases with arousal. This is in accordance with the findings of [90] in which arousal and valence scores in response to the IAPS and the International Affective Digitized Sound system (IADS) showed a parabolic or heart shape distribution.

Three participants out of ten were female. The participants were from 20 to 40 years old ($M = 29.3, SD = 5.4$). The difference between arousal and valence scores given by the participants to all the videos was studied by means of a multi-way ANOVA was performed on arousal and valence scores considering three factors, the video scenes, the participants and the order in which the videos were shown to the participants during sessions. The effect of the order in which the videos were presented to users on ratings was not significant. However, there was a significant difference on average valence scores between different participants ($F(9) = 18.53, p < 1 \times 10^{-5}$) and different videos ($F(63) = 12.17, p < 1 \times 10^{-5}$). There was also a significant difference on average arousal scores between different participants ($F(9) = 19.44, p < 1 \times 10^{-5}$) and different videos ($F(63) = 3.23, p < 1 \times 10^{-5}$). These differences can be originated from different personal experiences and memories concerning different movies as well as participants' mood and background.

3.1.2.2 Web-based annotated movie scenes dataset [2]

In order to find videos eliciting emotions from the whole spectrum of possible emotions, a user study was conducted to annotate a set of manually preselected movie scenes. The dataset has been extracted from 16 full length Hollywood movies which are listed in Table 3.2 (mostly popular movies). We extracted and chose these video scenes from movies in the same way as the video clips in the previous Section. 155 short clips, which are about one to two minutes long, were manually selected from these movies to form the dataset.

A web-based annotation system has been launched to assess participants' felt emotion. In this system, a user signs up giving his/her personal information including gender, age, and email address. The system also asked optional information like, cultural background and origin which helped the system to form a profile of the user. Fig. 3.3 shows a snapshot of the assessment interface where a video clip is being shown. After watching each video clip, the participant expressed his/her felt emotion using arousal and valence, quantized in nine levels. The participants

[2] Jeremy Davis contributed to the development of this dataset by implementing the web-based platform.

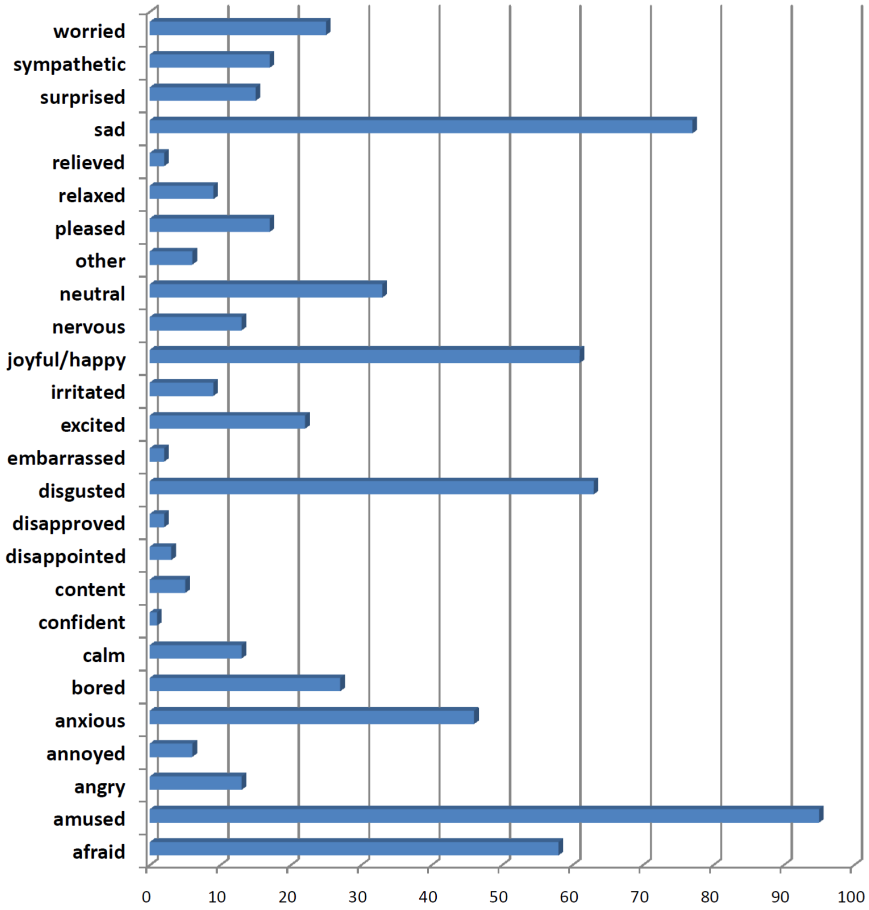
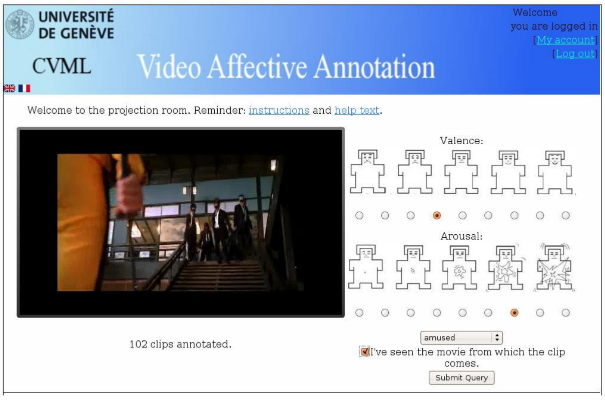


Figure 3.2: Total number of keywords reported by 10 participants to 64 video clips

Table 3.2: The video clips were extracted from the listed movies

Drama movies	**Comedy movies**
The pianist, Hotel Rwanda, Apocalypse now, American history X, Hannibal	Man on the moon, Mr. Bean's holiday, Love actually
Horror movies	**Action movies**
Silent hill, 28 days later, The shining	Kill Bill Vol. I, Kill Bill Vol. II, Platoon, The thin red line, Gangs of New York



Welcome to the projection room. Reminder: instructions and help text.

○ ○ ○ ● ○ ○ ○ ○ ○

○ ○ ○ ○ ○ ○ ● ○ ○

102 clips annotated.

☑ I've seen the movie from which the clip comes.

Figure 3.3: A snapshot of the affective annotation platform.

also chose the emotional label manifesting his/her felt emotion. The emotion labels are afraid, amused, anxious, disgusted, joyful, neutral, and sad. These labels have been chosen based on the labels assessed in our previous experiments (see subsection 3.1.2.1). During the previous experiments, the laboratory based experiment, we asked 10 participants to freely express their emotions, elicited by movie scenes with words. These emotional keywords were the ones which appeared more frequently [10] (see Fig. 3.2). Note that they roughly correspond to the six basic "Ekman's emotions" [42].

Initially, 82 participants signed up to annotate the videos. From these 82 participants, 42 participants annotated at least 10 clips. Participants were from 20 to 50 years old ($M = 26.9$, $SD = 6.1$). Out of the 42 participants, 27 were male, and 15 were female with different cultural backgrounds living in four different continents. The results of a multi-way ANOVA on arousal scores as the dependent variable and participant, video clip, and time of the day as effects showed that the average arousal scores have a significant difference for different participants ($F(41) = 3.23, p < 1 \times 10^{-5}$), video clips ($F(154) = 5.35, p < 1 \times 10^{-5}$) and times of the day ($F(7) = 2.69, p < 0.01$). A day has been divided into eight time interval, early morning (6:00 to 9:00), morning (9:00 to 11:30), noon (11:30 to 13:00), afternoon (13:00 to 16:30), evening (16:30 to 19:30), late evening (19:30 to 22:30), night (22:30 to 24:00) and after midnight (00:00 to 6:00).

The average arousal scores in different time periods are shown in Fig. 3.6. The average arousal scores given to all videos is increasing from early in the morning till noon. Then it decreases till it bounces back for late evening and night. Female participants, on average, gave higher arousal scores to the videos (see Fig. 3.4). A Wilcoxon test showed that the difference between female and male participants' arousal scores was significant ($p = 3 \times 10^{-16}$).

Studying the variance of valence is more tricky since the dataset roughly has a balanced set of pleasant and unpleasant videos. Looking at the average valence scores, no significant difference can be observed between the ratings given by different gender groups or in different time intervals. We therefore decided to compute the absolute valence score and study the effect of the absolute valence score after being centered. This means extremely pleasant or unpleasant scores will be treated in the same way, and we defined absolute centered valence score as a measure of emotion strength. Looking at this absolute score, female participants reported significantly stronger emotions (see Fig. 3.5). The result of a Wilcoxon test between the absolute valence scores of female and male participant showed the significance ($p = 0.002$). The results of a multi-way ANOVA on absolute valence scores as the variable and participant, video clip, and time of the day as effects showed that the average arousal scores have a significant difference for different participants ($F(41) = 4.76, p < 1 \times 10^{-5}$), video clips ($F(154) = 4.89, p < 1 \times 10^{-5}$) and times of the day ($F(7) = 2.68, p < 0.01$). The absolute valence, which is resulted by folding the valence arousal plane, is correlated with arousal.

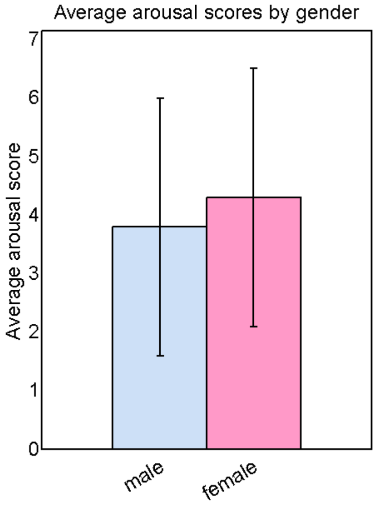


Figure 3.4: Average arousal scores, given by male and female participants.

3.1.2.3 Boredom detection dataset using crowdsourcing[3]

Developing video processing algorithms capable of predicting viewer boredom requires suitable corpora for development and testing. A video dataset has been gathered in the context of the MediaEval[4] 2010 Affect Task for boredom prediction of Internet videos. Standard limitations

[3] This dataset was developed in Collaboration with Martha Larson from Delft University of Technology, the Netherlands.

[4] http://www.multimediaeval.org

on viewer affective response annotation are overcome by making use of crowdsourcing. Using MTurk [5], we rapidly gathered self-reported boredom scores from a large user group that is demographically diverse also represented our target population (Internet video viewers). Ultimately, this dataset can be used by boredom-prediction algorithms to improve multimedia retrieval and recommendation. Relatively little research has investigated topic-independent factors that contribute to the relevance of multimedia content to the user information need.

For the purpose of the Affect Task and related research, a simple definition of boredom was adopted. Boredom was taken to be related to the viewer's sense of keeping focus of attention and to be related to the apparent passage of time [91]. Boredom is a negative feeling associated with viewer perceptions of the viewer-perceived quality (viewer appeal) of the video being low.

The dataset selected for the corpus is Bill's Travel Project, a travelogue series called "My Name is Bill" created by the film maker Bill Bowles [6] (see Fig. 3.8). The series consists of 126 videos between two to five minutes in length. This data was chosen since it represents the sort of multimedia content that has risen to prominence on the Internet. Bill's travelogue follows the format of a daily episode related to his activities and as such is comparable to "video journals" that are created by many video bloggers. The results of analysis on video series such as "Bill's Travel Project" can extend to other video bloggers, and also perhaps to other sorts of semi-professional user generated video content. Because the main goal of this study was to study the effect of content related features, by using one series, the effect of high variance between content generated by multiple film makers in different genres was avoided.

The design of the utilized crowdsourcing strategy was inspired by existing crowdsourcing literature, for example, [92], online articles and blog posts about crowdsourcing such as "Behind the enemy lines" blog [7], and reflecting on our past experience regarding collecting annotations online. A two-step approach was taken for our data collection. The first step was the pilot that consisted of a single micro-task or Human Intelligent Task (HIT) involving one video and

[5] http://www.mturk.com

[6] http://www.mynameisbill.com

[7] http://behind-the-enemy-lines.blogspot.com

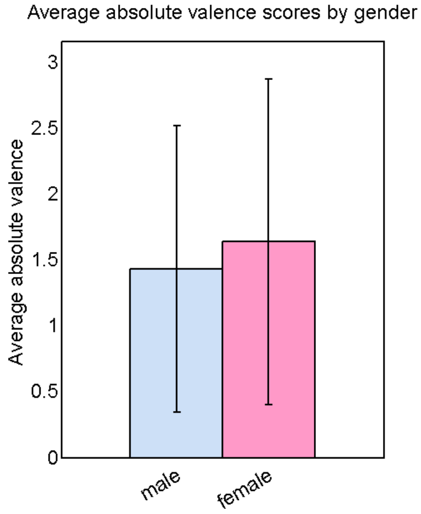


Figure 3.5: Average absolute valence scores given by male and female participants.

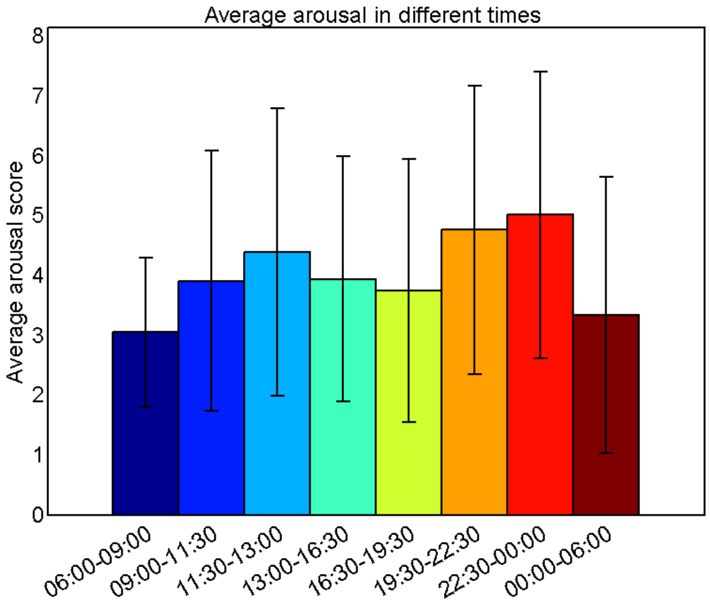


Figure 3.6: Average arousal scores in different times of the day.

would be used for the purpose of recruiting and screening MTurk users (referred to as “workers”). The second step was the main task and involved a series of 125 micro-tasks, one for each of the remaining videos in the collection. Workers were paid 30 US dollar cents for each accomplished HIT.

The pilot contained three components corresponding to qualities that were required from our recruits. The first section contained questions about the personal background (age, gender, cultural background). Using MTurk’s ability to block workers from certain countries, the geographical location of participants was partly limited from south Asian countries to maintain the

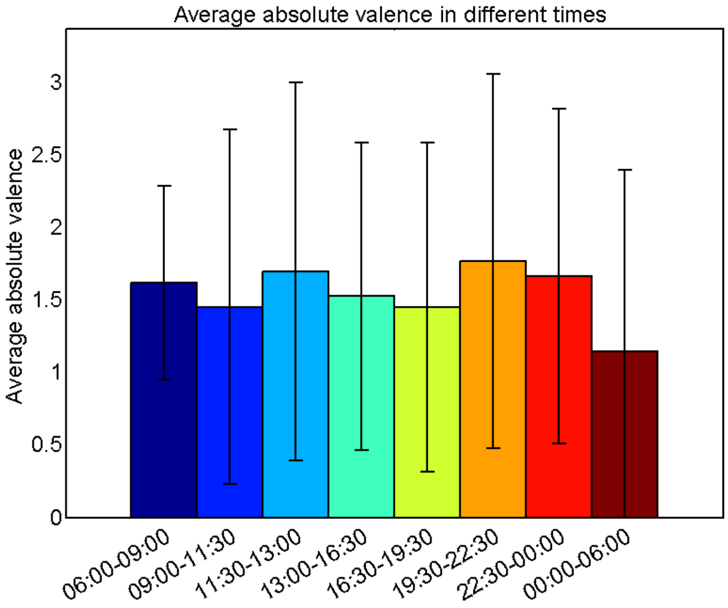


Figure 3.7: Average absolute valence scores in different times of the day.

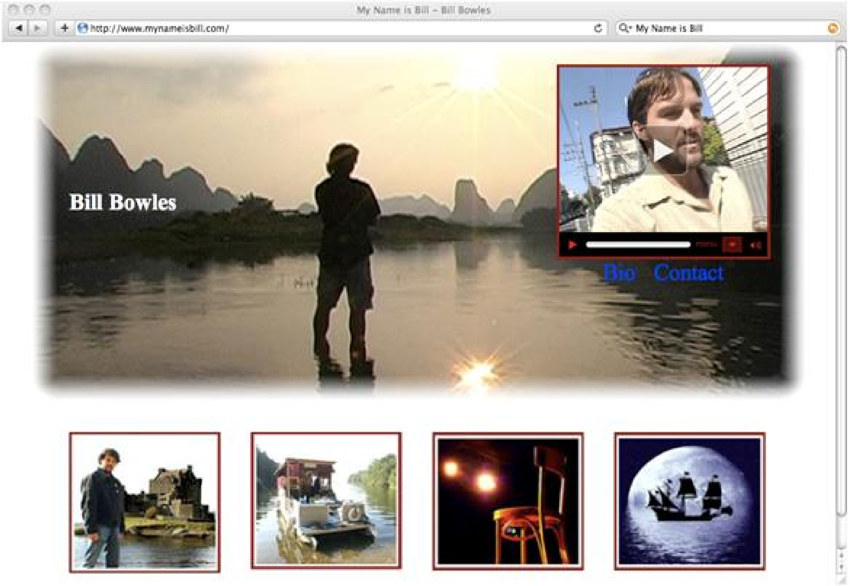
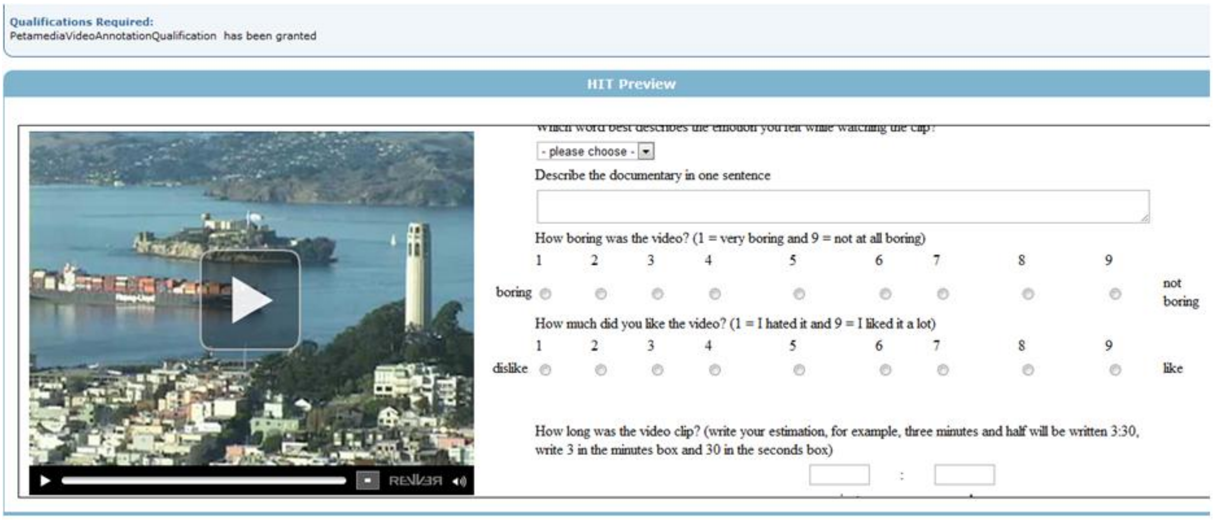


Figure 3.8: A snapshot of Bill Bowel's travel projects website.

overall balance. The second section contained questions about viewing habits: workers were asked whether they were regular viewers of Internet videos. The third section tested their seriousness by asking them to watch the video, select a word that reflected their mood at the moment and also write a summary. The summary constituted a "verifiable" question, recommended by [92]. The summary offered several possibilities for verification. Its length and whether it contained well-formulated sentences gave us an indication of the level of care that the worker devoted to the HIT. Also, the descriptive content indicated whether the worker had watched the entire video, or merely the beginning. A final question inquired if they were interested in performing further HITs of the same sort. In order to hide the main goal of the study from workers, the video description edit box was placed prominently in the HIT.

The workers were chosen for the main task from the participants of the pilot by considering the quality of their description and choosing a diverse group of respondents. The qualification was only granted to the participants who answered all the questions completely. The workers were invited to do the main study by sending them an invitation e-mail via their ID number on the MTurk platform. The e-mail informed the users that our qualification was granted to them. Use of a qualification served to limit those workers that carry out the HIT to the invited workers.

Each HIT in the main study consisted of three parts. In the first part, the workers were asked to specify the time of the day, which gave us a rough estimate of how tired they were. Also, the workers were asked to choose a mood word from a drop down list that best expressed their reaction to an imaginary word (none word), such as those used in [93]. The mood words were pleased, helpless, energetic, nervous, passive, relaxed, and aggressive. The answers to these questions gave us an estimate of their underlying mood. In the second part, they were asked to watch the video and give some simple responses to the following questions. They were asked to choose the word that best represented the emotion they felt while watching a video from a second list of emotion words in the drop down list. The emotion list contained Ekman six basic emotions [42]; namely, sadness, joy, anger, fear, surprise, and disgust, in addition to boredom, anxiety, neutral and amusement, which cover the entire affective space, as defined by the conventional dimensions of valence and arousal [36]. The emotion and mood word lists contained different items, which were intended to disassociate them for the user. Next, they were asked to provide a rating specifying how boring they found the video and how much they liked the video, both on a nine point scale. Then, they were asked to estimate how long the video lasted. Here, we had to rely on their full cooperation in order not to cheat and look at the video timeline. Finally, they were asked to describe the contents of the video in one sentence. We emphasized the description of the video rather than the mood word or the rating, in order to conceal the main purpose of the HIT. Quality control of the responses was carried out by checking the description of the video and also by ensuring that the time that they took to complete the HIT was reasonable. A snapshot of the main HIT is shown in Fig. 3.9.



Qualifications Required:
PetamediaVideoAnnotationQualification has been granted

HIT Preview

Which word best describes the emotion you felt while watching the clip?

- please choose -

Describe the documentary in one sentence

How boring was the video? (1 = very boring and 9 = not at all boring)

	1	2	3	4	5	6	7	8	9	
boring	○	○	○	○	○	○	○	○	○	not boring

How much did you like the video? (1 = I hated it and 9 = I liked it a lot)

	1	2	3	4	5	6	7	8	9	
dislike	○	○	○	○	○	○	○	○	○	like

How long was the video clip? (write your estimation, for example, three minutes and half will be written 3:30, write 3 in the minutes box and 30 in the seconds box)

:

Figure 3.9: A snapshot of the main HIT on MTurk.

Our pilot HIT was initially published for 100 workers and finished in the course of a single weekend. We re-published the HIT for more workers when we realized we needed more people in order to have an adequate number of task participants. Only workers with the HIT acceptance rate of 95% or higher were admitted to participate in the pilot HIT. In total, 169 workers completed our pilot HIT, 87.6% of which reported that they watch videos on the Internet. We took this response as confirmation that our tasks participants were close to the target audience of our research. Out of 169 workers, 105 were male, and 62 were female and two did not report their gender. Their age average was 30.48 with the standard deviation of 12.39. The workers in the pilot HITs identified themselves by different cultural backgrounds from North American, Caucasian to South and East Asian. Having such a group of participants with a high diversity in their cultural background would have been difficult without using the crowdsourcing platforms. Of the 169 pilot participants, 162 were interested in carrying out similar HITs. Out of the interested group, 79 workers were determined to be qualified and were assigned our task-specific qualification within MTurk. This means only 46.7% of the workers who did the pilot HIT were able to answer all the questions and had the profile we required for the main task.

In total, 32 workers have participated and also annotated more than 60 of the 125 videos in the main task HIT series. This means only 18.9% of the participants in the pilot and 39.0% of the qualified participants committed to doing the main task HIT series seriously. Of this group of 32 serious participants, 18 are male and 11 are female with ages ranging from 18 to 81 ($M = 34.9, SD = 14.7$).

To evaluate the quality of the annotations, the time spent for each HIT was compared to the video length. In 81.8% of the completed HITs, the working duration for each HIT was longer than the video length. This means that in 18.2% of the HITs we have doubts if the workers fully watched the videos. This shows the importance of having workers with the right qualifications and trustworthy pool of workers in annotation or evaluation hits. Rejecting those HITs reduced the number of workers who carried out more than 60 videos in the main series of HIT to 25 from which 17 are male and, 8 are female ages ranging from 19 to 59 ($M = 33.9, SD = 11.8$).

We asked three questions for each video to assess the level of boredom. First, how boring the video was on a nine-point scale from the most to the least boring. Second, how much the user liked the video on the nine-point scale and third how long the video was. Boredom was shown to have on average a strong negative correlation, $\rho = -0.86$ with liking scores. The time perception did not show a significant correlation for all users, and it varied from 0.4 down to -0.27. Although a positive correlation was expected from boredom scores, and the perception of time seven participants' boredom scores have negative correlations with the time perception.

The correlation between the order of watching the videos for each participant and the boredom ratings was also examined. No positive linear correlation was found between the order and boredom score. This means that watching more videos did not increase the level of boredom and in contrary for 2 of participant it decreased their boredom. Additionally, the correlation between the video length and boredom scores was investigated. No positive correlation was found between the boredom scores and videos' duration. We can conclude that the lengthy videos are not necessarily perceived as more boring than the shorter videos.

To measure the inter-annotator agreement, the Spearman correlation between participants' pairwise boredom scores was computed. The average significant correlation coefficient was very low, $\rho = 0.05$. There were even cases where the correlation coefficients were negative, which shows complete disagreement between participants. For each worker, we then grouped videos into two rough categories, above and below the mean boredom score of that worker. We computed the average pair-wise Cohen's kappa for these categories and here found only slight agreement ($\kappa = 0.01 \pm 0.04$). We also compared agreement on the emotion words workers associated with viewers. Here, again Cohen's kappa indicated only slight agreement ($\kappa = 0.05 \pm 0.06$). The strong correlations suggest that it is indeed beneficial to investigate personalized approaches to affective response prediction.

In order to obtain the best estimation of mood out of the mood words, first the responses of each participant were clustered into the three hours time intervals. In each three hours interval, the most frequent chosen mood word was selected as the dominant mood. After calculating the dominant moods, we found that using the implicit mood assessment none of the participants had the "relaxed" mood as her dominant mood.

The average boredom scores in different moods are shown in Fig. 3.10. The boredom scores were on average lower for passive mood and higher in energetic, nervous and pleased. Moods were categorized into two groups, positive, such as pleased, energetic, relaxed and negative, such as, helpless, nervous, passive, and aggressive. On average, participants gave higher ratings to videos while they were in positive moods (see Fig. 3.11). The statistical significance of the difference between ratings in positive and negative moods was examined by a Wilcoxon test and was found significant ($p = 4 \times 10^{-8}$). The effect of four different factors on boredom scores was investigated with a four way ANOVA. The effects were mood, time of the day, videos and participants. The effect of the time of the day on boredom scores was not significant. Participants' mood had a significant effect on the ratings ($F(6) = 5.55, p < 1 \times 10^{-4}$). The interaction between every two factors was investigated to check whether the observed difference was as a result of having special videos for every mood. The interaction in two way ANOVA between the videos and moods was not significant. Therefore, the effect of mood on boredom scores was independent of the effect of videos.

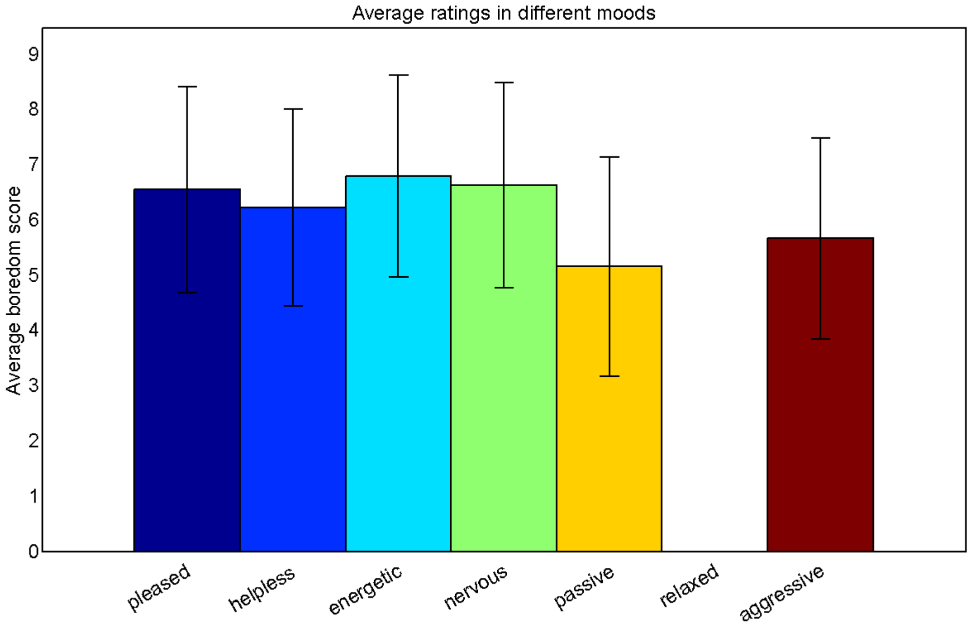


Figure 3.10: Average boredom scores in different moods.

3.1.2.4 Music videos dataset [8]

120 music videos were initially selected with the goal of having videos with emotions that are uniformly covering the arousal-valence space. 60 of these were selected manually and 60 were selected using the last.fm website for music recommendation by searching on a list of emotional keywords. The music videos were then segmented into one minute segments with 55 seconds overlap between segments. Arousal and valence of the one minute long segments were computed using the method proposed by Soleymani et al. [19] which is trained on movie scenes. In this method, a linear regression was used to compute arousal for each shot in movies. Informative features for arousal estimation include loudness and energy of the audio signals, motion component, visual excitement and shot duration. The same approach was used to compute valence. Other

[8] This database was developed in collaboration with, Sander Koelstra, Christian Mühl, Ashkan Yazdani, and Jong-Seok Lee in the context of Petamedia European network of excellence.

content features such as color variance and key lighting that have been shown to be correlated with valence [50] were utilized for valence estimation. The emotional highlight score of the i-th segment e_i was computed using the following equation:

$$e_i = \sqrt{a_i^2 + v_i^2} \tag{3.1}$$

The arousal, a_i, and valence, v_i, ranged between -10 and 10. Therefore, a smaller emotional highlight score (e_i) is closer to the neutral state. For each video, the one minute long segment with the highest emotional highlight score was chosen to be extracted for the experiment. For a few clips, the automatic affective highlight detection was manually overridden. This was done only for songs with segments that are particularly characteristic of the song, well-known to the public, and most likely to elicit emotional reactions. Given the 120 one-minute music video segments, the final selection of 40 videos used in the experiment was made on the basis of subjective ratings. Each video was rated by 14-16 volunteers using an online self-assessment tool. Valence and arousal was rated on a 9-point discrete scale. To maximize the strength of elicited emotions, we selected those videos that had the strongest volunteer ratings and at the same time a small variation. For each video, we calculated a normalized arousal and valence score by taking the mean rating divided by the standard deviation. Then, for each quadrant in the normalized valence-arousal space, we selected the 10 videos that lay the closest to the extreme corner of the quadrant.

After watching each video, participants reported their emotion by means of continuous ratings ranging from 1 to 9. Although they were able to choose any point on a continuous scale participants tended to click under displayed numbers (see the red bars on Fig. 3.12). The blue bars on Fig. 3.12 show the ratings' histograms quantized in nine levels. From the blue bars, we can see that the distribution of the ratings is skewed towards higher scores.

The average ratings of the videos are shown in Fig. 3.13. According to the average ratings, the videos are well covering the whole arousal and valence plane on four quadrants; namely, low arousal, high valence (LAHV), low arousal, low valence (LALV), high arousal, low valence

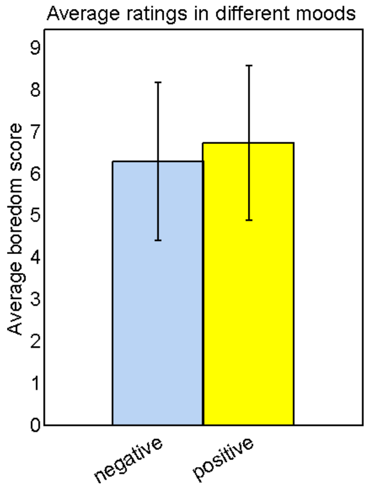


Figure 3.11: Average boredom scores in positive and negative moods.

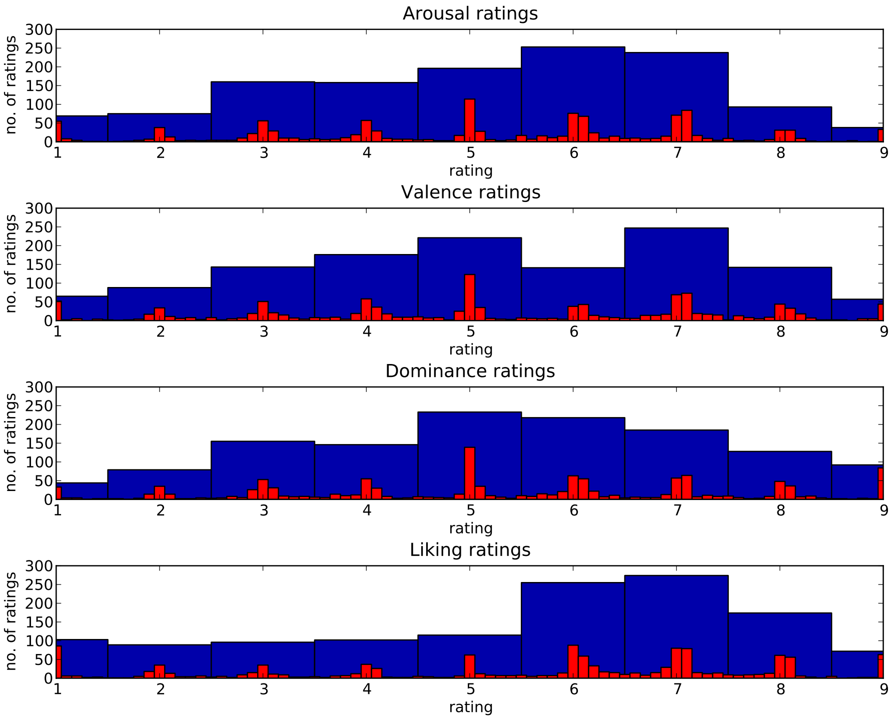


Figure 3.12: Histogram of arousal, valence, dominance and liking ratings given to all videos by the 32 participants. The blue bars are the histogram of the quantized ratings in nine levels. The red bars are showing the ratings quantized in 80 levels (the quantization step is equal to 0.1). This figure was originally created by Sander Koelstra and published in [2].

(HALV) and high arousal, high valence (HAHV). The orientation of the triangles represents the emotional quadrant represents the video clips' quadrants based on the preliminary study on a limited number of volunteers using an online platform. The results show that the expected emotions are in strong agreement with reported emotions (i.e. online volunteers usually place the video in the same quadrant as participants in the experiment). The average ratings for dominance ratings are also visible in Fig. 3.13. The liking ratings, which are encoded in colors, are visually shown to be correlated with valence.

In order to measure inter-annotation agreement between different participants, we computed the pair-wise Cohen's kappa between self reports after quantizing the ratings into nine levels. A very weak agreement was found on emotional feedbacks with $\kappa = 0.02 \pm 0.06$ for arousal, $\kappa = 0.08 \pm 0.08$ for valence, and $\kappa = 0.05 \pm 0.08$ for liking ratings. A paired t-test was performed on the κ values of valence ratings in comparison to liking and arousal. The t-test results showed that, on average, the agreement on valence ratings is significantly higher than agreement on arousal ($p = 2.0 \times 10^{-20}$) and liking rating ($p = 4.5 \times 10^{-7}$).

3.2 Affective Expression Databases

Creating affective databases is the first step in any emotion recognition study. Recent advances of emotion recognition studies have created novel databases containing emotional expressions in different modalities. These databases mostly includes speech, visual, or audio-visual data [94, 49, 95, 96, 97]. The visual modality of the emotional databases includes face- and/or body

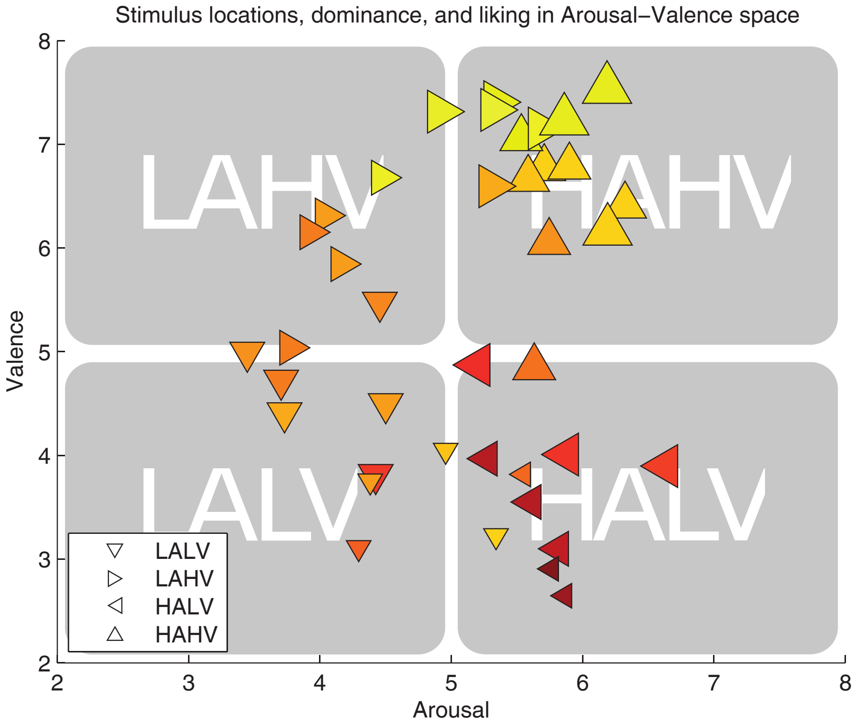


Figure 3.13: The mean video ratings from experiment participants on the arousal-valence plane. The orientation of the triangle indicates the quadrant for which the video was selected by the online ratings. Liking is encoded by color: dark red is low liking and bright yellow is high liking. Dominance is encoded by symbol size: small symbols stand for low dominance and big for high dominance. This figure was originally created by Christian Mühl and published in [3].

gestures. The audio modality carries acted or genuine emotional speech in different languages. In the last decade, most of the databases consisted only of acted or deliberately expressed emotions. This is nevertheless changing by more recent trend of sharing spontaneous and natural emotional databases such as in [95, 97, 96, 3]. We here only review the publicly available spontaneous or naturalistic databases and refer the reader for posed, audio and audio-visual databases to the following reviews [98, 72, 94].

3.2.1 Existing databases

One of the notable databases with spontaneous reactions is the Belfast database (BE) created by Cowie et al. [49]. The BE database includes spontaneous reactions in TV talk shows. Although the database is very rich in body gestures and facial expressions, the variety in the background and quality makes the data a challenging dataset of automated emotion recognition. The usual databases are recorded with a constant background (blue or green curtain) with a fixed camera and lighting. Despite the existence of a large number of studies on emotional expressions and their databases, there are only few open and publicly available databases which include bodily responses, in addition to audio-visual data. BE database was later included in a much larger ensemble of databases of the HUMAINE projects [95]. HUMAINE database consists of three naturalistic databases and six induced reaction databases. Databases vary in size from 8 to 125 participants and in modalities, from only audio-visual to peripheral physiological signals. These databases were developed independently in different sites and collected under HUMAINE project.

Pantic et al. created a web-based emotional database of posed and spontaneous facial expressions with both static images, and videos [94, 99]. Their database, called MMI database, consists of images and videos captured from both frontal and profile view. The MMI database includes data from 61 adults acting different basic emotions and 25 adults reacting to emotional videos. This web-based database gives an option of searching in the corpus and is downloadable on the Internet [9].

The "Vera am Mittag" (VAM) audio-visual database [96] is another example of developing a database using spontaneous naturalistic reactions during a talk show. 12 hours of audio-visual recordings from a German talk show with the same name was segmented and annotated. The segments were annotated using valence, activation and dominance dimensions. The audio-visual signals consist of the video and utterances recorded from 104 different speakers. In addition to the audio-visual database, an audio database and a video database were also developed. The VAM-Audio includes utterances from 19 speakers. Only high quality and low noise segments of the emotional audio content were selected for VAM-Audio. In VAM-Faces, the segments in videos which included faces were separately segmented and annotated. The VAM-Faces includes 1867 images from 20 speakers.

Comparing to audio-visual databases, there are fewer publicly available affective, physiological databases. Healey recorded one of the first affective, physiological datasets at Massachusetts Institute of Technology (MIT), which has reactions of 17 drivers under different levels of stress [59, 65]. She recorded 24 subjects driving around Boston and annotated the dataset by the drivers' stress level. Out of these 24 subjects' responses, 17 are available on the publicly available database. Her recordings include ECG, GSR - recorded from hands and feet -, EMG - from right Trapezius - as well as the respiration pattern. The database of stress recognition in drivers is publicly available on Internet from Physionet [10].

One of the few databases which includes both peripheral and central nervous system physiological responses and facial expressions, is the enterface 2005 emotional database recorded by Savran et al. [100]. This database includes two sets. The first one has EEG, peripheral physiological signals, functional Near InfraRed Spectroscopy (fNIRS) from 5 male subjects. Second dataset only has fNIRS and facial videos from 16 subjects from both genders. The first database recorded spontaneous responses to the emotional images from the IAPS [56] and the second database was recorded in a selected set of images retrieved from arbitrary sources on the internet.

The characteristics of the reviewed databases are summarized in Table 3.3.

[9] http://www.mmifacedb.com/

[10] http://www.physionet.org/pn3/drivedb/

Table 3.3: The summary of the characteristics of the reviewed emotional databases. The last two columns are the developed databases.

Database	**No Part.**	**Posed or Spon.**	**Induced or Natural**	**Audio**	**Visual**	**Peripheral physio.**	**EEG**	**Eye gaze**
MIT [59]	17	Spon.	Natural	No	No	Yes	No	No
eNTERFACE05 [100]	5,16	Spon.	Induced	No	Yes	Yes	Yes	No
MMI [94]	61,29	Posed & spon.	Induced	No	Yes	No	No	No
HUMAINE [95]	Multiple	Spon.	Both	Yes	Yes	Yes	No	No
VAM [96]	19	Spon.	Natural	Yes	Yes	No	No	No
SEMAINE [97]	20	Spon.	Induced	Yes	Yes	No	No	No
DEAP [3]	32	Spon.	Induced	No	Yes*	Yes	Yes	No
MAHNOB-HCI [20]	30	Spon.	Induced	Yes	Yes	Yes	Yes	Yes

*Only for 22 participants.

Spon.: spontaneous, Part.: participant

3.2.2 Developed emotional response corpora

I contributed to the recordings in the three experiments which were conducted to record participants' emotional reactions to videos. Physiological responses and facial expressions (only in the second and third experiments) were recorded while participants were watching video clips extracted from movies or online repositories. The nature of the video clips varied from scenes from famous commercially produced movies (the first experiment) to the music clips (the third experiment). The goal of these recordings was to study the correlations and the intensity of emotional reactions by viewers in response to videos. The experimental setup, apparatus and recorded modalities for each experiment is provided in detail.

3.2.2.1 Movie experiment

The video scenes selected in the Section 3.1.2.1 were used to elicit emotions in the same experiment. Peripheral signals and facial expression EMG signals were recorded for emotion assessment. EMG signals from the right Zygomaticus major muscle (smile, laughter) and right Frontalis muscle (attention, surprise) were used as indicators of facial expressions (see Fig. 3.14). GSR, skin temperature, breathing pattern (using a respiration belt) and BVP (using a plethysmograph) were also recorded. All physiological data was acquired via a Biosemi Active II system with active electrodes, from Biosemi Systems[11]. The data were recorded with a sampling frequency of 1024 Hz in a sound-isolated Faraday cage. Examples of recorded physiological signals in a surprising and funny scene are given in Figures 3.16 and 3.15. The GSR and respiration signals were respectively smoothed by a 512 and a 256 points averaging filters to reduce the high frequency noise. EMG signals were filtered by a Butterworth band pass filter with a lower cutoff frequency of 4 Hz and a higher cutoff frequency of 40 Hz.

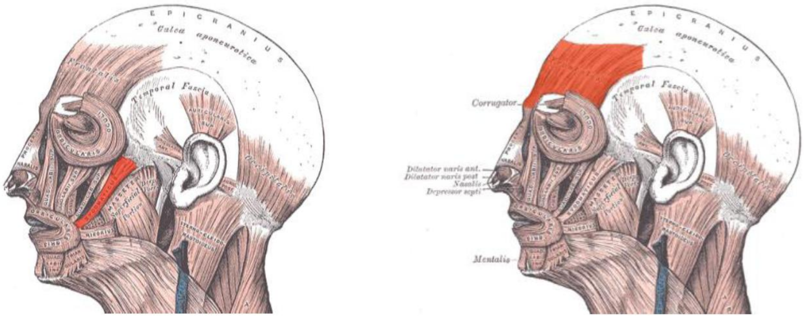


Figure 3.14: facial EMG signals were recorded from Zygomaticus major (left), and Frontalis muscles (right). These images are retrieved from Wikimedia.

To reduce the mental load of the participants, the protocol divided the show into 2 sessions of 32 movie scenes each. Each of these sessions lasted approximately two hours, including setup. Eight healthy participants (three female and five male, from 22 to 40 years old) participated in the experiment. Thus, after finishing the experiment three types of affective information about each movie clip were available:

- multimedia content-based information extracted from audio and video signals;

[11] http://www.biosemi.com

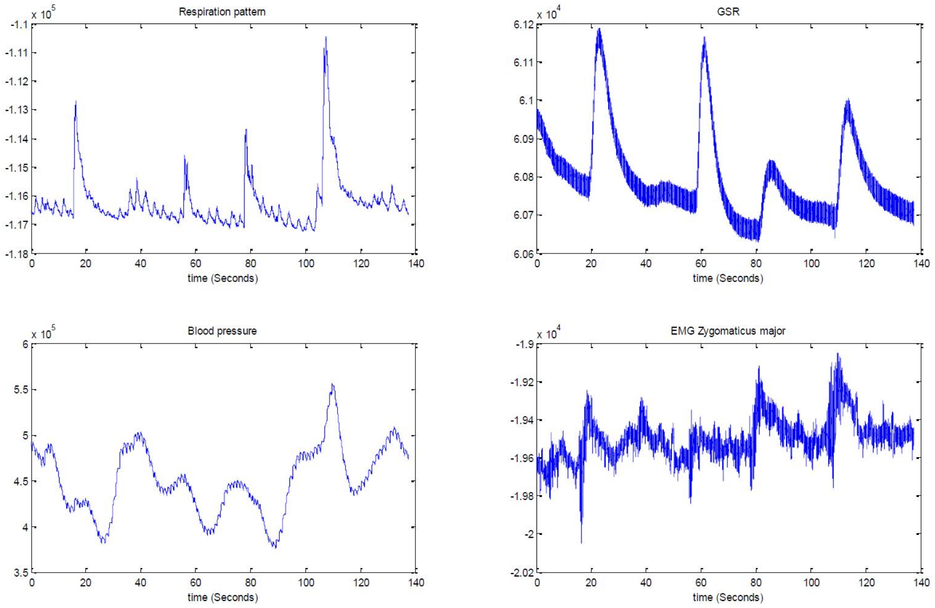


Figure 3.15: Participant 1 physiological responses to a comedy scene: respiration pattern (top-left), GSR (top-right), blood pressure (bottom left), and Zygomaticus major EMG (bottom-right). The effect of the spectator's laugh can be seen on the respiration pattern and EMG signal.

- physiological responses from spectators' bodily reactions (due to the autonomous nervous system) and facial expressions;
- self-assessed arousal and valence, used as 'ground truth' for the true feelings of the spectator.

Since video scenes were showed in random order, the average valence-arousal values over participants in the self-assessed vectors (64 elements each) do not depend on the order in which

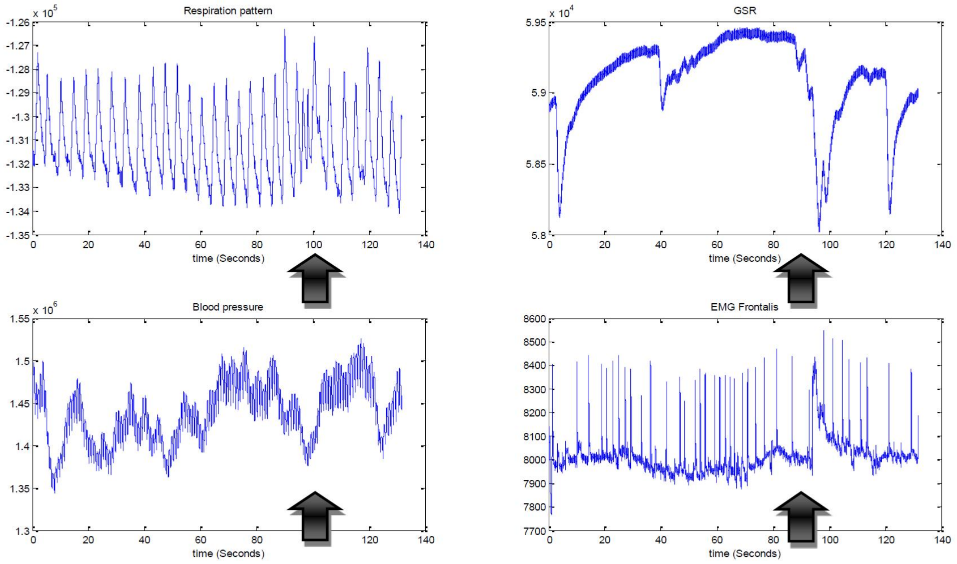


Figure 3.16: Participant 2 physiological responses to a surprising action scene: respiration pattern (top-left), GSR (top-right), blood pressure (bottom left), and Frontalis EMG (bottom-right). The surprise moment is indicated by an arrow.

scenes were presented.

The participants were first informed about the experiment, the meaning of arousal and valence, the self-assessment procedure, and the video content. In emotional-affective experiments, the bias of the emotional state (participants' mood) needs to be removed. To allow leveling of feature values over time a baseline is recorded at each trial start by showing one short 30s. neutral clip randomly selected from clips provided by the Stanford psychophysiology laboratory [9].

Each trial started with the user pressing the "I am ready" key which started the neutral clip playing. After watching the neutral clip, one of the movie scenes was played. After watching the movie scene, the participant filled in the self-assessment form which popped up automatically. In total, the time interval between the starts of consecutive trials was approximately three to four minutes. This interval included playing the neutral clip, playing the selected scene, performing the self-assessment, and the participant-controlled rest time. In the self-assessment step for evaluating arousal and valence, the SAM Manikin pictures with a slider to facilitate self-assessment of valence and arousal were used (see Fig. 3.17). The sliders correspond to a numerical range of [0, 1] while the numerical scale was not shown to the participants.

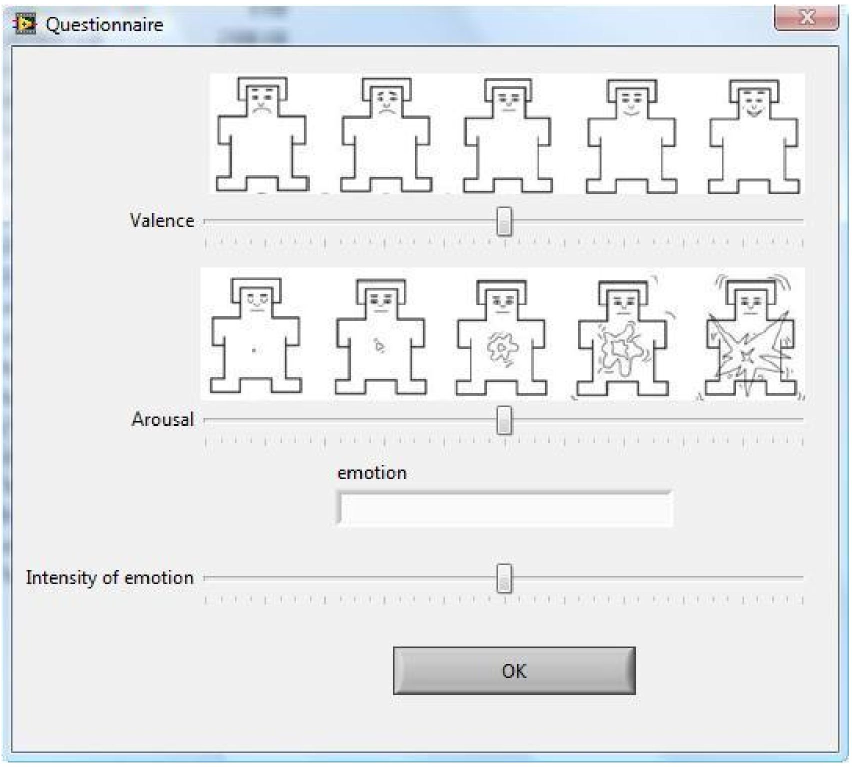


Figure 3.17: Arousal and valence self-assessment: SAM manikins and sliders.

3.2.2.2 Multimodal movie experiment (MAHNOB-HCI) [12]

Multi-modal recording setup was arranged to record facial videos, audio and vocal expressions, eye gaze, and physiological signals simultaneously (see Fig. 3.20). The experiment was

[12] This database is developed in collaboration with Maja Pantic, Imperial College London and available online at http://mahnob-db.eu/hct-tagging

controlled by the Tobii studio software (http://www.tobii.com). In order to synchronize different modalities, device generated time stamps were recorded along with audio and physiological signals. These time stamps consist of time series with square shaped periodic signal (60Hz) representing the moments when the cameras' shutters were open to capture each frame. The synchronization method and hardware setup details are given in Lichtenauer et al. [101].

The Biosemi active II system[13] with active electrodes was used for physiological signals acquisition. Physiological signals including ECG, EEG (32 channels), respiration amplitude, and skin temperature were recorded while the videos were shown to the participants.

Electroencephalogram signals were recorded with a 1024Hz sampling rate and later downsampled to 256Hz to reduce the memory and processing costs. EEG signals were recorded using active AgCl electrodes placed according to the international 10-20 system. The layout of EEG electrodes on the cap is shown in Fig. 3.18. Motion and moscular artifacts were kept at a minimum level by instructing the participants to minimize their movements while videos were playing. The EEG signals recorded by Biosemi active electrodes are recorded referenced to Common Mode Sense (CMS) electrode as a part of its feedback loop. In order to gain the full Common-Mode Rejection Ratio (CMRR) at 50Hz, EEG signals should be re-referenced to a common reference. EEG signals were thus re-referenced to the average reference to increase signal to noise ratio.

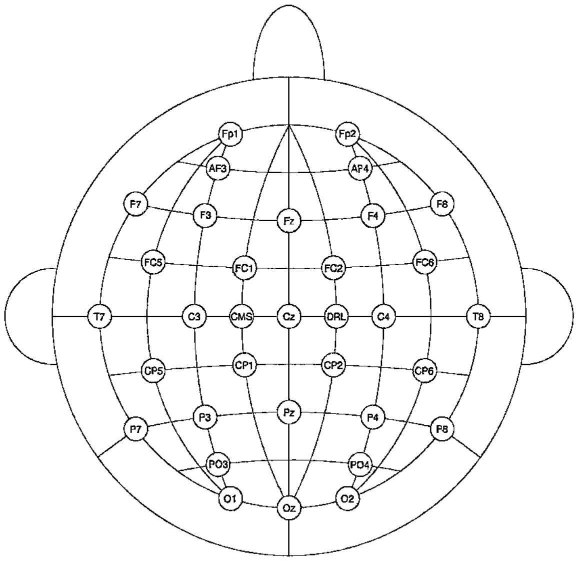


Figure 3.18: The EEG cap layout for 32 EEG in addition to two reference electrodes. Retrieved from Biosemi website (http://www.biosemi.com).

30 participants with different cultural and education backgrounds volunteered to participate in response to a campus wide call for volunteers at Imperial College, London. Out of the 30 young healthy adult participants, 17 were female and 13 were male; ages varied between 19 to 40 years old (M = 26.06, SD = 4.39). Participants had different educational background from undergraduate students to post-docs with different English proficiency from intermediate to native speakers.

[13] http://www.biosemi.com

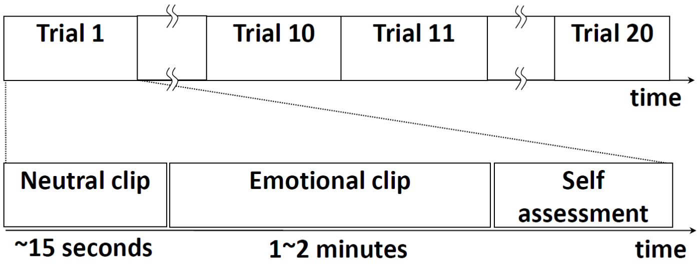


Figure 3.19: Each trial started by a 15s neutral clip and continued by playing one emotional clip. The self-assessment was done at the end of each trial. There were 20 trials in each session of experiment.

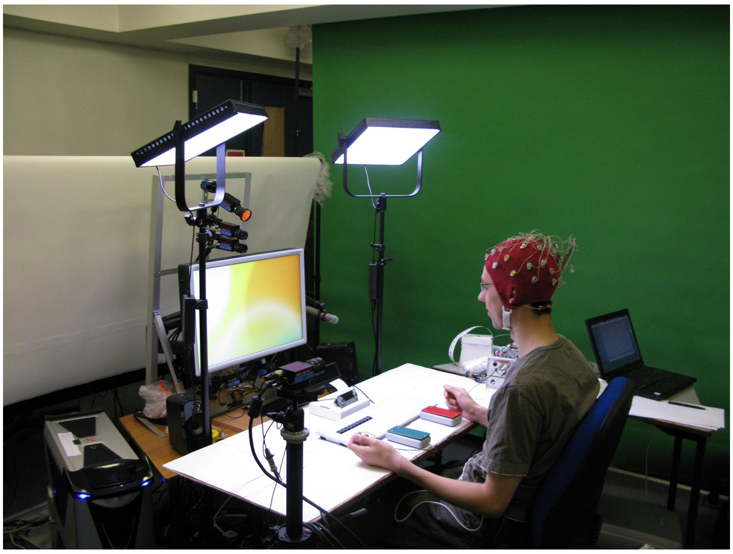


Figure 3.20: In the experimental setup, six video cameras were recording facial expressions. The modified keyboard is visible in front of the participant.

The participants were informed about the experiment and their rights with a verbal introduction, by email and through a consent form. Participants were trained about the interface before the experiment and during the setup time. The participants were also introduced to the meaning of arousal, valence, dominance and predictability in the self-assessment procedure, and to the nature of the video content.

To reduce the emotional bias, before each emotional video a short, neutral clip randomly selected from the clips provided by the Stanford psychophysiology laboratory [9] was shown to the participants.

After watching a short, neutral clip, one of the 20 video clips was played. Video clips were played from the dataset in random order. After watching the video clip, the participant filled in the self-assessment form which appeared automatically. Five multiple choice questions were asked during the self-report for each video. The five questions were 1. emotional label/tag; 2. arousal level; 3. valence level; 4. dominance level; 5. predictability level. These questions were chosen based on the emotional dimensions suggested by [37]. The emotional labels included neutral, anxiety, amusement, in addition to the Ekman six basic emotions; namely, sadness, joy, disgust, anger, surprise, fear. To simplify the interface, a keyboard was provided with only nine numerical keys and the participant could answer each question by pressing one of the nine keys. Questions 2 to 5 were on a nine points scale. Dominance and predictability responses were not used in this thesis. In total, the time interval between the start of a trial and the end of the self-reporting phase was approximately two and half minutes. This interval included playing the neutral clip, playing the emotional clip, performing the self-assessment. Running of the whole protocol took on average 50 minutes, in addition to 30 minutes setup time (see Fig. 3.19).

Automatic analysis of facial expression is an interesting topic from both scientific and practical point of view. It attracted the interest of many researchers since such systems will have numerous applications in behavioral science, medicine, security, and human-computer interaction. To develop and evaluate such applications, large collections of training and test data are needed [102, 103]. In this database, we were interested in studying the spontaneous responses of participants while watching video clips. This can be used later for emotional implicit tagging of multimedia content.

Fig. 3.22 shows the synchronized views from the six different cameras. The two close up cameras above the screen give a near-frontal view of the face in color 3.22(a) or monochrome

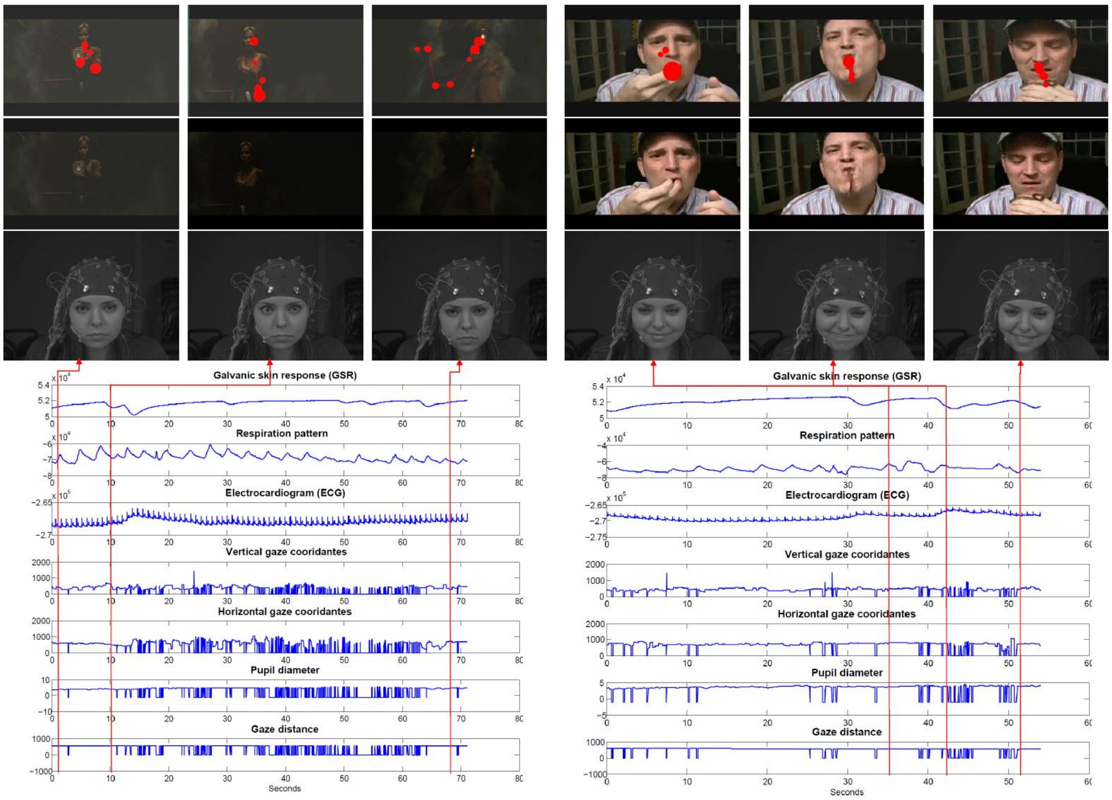
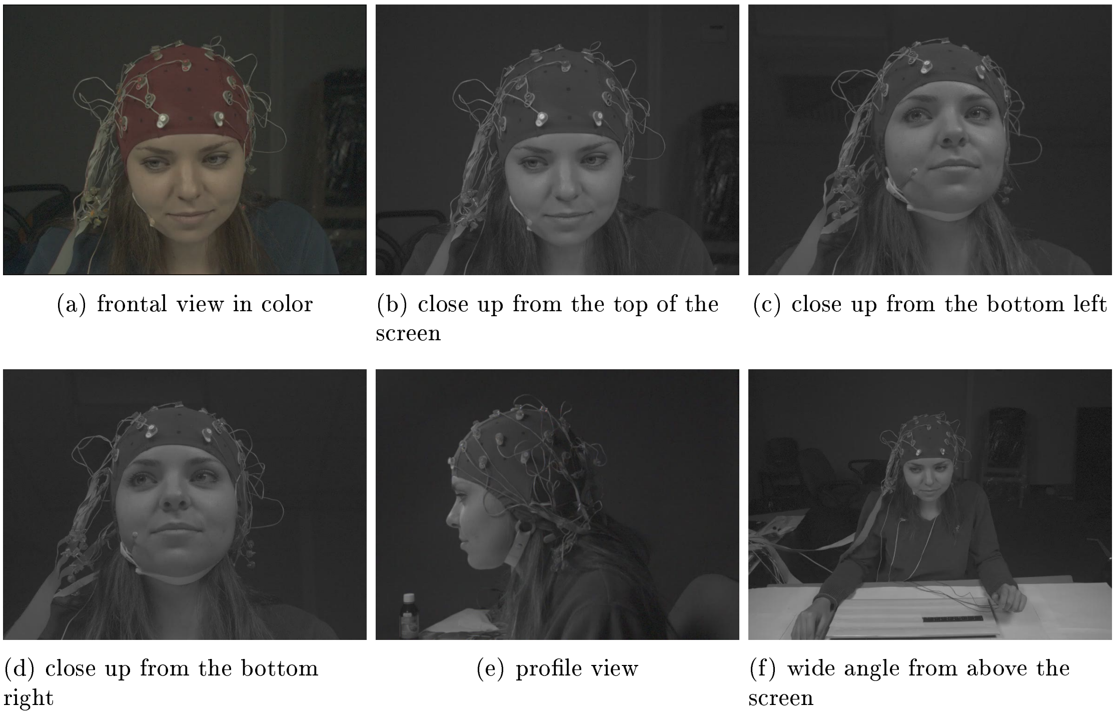


Figure 3.21: Two examples of natural expressions to a fearful (on the left) and disgusting (on the right) video. The snapshots of the stimuli videos with eye gaze overlaid and without eye gaze overlaid, frontally captured video, raw physiological signals and raw eye gaze data are shown. On the first row, the red circles show the fixation points and their radius indicate the time spent in each fixation point. The red lines indicate the moments were each of the snapshots were captured. The eye gaze tracker signals have the value -1 when eye gaze data is not available (e.g., during blinking moments).



(a) frontal view in color (b) close up from the top of the screen (c) close up from the bottom left

(d) close up from the bottom right (e) profile view (f) wide angle from above the screen

Figure 3.22: Snapshots of videos captured from 6 cameras recording facial expressions and head pose.

3.22(b). The monochrome views have a better sharpness and less motion blur than the color camera. The two views from the bottom of the screen, 3.22(c) and 3.22(d), give a close up view that may be more useful during down-facing head poses, and make it possible to apply passive stereo imaging. For this, the intrinsic and extrinsic parameters of all cameras have been calibrated. Linear polarizing filters were applied with the two bottom cameras, in order to reduce the reflection of the computer screen in eyes and glasses. The profile view 3.22(e) can be used to extract backward-forward head/body movements, or to aid the extraction of facial expressions, together with the other cameras. The wide-angle view 3.22(f) captures the upper body, arms and hands, which can also carry valuable information about a person's affective state.

Although we did not explicitly asked the participant to express or talk during the experiments, there exist some natural utterances and laughters in the recorded audio signals. Petridis and Pantic showed that it is possible to tag videos by the level of hilarity using viewer's laughter [104]. Different types of laughter can be an indicator of the level of hilarity of multimedia content.

3.2.2.3 Music videos experiment (DEAP) [14]

A set of experiments were conducted to record participants' responses to music videos. The goal of the experiments were to use the collected data to train classifiers to be used for an emotionally aware music recommendations system. The experiments were performed in the laboratory environment with controlled illumination. EEG and peripheral physiological signals were

[14] The database for emotion analysis using physiological signals (DEAP) was developed in collaboration with, Sander Koelstra, Christian Mühl, Ashkan Yazdani, and Jong-Seok Lee in the context of Petamedia European network of excellence.

recorded using a Biosemi Active II system[15]. The "Presentation" software by Neuro-behavioral systems[16] was used for running the protocol, playing the videos and recording the self-reports. The computer which was running the "Presentation" software sent hardware triggers via parallel port to the Biosemi system. These triggers were later used for synchronization and segmentation of the signals.

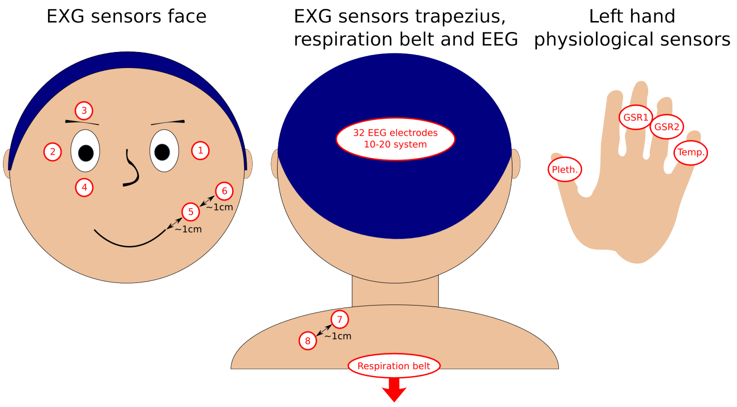

EXG sensors face
EXG sensors trapezius, respiration belt and EEG
Left hand physiological sensors
3
2
1
4
6
5
~1cm
~1cm
32 EEG electrodes 10-20 system
7
8
~1cm
Respiration belt
GSR1
GSR2
Temp.
Pleth.

Figure 3.23: Placement of peripheral physiological sensors. Four electrodes were used to record EOG and 4 for EMG (Zygomaticus major and Trapezius muscles). In addition, GSR, BVP, temperature and respiration were measured. This figure was originally created by Christian Mühl and published in [2].

Physiological signals were all recorded at a sampling rate of 512 Hz using active AgCl electrodes. 32 EEG electrodes were placed on participants' scalp according to the international 10-20 system. The following peripheral nervous system signals were acquired: GSR, respiration amplitude, skin temperature, ECG, BVP by plethysmograph, EMG of Zygomaticus and Trapezius muscles, and electrooculogram (EOG).

32 Healthy participants (16 male and 16 female), aged between 19 and 37 ($M = 27.2, SD = 4.4$), volunteered to participate in the experiment for a small compensation. Each participant signed a consent form and filled in a questionnaire before the experiment. They were then given a set of instructions about the protocol and the meaning of the different dimensions they were going to report during emotional self-assessment. Next, the electrodes were attached to the participant's body. This electrode setup took about 20 minutes. After this setup step, the participant watched and performed a training trial to allow the participants to get familiar with the experimental protocol. This training trial comprised of displaying a short video and asking the participants to rate the video accordingly, once the video was finished. The experiment started with a 2 minute rest period in which the participants were asked to relax for baseline recording. The experiment consisted of 40 trials each of which contained:

1. two second screen displaying the current trial number;
2. a five second baseline recording (a fixation cross was displayed);
3. playing of the one minute stimulus video;
4. emotional self-assessment using arousal, valence, like/dislike rating and dominance dimensions.

[15] http://www.biosemi.com
[16] http://www.neurobs.com

Participants also had the option to take a break after the 20th trial, and they were offered soft drink and snack. Fig. 3.24 shows a participant shortly before the start of the experiment.

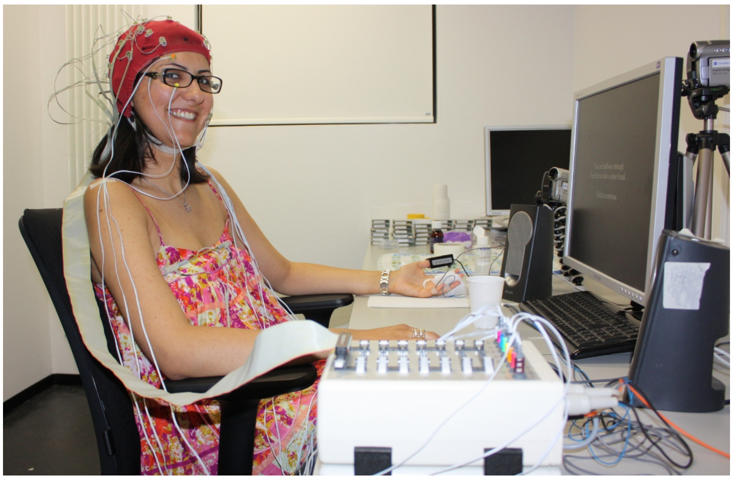


Figure 3.24: A participant shortly before the experiment.

Participants reported their felt emotions using different levels of arousal, valence, like/dislike rating and dominance. To facilitate understanding of different dimensions scales, SAM [40] were displayed (see Fig. 3.25). For the like/dislike rating scale, thumbs down/thumbs up symbols were displayed. The manikins were displayed in the center of the screen with the numbers 1-9 printed below. Participants could only move the mouse horizontally below the numbers and clicked to report their felt emotions.

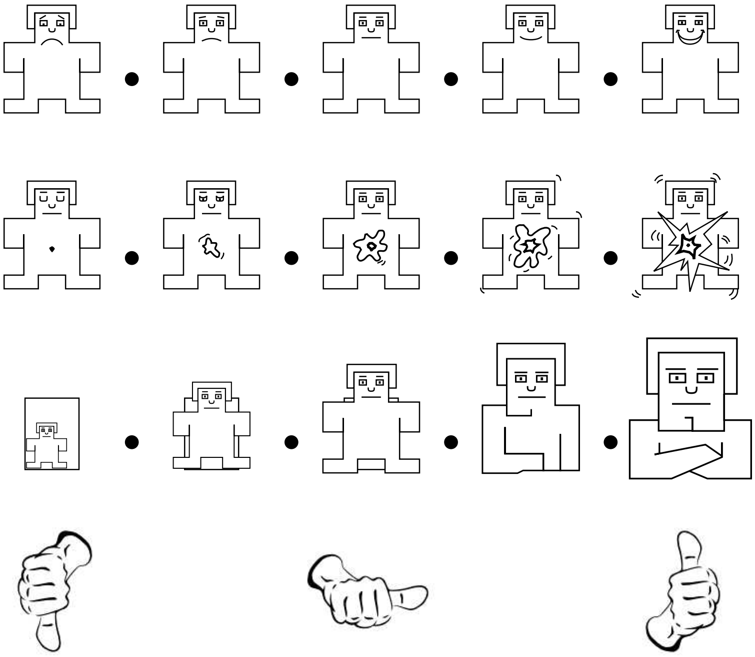


Figure 3.25: Images used for self-assessment. from top: Valence SAM, Arousal SAM, Dominance SAM, Liking.

Finally, in a post experiment questionnaire, participants were asked to rate their familiarity with each of the songs on a scale of 1 ("Never heard it before the experiment") to 5 ("Knew the song very well"). The analysis of ratings are given in subsection 3.1.2.4.

3.2.3 Recommendations

Inducing emotions and recording affective reactions is a challenging task. Special attention needs to be paid to several crucial factors, such as stimuli that are used, the laboratory environment, as well as the recruitment of participants. Our experience can be distilled into a list of recommendations that will enable the development of additional corpora to proceed smoothly.

- The experiment environment in the laboratory should be kept isolated from the outside environment. The participants should not be able to see the examiners or hear the noise from the outside. The light and temperature should be controlled to avoid variation in physiological reactions due to non controlled parameters.
- Choosing the right stimuli material is an important factor in any affective study. They should be long enough to induce emotions and short enough to prevent boredom. Furthermore, to be sure variation in stimuli length does not introduce variance in the measurements between emotional and non-emotional stimuli, we suggest the stimuli durations to be equal. The mixture of contradicting emotions can make problems for self-assessments. If the goal of the study is to find only one tag or recognize one specific emotional response, we recommend using videos which do not induce multiple emotions.
- A correct participant recruitment can make a significant difference in the results. A motivated participant with the right skills for filling the questionnaire on a computer is desirable, due to the nature of affective experiments. The rewards can make the participants more motivated and responsible. However, cash compensations might attract participants who are not motivated or lack desired communication skills. Therefore, rewarded recruitment should be done with considerations of desired population, e.g., gender balance, age distribution, etc.
- A complex apparatus is more likely to fail at any moment during experiments. Sometimes, a recording failure may reveal itself after the recordings have been done. Thus, planning of extra recording sessions is recommended.
- Contact lenses usually cause participants to blink more, which introduces a higher level of artifacts on EEG signals. Therefore, the participants with visual correction should avoid using contact lenses as much as possible. Thick glasses affect the eye gaze tracker performance. In the experiments in which both these modalities are recorded, recruiting participants with no visual correction is advised.
- Properly attending to participants takes an prominent part of one's attention, which can easily lead to forgetting parts of complicated technical protocols. Therefore, operation of the recording equipment during the experiments should be made as simple as possible (preferably just by pressing a single button). Alternatively, the tasks of controlling and monitoring correct data collection and attending to participants, can be divided between multiple laboratory assistants with carefully defined procedures.

3.3 Summmary

In this Chapter, an overview of relevant existing affective databases was given. Developed databases, in the course of our studies including stimuli videos and recorded responses, were introduced. Effect of different parameters such as circadian cycles, gender and mood on emotional ratings were investigated. In the following Chapter 4, the analysis carried out from bodily responses which were recorded in the developed databases will be provided.

Chapter 4

Emotion Detection from Bodily Responses

Emotional experience of viewers in response to videos can be identified from their expressions and physiological responses. In this chapter, I report on the studies, conducted during my doctoral studies, to detect emotions in response to videos. First, I give a brief introduction to the bodily signals. Then, the methods and results of emotion detection analyzed and validated from the three recorded experiments will be discussed.

4.1 Bodily Signals and Sensors

4.1.1 Physiological signals and sensors

In this Section, I provide a brief description of the peripheral physiological sensors, which were used, and signals, which were recorded during my studies. All the physiological signals including EEG signals were recorded using a Biosemi Active two MK1 [1] system. Biosemi Active II systems can record physiological signals with a sampling rate up to 16kHz using 24 bits A/D converters (see Fig. 4.1). The A/D box connects to a Personal Computer (PC) via Universal Serial Bus (USB) connection and is insulated from the acquisition system using an optical fiber. A participant wearing the EEG head cap with Biosemi sensors attached is shown in Fig. 4.2.

4.1.1.1 Electromyogram

EMG are the signals generated by skeletal muscles, which can be recorded by means of electrodes attached to the skin covering those muscles. Normally, a pair of electrode are attached along the muscle to record the difference of the electrical potential between two points. An EMG signal is the electrical potential generated by muscle cells, which is caused by the electrical or neurological activation of muscular cells. Facial and gesture emotional expressions activate different muscles. The muscular activity related to emotional responses can be measured by electromyography. For example, smiling activates Zygomaticus major and Frontalis muscles activity is a sign of attention or stress [105].

[1] http://www.biosemi.com

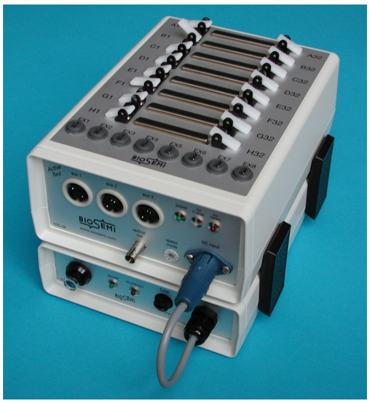


Figure 4.1: Biosemi active two box and battery.

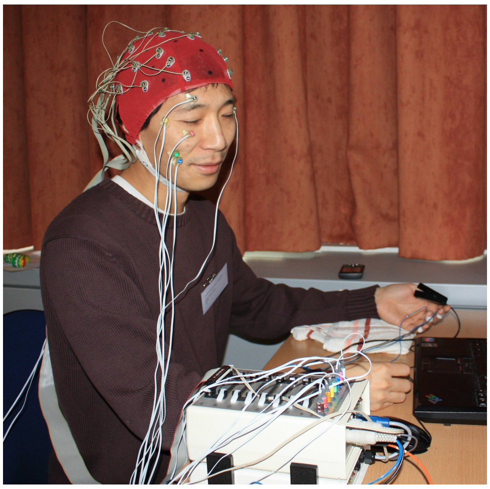


Figure 4.2: A participant wearing an EEG head cap with 32 electrodes. EOG and EMG flat electrodes are attached on the participant's face. The temperature, GSR sensors and plethysmograph are attached to the participant's left hand.

4.1.1.2 Galvanic skin response

GSR is a physiological signal which measures electrical conductance or resistance between two points on the skin. Skin's electrical conductance varies with its level of moisture which is mostly caused by sweat glands. Sweat glands are controlled by the sympathetic nervous system and their activity changes by emotional arousal [106]. Lang et al. discovered that the mean value of GSR is related to the level of motional arousal [57]. GSR is also known by Skin conductance (SC), electrodermal response (EDR), psychogalvanic reflex (PGR), skin conductance response (SCR). In my studies, GSR was measured by positioning two electrodes on the distal phalanges of the middle and index fingers and passing a negligible current through the body. An example of GSR response is shown in Fig. 4.4.

4.1.1.3 Electrocardiogram and blood volume pulse

Blood Volume Pulse (BVP) is the volume of blood in peripheral vessels measured by photoplethysmograph or plethysmograph. A photoplethysmograph consists of an infra-red emitter and detector. The amount of reflected infra-red light from the skin (usually finger) corresponds

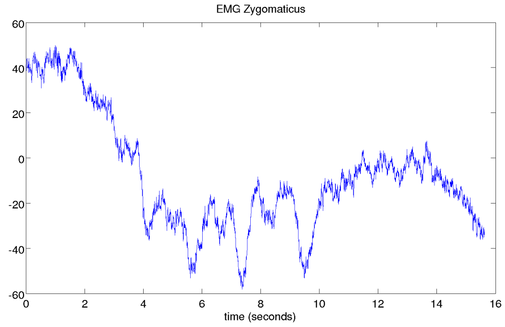


Figure 4.3: An EMG signal recorded from the zygomaticus major muscle.

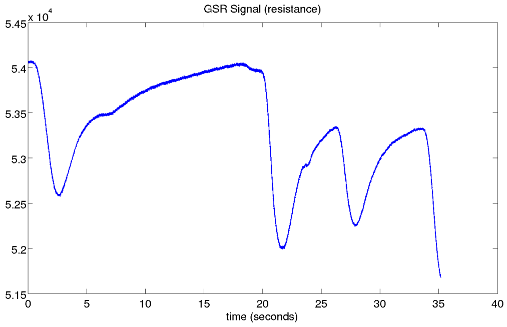


Figure 4.4: A raw GSR signal. The noisy GSR signal is visible with its tonic and phasic responses. Tonic response corresponds to the overall level of GSR whereas phasic response is the abrupt changes (drops in the current figure) due to the stimuli.

to the volume of blood in peripheral vessels. It is possible to derive heart rate by detecting peaks on BVP signals (see Fig. 4.5). BVP is also a relative measure of blood pressure and its level can be used as an indirect measure of blood pressure.

Because of the nature of muscular activity, heart muscle activity generates an electrical potential difference on the skin. These changes can be measured by placing electrodes on one's chest. Electrocardiography (ECG or EKG) is the electrical measurements by electrodes on the participant's chest, which mostly originate from heart activities. ECG signals depending on electrodes' placement can be interpreted and used to detect heart rate (HR) and heart rate variability (HRV) (see Fig. 4.6). HR and HRV correlate with emotional changes. Pleasantness of stimuli can increase peak heart rate response [57], and HRV decreases with fear, sadness, and

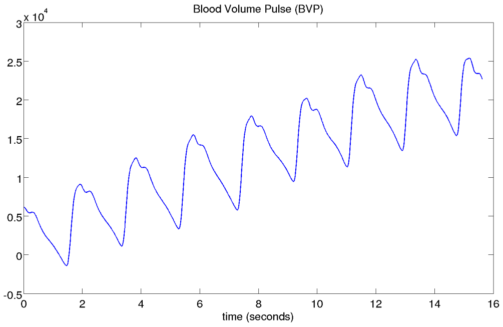


Figure 4.5: A BVP signal.

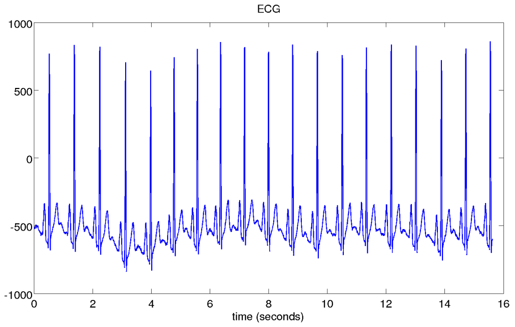


Figure 4.6: An ECG signal recorded from electrodes placed on a participant's upper corner of chest below the clavicle bone and the abdomen below the last rib. The strong contraction of the left ventricle generated large peaks visible in the signal.

happiness [107]. In addition to the HR and HRV features, spectral features derived from HRV were shown to be a useful feature in emotion assessment [108].

4.1.1.4 Respiration

Respiration depth or amplitude can be measured by measuring the expansion of the chest or abdomen circumference. A flexible belt with a piezo-electric crystal sensor which measures the belt's expansion can measure respiration patterns. Respiration rate and depth can be measured from the resulting signal (see Fig. 4.7). Respiration pattern (RSP) has been also shown to change in different emotional responses. Slow respiration is linked to relaxation whereas irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear [8, 107].

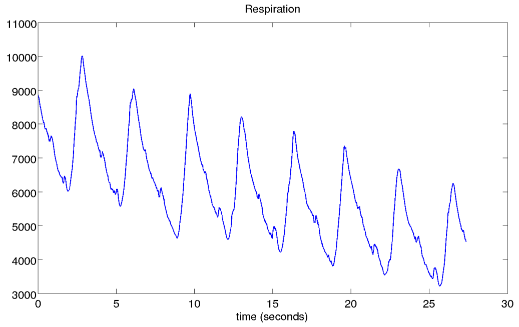


Figure 4.7: A RSP signal is shown. Counting the signal peaks provides us with a measure of respiration rate.

4.1.1.5 Skin temperature

Skin temperature also changes in different emotional states [109]. Although temperature changes are slower than other signals, they can be an indicator of emotional responses [110]. Skin temperature is measured by placing a temperature sensor directly in contact with participants' skin, e.g., their finger.

4.1.1.6 EEG

Psychological studies regarding the relations between emotions and the brain are uncovering the strong implication of cognitive processes in emotions [111, 112] (see Section 2.2.1). As a result, the EEG signals carry valuable information about the participants' felt emotions.

EEG signals are recorded by measuring electrical potential by placing electrodes on one's scalp. EEG signals result from the ensemble of the activation of millions of neurons in the brain. A famous electrode placing system is the international 10-20 system, which was employed in all the studies in this thesis. Electrodes are placed on approximately fixed positions according to their distance between inion and nosion point of a participant. The spiking neurons' electrical potentials (in the order of 10 μVolts) are very weak compared to muscular activities (order of 200 μVolts). Therefore, EEG signals can be easily contaminated by facial muscle activities, eye movements and environment noise. The unwanted artifacts, 50Hz or 60Hz power interference, EEG signals are usually recorded via active electrodes, which amplify the EEG signals on the electrodes to increase their signal to noise ration (SNR). The second solution to avoid environment noise is to record EEG signals in rooms which are isolated and shielded to electromagnetic fields, i.e, Faraday cages.

4.1.2 Eye gaze and pupil diameter

Eye gaze trackers use the reflection of infrared or near infrared red light from retina detected by cameras to track eye gaze. This allows measuring the pupil diameter with the eye gaze simultaneously. The eye gaze tracker technology utilized in this research, Tobii X120 [2] eye gaze tracker, provides the position of the projected eye gaze on the screen, the pupil diameter, the moments when the eyes were closed and the instantaneous distance of the participant's eyes to the gaze tracker device (see Fig. 4.8). Pupil diameter has been shown to change in different emotional states [69, 70].

Figure 4.8: The Tobii X120 eye gaze tracker which was used in this study.

[2] http://www.tobii.com

4.2 Peripheral Physiological Signals for Affective Tagging

4.2.1 Video affective characterization using peripheral physiological responses

4.2.1.1 Overview

A video dataset of 64 movie scenes from 8 movies was manually selected. Experiments were conducted during which physiological signals were recorded from spectators. After each scene, the spectator self-assessed his/her valence-arousal levels. To reduce the mental load of the participants, the protocol divided the show into 2 sessions of 32 movie scenes each. Each of these sessions lasted approximately two hours, including setup. The processed data include the responses of eight healthy participants (three female and five male, from 22 to 40 years old) participated in the experiment. The details on the experimental setup and protocol are given in detail in Section 3.2.2.1.

After finishing the experiment three types of affective information about each movie clip were available:

– multimedia content-based information extracted from audio and video signals;

– physiological responses from spectators' bodily reactions (due to the autonomous nervous system) and from facial expressions;

– self-assessed arousal and valence, used as "ground truth" for the true feelings of the spectator.

Next, we aimed at demonstrating how those true feelings about the movie scenes can be obtained by using the information that is either extracted from audio and video signals or contained within the recorded physiological signals. To this end, features that were likely to be influenced by affect were extracted from the audio and video content as well as from the physiological signals. Thus, each feature from feature vectors of 64 samples highlights a single characteristic (for instance, average sound energy) of the movie scenes. In a similar way, feature vectors were extracted from the physiological signals. As one may expect, a single feature, e.g., average sound energy, may not be equally relevant to the affective feelings of different participants. In order to personalize the set of all extracted features, an additional operation called relevant-feature selection has been implemented. During the relevant-feature selection for arousal, the correlation between the single-feature vectors and the self-assessed arousal vector is determined. Only the features with high absolute correlation coefficient ($|\rho| > 0.25$) were subsequently used for estimating arousal. A similar procedure was performed for valence. It will be shown that accurate estimates of the self-assessed arousal and valence can be obtained based on the relevant feature vectors for physiological signals as well as from the relevant feature vectors for audio and video information.

4.2.1.2 Feature extraction

Galvanic Skin Response (GSR), Blood Volume Pulse (BVP), ElectroMyoGram (EMG), skin temperature, and respiration pattern were recorded in the experiments. The placement of electrodes on the face (for EMG) and ground electrodes on the left hand enabled us to record an ECG signal. Using the electrocardiogram, the pulse transit time was computed as a feature. In addition to the heart rate and heart rate variability features, the multi-scale entropy (MSE) of the heart rate variability was computed from ECG signals. The MSE of the heart rate was shown to be a useful feature in emotion assessment [8].

Regarding the EMG signals, the Frontalis muscles activity is a sign of attention or stress in facial expressions. The activity of the Zygomaticus major was also monitored, since this muscle is active when the user is laughing or smiling. Most of the power in the spectrum of an EMG during muscle contraction is in the frequency range between 4 to 40 Hz. Thus, the muscle activity features were obtained from the energy of EMG signals in this frequency range for the different muscles. The rate of eye blinking is another feature, which is correlated with anxiety [113]. Eye-blinking affects the EMG signal that is recorded over the Frontalis muscle and results in easily detectable peaks in that signal. All the extracted features are listed in Table 4.1.

Table 4.1: This table lists all 66 features extracted from physiological signals. Number of features extracted from each channel is given in brackets.

Peripheral Signal	**Extracted features**
GSR (16)	Average skin resistance, average of derivative, mean of derivative for negative values only(average decrease rate during decay time), proportion of negative samples in the derivative vs. all samples, Number of local minima in the GSR signal, average rising time of the GSR signal, kurtosis, skewness, spectral power in the bands([0-0.1]Hz, [0.1-0.2]Hz, [0.3-0.4]Hz)
BVP (20) (64)	Average BVP, heat rate, heart rate derivative, heart rate variability, standard deviation of heart rate, ECG multi-scale entropy (4 levels), finger pulse transit time, kurtosis and skewness of the heart rate, energy ratio between the frequency bands [0, 0.08]Hz and [0.15, 5]Hz, spectral power in the bands ([0-0.1]Hz, [0.1-0.2]Hz, [0.3-0.4]Hz)
RSP (10)	Band energy ratio (energy ratio between the lower (0.05-0.25Hz) and the higher (0.25-5Hz) bands), average respiration signal, mean of derivative (variation of the respiration signal), standard deviation, dynamic range or greatest breath, breathing rhythm (spectral centroid), breathing rate, spectral power in the bands ([0-0.1]Hz, [0.1-0.2]Hz, [0.3-0.4]Hz)
Skin temperature (11)	Range, average, minimum, maximum, standard deviation, kurtosis, skewness, spectral power in the bands([0-0.1]Hz, [0.1-0.2]Hz, [0.3-0.4]Hz)
EMG Zygomaticus and Frontalis (9)	Energy, average, standard deviation of energy, variance, Rate of eye blinking per second, extracted from the Frontalis EMG

4.2.1.3 Feature selection and regression

The relevance of features for affect was determined using Spearman correlation between each extracted feature and the users' self-assessment. I now demonstrate how user-felt arousal and valence can be estimated, based on the physiological which were found to have a significant correlation with the self-assessed valence and arousal. For each participant, a training set of 63 scenes was formed by selecting 63 of the 64 movie scenes and the corresponding feature values. The remaining scene served as a test set. In order to obtain an estimate, based on the significantly correlated features, of the user's valence and arousal, all significantly correlated features are weighted and summed as is indicated in Equation 4.1, where $\hat{y}(j)$ is the estimate of valence-arousal grade, j is the indexing number of a specific movie scene $1, 2, .., 64$, $x_i(j)$ is the feature vector corresponding to the i-th significantly correlated feature, N_s is the total number of significant features for this participant, and w_i is the weight that corresponds to the i-th feature.

$$\hat{y}(j) = \sum_{i=1}^{N_s} w_i x_i(j) + w_0 \tag{4.1}$$

In order to determine the optimum $\hat{y}$, the weights in Equation 4.1 were computed by means of a RVM from the Tipping RVM toolbox [114]. This procedure was applied on the user self assessed valence-arousal, $y(j)$, and on the feature-estimated valence-arousal, $\hat{y}(j)$, over all movie scenes. This procedure is performed two times for optimizing the weights corresponding to:

- physiological features when estimating valence,
- physiological features when estimating arousal,

In a first step, weights are computed from the training set. In the second step, the obtained weights were applied to the test set, and the mean absolute error between the resulting estimated valence-arousal grades and self assessed valence-arousal was calculated. These two steps were repeated 64 times. Each time the 63 movie scenes of the training set were selected from the total of 64 scenes while the single, remaining scene served as the test set. The results from this cross-validation will be presented in the following Section.

4.2.1.4 Experimental results

To video dataset presented in Section 3.1.2.1 was used as stimuli. Peripheral signals and facial expression EMG signals were recorded for emotion assessment. EMG signals from the right Zygomaticus major muscle (smile, laughter) and right Frontalis muscle (attention, surprise) were used as indicators of facial expressions. GSR, skin temperature, breathing pattern (using a respiration belt) and BVP (using a plethysmograph) were also recorded. All physiological data was acquired via a Biosemi Active-two system with active electrodes, from Biosemi Systems (http://www.biosemi.com). The data were recorded with a sampling frequency of 1024 Hz in a sound-isolated Faraday cage. The GSR and respiration signals were respectively smoothed by a 512 and a 256 points averaging filters to reduce the high frequency noise. EMG signals were filtered by a Butterworth band pass filter with a lower cutoff frequency of 4 Hz and a higher cutoff frequency of 40 Hz.

The correlations between physiological features and self assessments were determined. Table 4.2 shows, for each participant, the features which had the highest absolute correlations with that participant's self-assessments of valence and arousal.

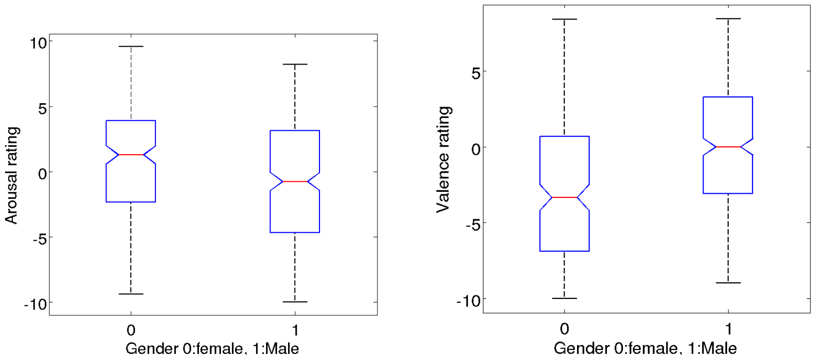
Table 4.2: Physiological features with the highest absolute correlation with self assessments for participants 1 to 8.

Part.	**Arousal**	ρ	**Valence**	ρ
1	GSR Skewness	0.43	EMG Zygomaticus (sum of absolute)	0.74
2	EMG Frontalis (sum of absolute)	0.66	EMG Frontalis (sum of absolute)	-0.73
3	GSR power spectral density 0.1-0.2 Hz band	0.48	EMG Zygomaticus(sum of absolute)	0.53
4	EMG Zygomaticus average	0.32	EMG Zygomaticus(sum of absolute)	0.49
5	EMG Frontalis (sum of absolute)	0.38	EMG Frontalis (sum of absolute)	-0.49
6	Plethysmograph multi-scale entropy (2nd)	0.42	EMG Zygomaticus (sum of absolute)	0.56
7	GSR standard deviation	0.55	EMG Zygomaticus (sum of absolute)	0.71
8	BVP (average)	-0.33	EMG Zygomaticus (sum of absolute)	0.64

For physiological signals, the variation of correlated features over different subjects illustrates the difference between participants' responses. While GSR features are more informative regarding the arousal level of participants 1, 3, and 7, EMG signals were more important to estimate arousal in participants 2, 4, and 5.

The difference between the self assessments of male and female participants was investigated by means of a one way ANOVA test of variance applied on these assessments. The difference between the two genders group self assessments was found to be significant for gender groups' valence ($F = 50.6,\ p < 0.005$) and arousal ($F = 11.9,\ p < 0.005$), and for participants' valence ($F = 10.3,\ p < 0.005$) and arousal ($F = 20.3,\ p < 0.005$). The female participants reported lower valence and higher arousal on average. Comparison with assessed valence showed that this gender difference comes from the fact that female participants reported higher level of unpleasantness and excitement. Rottenberg et al.[9] showed that female participants reported more intense emotions in response to emotional movie scenes. The female participants' emotional responses in our study were also stronger compared with those from male participants. Fig. 4.9 shows the results of this one way ANOVA test on the two gender groups' valence self assessments.

The accuracy of the estimated valence and arousal is evaluated by computing the mean absolute error between the estimates and the self-assessments of either valence or arousal (Table 4.3). The mean absolute error (E_{MAE}) was calculated from a leave-one-out cross validation on



(a) arousal, $p = 6 \times 10^{-4}$, $F = 11.9$ (b) valence, $p = 3.8 \times 10^{-12}$, $F = 50.6$

Figure 4.9: Results of the one way ANOVA test on the self assessments showing significant differences between the average assessments levels of the two gender groups.

Table 4.3: Mean absolute error (E_{MAE}), and Euclidean distance (E_{ED}) between estimated valence-arousal grades and self assessments (participants 1 to 8).

Part.	E_{MAE} **Arousal estimated from physiological features**	E_{MAE} **Valence estimated from physiological features**	E_{ED} **Physiological features**
1	0.17	0.10	0.21
2	0.12	0.09	0.16
3	0.15	0.11	0.21
4	0.15	0.12	0.20
5	0.15	0.14	0.22
6	0.18	0.18	0.27
7	0.16	0.11	0.21
8	0.16	0.07	0.18
Average	0.15	0.11	0.21
Random level	**∼0.4**	**∼0.4**	**∼0.5**

64 video clips for each participant.

$$E_{MAE} = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}=64} |\hat{y}_j - y_j| \tag{4.2}$$

E_{MAE} was computed from Equation 4.2 where N_{test} is the number of test samples (64) and $\hat{y}_j$ is the estimated valence-arousal for the j-th sample in test set. E_{MAE} values are shown in Table 4.3; all E_{MAE} values are considerably smaller than a random level determination of

E_{MAE} (which is around 0.4, and is estimated by generating random measurements of valence and arousal). While E_{MAE} separately considers valence and arousal determinations, a more global performance measure can be defined. Considering valence and arousal as coordinates in the 2-D valence-arousal space, the overall accuracy of the estimated, joint valence-arousal grades was evaluated by computing the Euclidean distance (E_{ED}) between the estimated points and the self-assessments (ground truth). This Euclidean distance is a useful indicator of the system's performance for affect representation and affects similarity measurement, when using valence and arousal as indicators. With valence and arousal being expressed in normalized ranges [0-1], E_{ED} is computed as follows:

$$E_{ED} = \frac{1}{N_{test}} \sum_{j=1}^{N_{test}=64} \sqrt{(\hat{y}_j^{arousal} - y_j^{arousal})^2 + (\hat{y}_j^{valence} - y_j^{valence})^2} \tag{4.3}$$

E_{ED} values are shown in Table 4.3. It can in particular be observed that the average Euclidean distance results are all below random level (which is around 0.5). The E_{MAE} represents the distance of the determined emotion from the self-assessed emotion in the dimensions of arousal or valence. E_{MAE} is thus useful to compare each dimension's results. The E_{MAE} of arousal and valence shows that valence determination was more precise than arousal determination. The superior valence results might have been caused by the fact that the pleasantness of emotion is easier to understand and report for participants.

4.3 Multi-Modal Emotion Recognition in Response to Videos

A user-independent emotion recognition method with the goal of recovering affective tags for videos using EEG, pupillary response and gaze distance was developed and evaluated. We first selected 20 video clips with extrinsic emotional content from movies and online resources. Then EEG responses and eye gaze data were recorded from 24 participants while watching emotional video clips. Ground truth was defined based on the median arousal and valence scores given to clips in a preliminary study using an online questionnaire. Based on the participants' responses, three classes for each dimension were defined. The arousal classes were calm, medium aroused and activated, and the valence classes were unpleasant, neutral and pleasant. One of the three affective labels of either valence or arousal was determined by classification of bodily responses. The details on the experimental setup and protocol are given in detail in Section 3.2.2.2.

4.3.1 Preliminary study

In the preliminary study, 21 commercially produced movies were first segmented into their scenes. Scenes longer than two minutes were divided into shorter two minutes long excerpts. From these excerpts, 155 emotional video clips containing excerpts extracted from movies were manually selected. The 155 selected videos were shown to more than 50 participants; each video clip received 10 annotations on average [18]. The preliminary study was conducted utilizing an online affective annotation system in which the participants were able to use a web interface to report their emotions in response to the videos played by a web-based video player (see Fig. 3.3). In case of using videos from online repositories, the full length videos were used in the dataset.

In the preliminary study, the participants were thus asked to self-assess their emotion by reporting the felt arousal (ranging from calm to excited/activated) and valence (ranging from unpleasant to pleasant) on nine points scale as well as emotional keywords. 14 video clips were chosen based on the preliminary study from the clips which received the highest number of emotional keyword tags in different emotion categories, which are listed in the Table 4.3.1.1. Videos were selected to cover different emotional responses (see Fig. 4.10). Three other popular video clips from online resources were added to this set (two for joy/happiness and one for disgust). Three past weather forecast reports (retrieved from youtube.com) were also used as neutral emotion clips. The videos from online resources were added to the dataset to enable us to distribute some of the emotional video samples with the recorded multi-modal dataset described below. Table 4.3.1.1 gives the emotional labels, titles, and sources of the emotional video clips.

The median arousal and valence was used to determine ground truth labels with the following procedure. First, the values assessed by the online questionnaire were centered and then three equal length intervals were defined on the assessment range ($arousal, valence \in [1, 9]$). The labels assigned to all videos are given in Table 4.3.1.1. The distribution of online self emotions for the selected videos is shown in Fig. 4.10.

Regarding the self-reports, we computed the average pair-wise Cohen's kappa for keyword based annotations. A fair multi-rater agreement was found on emotional keywords (9 keywords) with $\kappa = 0.32$. The correlation between arousal and valence ratings between participants was also computed. The correlation between arousal and valence ratings given by different participants on nine points scales were $mean(\rho) = 0.45,\ SD(\rho) = 0.25$ and $mean(\rho) = 0.73,\ SD(\rho) = 0.12$ respectively. Therefore, there was a higher inter-rater agreement on valence comparing to arousal.

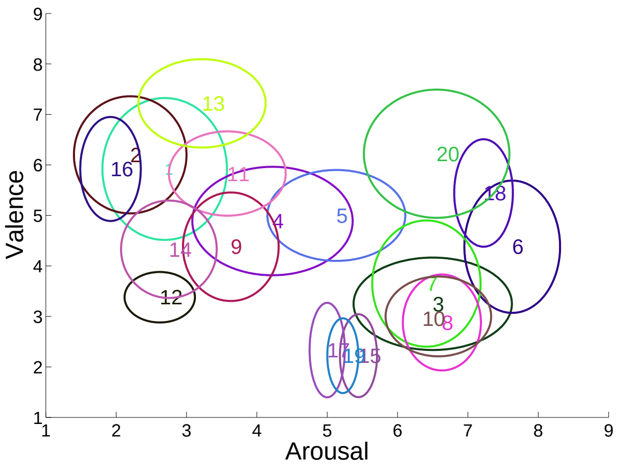


Figure 4.10: Stimulus videos are shown in the valence-arousal plane. The center of the ellipses represents the mean arousal, and valence and the horizontal and vertical radius represents the standard deviation of the online assessments. The clip codes are printed at the center of each ellipse.

Table 4.4: The video clips are listed the with their sources. The emotion labels are: calm (Cal.), medium aroused (Med.), activated (Act.), unpleasant (Unp.), neutral valence (Neu.), Pleasant (Pls.).

Code	**Emotion Labels**	**Video clips sources**
1	Act., Unp.	Hannibal
2	Act., Unp.	The Pianist
3	Med., Pls.	Mr. Bean's holiday
4	Act., Neu.	Ear worm (blip.tv)
5	Med., Neu.	Kill Bill VOL I
6	Med., Pls.	Love actually
7	Med., Pls.	Mr. Bean's holiday
8	Cal., Pls.	The thin red line
9	Med., Neu.	The shining
10	Med., Pls.	Love actually
11	Act., Unp.	The shining
12	Med., Unp.	Gangs of New York
13	Act., Unp.	Silent hill
14	Med., Unp.	The thin red line
15	Cal., Neu.	AccuWeather New York weather report (youtube.com)
16	Act., Unp.	American history X
17	Cal., Neu.	AccuWeather Detroit weather report (youtube.com)
18	Act., Pls.	Funny cats (youtube.com)
19	Cal., Neu.	AccuWeather Dallas weather report (youtube.com)
20	Act., Pls.	Funny (blip.tv)

4.3.1.1 Ground truth definition

The median arousal and valence was used to determine ground truth labels with the following procedure. First, the values assessed by the online questionnaire were centered and then three equal length intervals were defined on the assessment range ($arousal, valence \in [1, 9]$). The labels assigned to all videos are given in Table 4.3.1.1. The distribution of online self emotions for the selected videos is shown in Fig. 4.10.

Ultimately, 20 videos were selected to be shown which were between 35s to 117s long ($M = 81.4s$, $SD = 22.5s$). Psychologists recommended videos from one to ten minutes long for elicitation of a single emotion [44, 9]. Here, the video clips were kept as short as possible to avoid multiple emotions or habituation to the stimuli while keeping them long enough to observe the effect.

4.3.2 Multi-modal emotion recognition

4.3.2.1 EEG signals

Electroencephalogram signals were recorded with a 1024Hz sampling rate and later downsampled to 256Hz to reduce the memory and processing costs. EEG signals were recorded using active AgCl electrodes placed according to the international 10-20 system. The layout of EEG electrodes on the cap are shown in Fig. 4.11. The unwanted artifacts, trend and noise were reduced prior to extracting the features from EEG data by pre-processing the signals. Drift and noise reduction were done by applying a 4-45Hz band-pass filter. Other artifacts such as muscular activity was kept at a low level by instructing the participants to minimize their movements while videos were playing. Biosemi active electrodes record EEG signals referenced to common mode sense electrode (CMS) as a part of its feedback loop. In order to gain the full common-mode rejection ratio (CMRR) at 50Hz, EEG signals should be re-referenced to another reference. EEG signals were thus re-referenced to the average reference to maximize signal to noise ratio.

The spectral power of EEG signals in different bands was found to be correlated with emotions [115, 116, 61]. Power spectral density (PSD) from different bands were computed using fast Fourier transform (FFT) and Welch algorithm [117]. In this method, the signal is split into overlapping segments and the PSD is estimated by averaging the periodograms. The averaging of periodograms results in smoother power spectrum. The PSD of each electrode's EEG signals was estimated using 15s long windows with 50% overlapping.

The logarithms of the PSD from theta ($4Hz < f < 8Hz$), slow alpha ($8Hz < f < 10Hz$), alpha ($8Hz < f < 12Hz$), beta ($12Hz < f < 30Hz$) and gamma ($30Hz < f$) bands were extracted from all 32 electrodes as features. In addition to power spectral features, the difference between the spectral power of all the 14 symmetrical pairs on the right and left hemisphere was extracted to measure the possible asymmetry in the brain activities due to the valence of an emotional stimuli [118, 115]. The asymmetry features were extracted from all mentioned bands except slow alpha. The total number of EEG features of a trial for 32 electrodes is $14 \times 4 + 32 \times 5 = 216$ features. A list of extracted EEG features is given in Table 4.5.

4.3.2.2 Eye gaze data

The X120 Tobii [3] eye gaze tracker provides the position of the projected eye gaze on the screen, the pupil diameter, the moments when the eyes were closed and the instantaneous distance of the participant's eyes to the gaze tracker device positioned below the screen. The eye gaze data was sampled at 60Hz. The blinking moments are also extractable from eye gaze data. The eye gaze itself is highly dependent on the content and therefore, it was not used directly for emotion recognition. However, pupil diameter has been shown to change in different emotional states [69, 70].

A linear interpolation was used to replace the missing pupil diameter samples due to eye blinking. Then, the average diameter of right and left eye pupil was used as the pupil diameter time series. The major cause of pupil diameter variation comes from lighting; therefore, the

[3] http://www.tobii.com

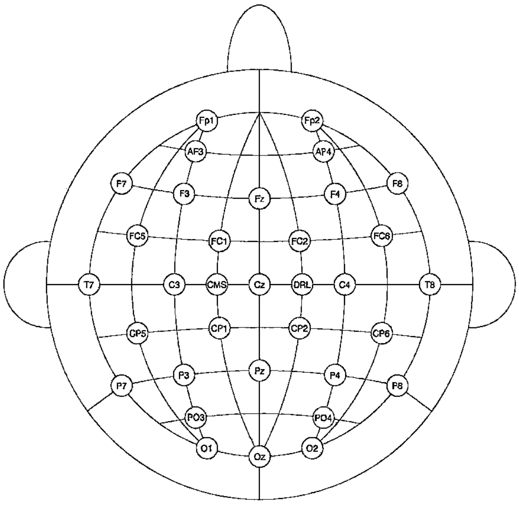


Figure 4.11: The EEG cap layout for 32 EEG in addition to two reference electrodes. Retrieved from Biosemi website (http://www.biosemi.com).

participants' responses to the same video (stimuli) in the controlled lighting environment follow similar patterns. There are different parametric models for pupillary light reflex [119, 120]. However, these parametric models are not error free and calculating their numerous parameters is rather difficult without specific light reflex experiment. It has been shown that the pupil diameter variation as a result of light changes with age and between different people [119]. Most of the participants in our experiment were young, in their twenties; therefore, the aging effect was assumed to be negligible. The difference between the magnitudes can be reduced by normalizing the pupil diameter time series. Consequently we extracted the light reflex using a non-parametric estimation from the data. This common lighting reflex pattern was estimated for each video using PCA.

If Y is the $M \times N_p$ matrix containing the centered and normalized pupillary responses to the same video from N_p participants and M samples, then Y consists of three components:

$$Y = X + Z + E \tag{4.4}$$

X is the lighting response which is the strongest effect on the signal. Z is the parasympathetic emotional and attentional response and E is the noise originated from measurement. These three components are originated from independent sources, and the decorrelating characteristic of PCA is able to separate these three. First, Y was decomposed using PCA into N_p components. The first principal component is assumed to be a close estimation of the lighting reflex. The normalized principal component was then removed from normalized time series. Then the remaining residual part includes $Z + E$.

$$Y = UDV^T \tag{4.5}$$

$$A_p = UD \tag{4.6}$$

Table 4.5: This table list all the features extracted from eye gaze data and EEG signals.

Eye gaze data	Extracted features
Pupil diameter	standard deviation, spectral power in the following bands:]0, 0.2]Hz,]0.2, 0.4]Hz,]0.4, 0,6]Hz and]0.6, 1]Hz
Gaze distance	approach time ratio, avoidance time ratio, approach rate
Eye blinking	blink depth, blinking rate, length of the longest blink, time spent with eyes closed
EEG	theta, slow alpha, alpha, beta, and gamma PSD for each electrode. The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma.

$$S_p = V^T \tag{4.7}$$

$$Y_1 = A_{p1} S_{p1} \tag{4.8}$$

$$Y_R = Y - Y_1 \tag{4.9}$$

If we decompose Y using singular value decomposition (SVD) U is a matrix with eigen vectors of YY^T as its column. D is a diagonal matrix whose diagonal values are the eigen values of YY^T. Finally, the columns of V are the eigen vectors of Y^TY (see Equation 4.5). From the principal components of Y, A_p, we can reconstruct the first principal component or the light reflex pattern Y_1 (see Equation 4.6). To remove the light reflex component, Y_1, from all the time series, it is enough to subtract it from the original data (see Equation 4.8 and 4.9). Y_R is the residual part which contains the emotional and attentional pattern, in addition to the noise.

After removing the linear trend, the power spectrum of the pupil diameter variation was computed. Standard deviation and spectral features were extracted from the pupil diameter. The Hippus effect is the small oscillations of eye pupil diameter between 0.05 to 0.3Hz and with the amplitude of 1 mm [119, 121]. Hippus effect has been shown to be present when one is relaxed or passive. In the presence of mental activity, the effect will disappear. The Hippus effect is extracted by the first two power spectral features which are covering up to 0.4 Hz. The rate of eye blinking is shown to be correlated with anxiety [113]. From the eye blinks, the eye blinking rate, the average and maximum blink duration were extracted as features. In addition to the eye blinking features, the amount of time the participants spent with his/her eyes closed was also used as a feature to detect possible eye closing due to unpleasant emotions.

Although the participants were asked not to move during the experiment, there were small head movements which manifested itself in the distance between participants' eyes and the eye gaze tracker. The distance of the participant to the screen and its changes provide valuable information about the participants' posture. The total change in the distance of the user to the gaze tracker, gaze distance, was calculated to measure the possible approach and avoidance phenomena. The amount of time the participant spent per trial getting close or far from the screen was computed as well. These features were named approach and avoidance ratio to represent the amount of time participant spent getting close or going far from the screen. The frequency of the participants' movement towards the screen during each trial, approach rate, was also extracted. Ultimately 12 features were extracted from the eye gaze data. A summary of all extracted features is given in Table 4.5.

All the extracted features were numerical. To reduce the between participant differences, it is necessary to normalize the features. Maximum-Minimum normalization was applied on each feature of the features set separately for each participant's signals. In this normalization, the minimum value for any given feature is subtracted from the same feature of a participant, and the results were divided by the difference between the maximum and minimum values.

4.3.2.3 Emotion classification

With the proposed inter-participant emotion recognition, the goal is to find the emotional class with the highest agreement within a population. The most popular emotional class or tag can satisfy a larger population of viewers in a video retrieval scenario. For each video from the dataset, the ground truth was thus defined by computing the median of arousal and valence scores given on a nine point scale. The median values were then categorized into three classes with equal intervals. According to this definition, we can name these classes calm, medium aroused, and activated for arousal and unpleasant, neutral, and pleasant for valence.

A SVM classifier with RBF kernel was employed to classify the samples using features from each of the two modalities. Prior to classification, a feature selection was used to select discriminative features as follows. First, a one way ANOVA test was done on only the training set for each feature with the class as the independent variable. Then any feature for which the ANOVA test was not significant ($p > 0.05$) was rejected. This feature selection criterion was hence re-calculated for each cross validation's iteration. A leave-one-participant-out cross validation technique was used to validate the user-independent classification performance. At each step of cross validation, the samples of one participant were taken out as test set, and the classifier was trained on the samples from the rest of the participants. This cross validation was employed to imitate the effect of introducing a new user to our emotion recognition system. This process was repeated for all participants' data.

4.3.2.4 Modality fusion strategy

Classification in different modalities can be fused at both feature level and decision level. We applied these two fusion strategies and reported their results. With the feature level fusion, the feature vectors from different modalities were concatenated to form a larger feature vector. The feature selection and classification methods were then applied to the new feature set. However with the decision level fusion, classification was performed on each modality separately and the classification outcomes were fused to generate the fusion results. In [62] feature level fusion of

EEG and peripheral physiological signals did not improve the single modality results. On the other hand, Chanel et al. [61] showed how a fusion strategy improved the emotion recognition accuracy by fusing the results from EEG and peripheral features at decision level. Our results (see Section 4.3.3) shows how in contrary to Feature Level Fusion (FLF), Decision Level Fusion (DLF) significantly outperforms the best single modality for arousal classification and do not underperform for valence classification.

In addition to the superior classification performance obtained by multi-modal strategy, in the absence of one of the modalities due to temporary problems or artifacts, the system can still continue working as a single modality emotion detection. The adaptability of the system to remove and add new modalities can be achieved without re-training the classifiers using the DLF. The adaptability and scalability of the DLF strategy gives it another advantage over FLF.

Here, we used two modalities, which are EEG and eye gaze data. The results of the classification over two modalities were fused to obtain the multi-modal fusion results. If the classifiers provide confidence measures on their decisions, combining decisions of classifiers can be done using a summation rule. The confidence measure summation fusion was used due to its simplicity and its proved performance for emotion recognition according to [61]. Other decision combination methods including product of confidence measures, decision template fusion, Dempster-Shafer, Bayesian belief integration [122], weighted sum and weighted product [123] did not give superior results.

The probabilistic outputs of classifiers are used as a measure of confidence. The sum rule is thus defined as follows for a given trial:

$$g_a = \frac{\sum_{q\in Q} P_q(\omega_a|x_i)}{\sum_{a=1}^{K} \sum_{q\in Q} P_q(\omega_a|x_i)} = \sum_{q\in Q} \frac{1}{|Q|} P_q(\omega_a|x_i) \tag{4.10}$$

In Equation 4.10, g_a is the summed confidence interval for affect class ω_a . Q is the ensemble of the classifiers chosen for fusion, $|Q|$ the number of such classifiers and $P_q(\omega_a|x_i)$ is the posterior probability of having class ω_a the sample is x_i according to classifier q. The final choice is done by selecting the class ω_a with the highest g_a. It can be observed that g_a can also be viewed as a confidence measure on the class, ω_a, given by the fusion of classifiers.

There are two problems employing SVM classifiers in this fusion scheme. First, they are intrinsically only two-class classifiers and secondly, their output is uncalibrated so that it is not directly usable as a confidence value in case one wants to combine outputs of different classifiers or modalities. To tackle the first problem, the one versus all approach is used where one classifier is trained for each class (N classifier to train), and the final choice is done by majority voting. For the second problem, Platt [124] proposes to model the probability of being in one of the two classes knowing the output value of the SVM by using a sigmoid fit, while Wu et al. [125] proposes a solution to extend this idea to multiple classes. In this study, we used the MATLAB libSVM implementation [126] of the Platt and Wu algorithms to obtain the posterior probabilities, $P_q(\omega_a|x_i)$.

4.3.3 Experimental results

The experiments were performed in a laboratory environment with controlled temperature and illumination; 24 participants viewed 20 video clips each. 467 samples were gathered over a potential dataset of $24 \times 20 = 480$ samples; the 13 missing ones were unavailable due to technical difficulties.

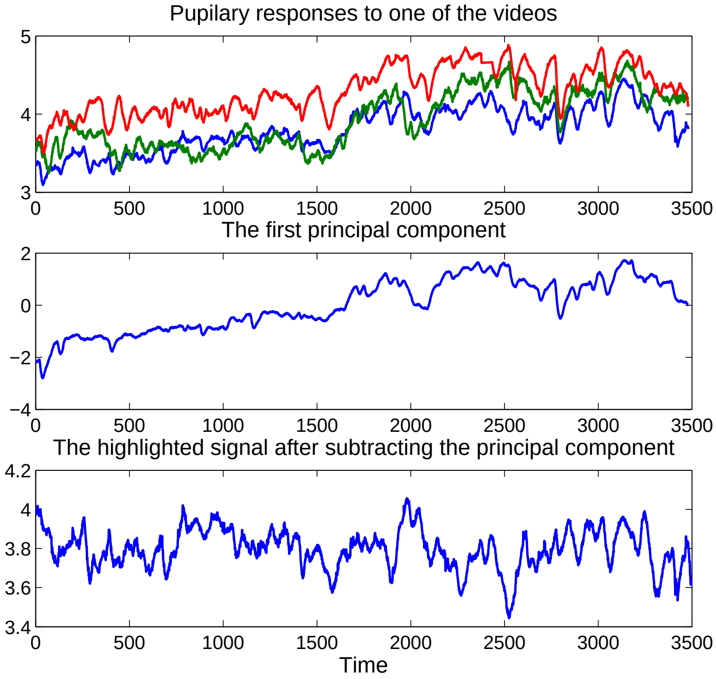


Figure 4.12: From top to bottom: on the first plot, there is an example of pupil diameter measures from three different participants in response to one video. The second plot shows the first principal component extracted by PCA from the time series shown in the first plot (the lighting effect). The bottom plot shows the pupil diameter of the blue signal in the first plot after reducing the lighting effect.

To find the best discriminative EEG features, the linear discrimination criterion was calculated. This parameter is the between class variance divided by within class variance for any given feature (see Table 4.6). For arousal classification, PSD in alpha bands of occipital electrodes was found to be the most discriminant features. In contrary for valence beta and gamma bands of temporal electrodes are more informative. The between class to within class variance ratios are higher for the best arousal EEG features. The higher linear discrimination criterion for best arousal features explains the superior classification rate for arousal dimension (see Table 4.7).

In order to study the discrimination abilities of the eye gaze data features, a one way analysis of variance test was performed on the features. The difference between the mean of features in different arousal or valence categories was found significant ($p < 0.05$). The significance of one way ANOVA shows that there is at least a significant difference between the means of the samples from two classes out of three. The box plots of four features namely, eye blinking rate, approach rate, maximum blink length, and standard deviation of pupillary responses are shown in Fig.

Table 4.6: 10 best EEG features for arousal and valence classification based on linear discrimination criterion. The between class variance to within class variance ratios, $\sigma^2_{bw}/\sigma^2_{wn}$ are also given.

Arousal classification			**Valence classification**		
Band	Electrode/s	$\sigma^2_{bw}/\sigma^2_{wn}$	Band	Electrode/s	$\sigma^2_{bw}/\sigma^2_{wn}$
Slow α	PO4	0.18	β	T8	0.08
α	PO4	0.17	γ	T8	0.08
θ	PO4	0.16	β	T7	0.07
Slow α	PO3	0.15	γ	T7	0.06
θ	Oz	0.14	γ	P8	0.05
Slow α	O2	0.14	γ	P7	0.05
Slow α	Oz	0.14	θ	Fp1	0.04
θ	O2	0.13	β	CP6	0.04
θ	FC6	0.13	β	P8	0.04
α	PO3	0.13	β	P7	0.04

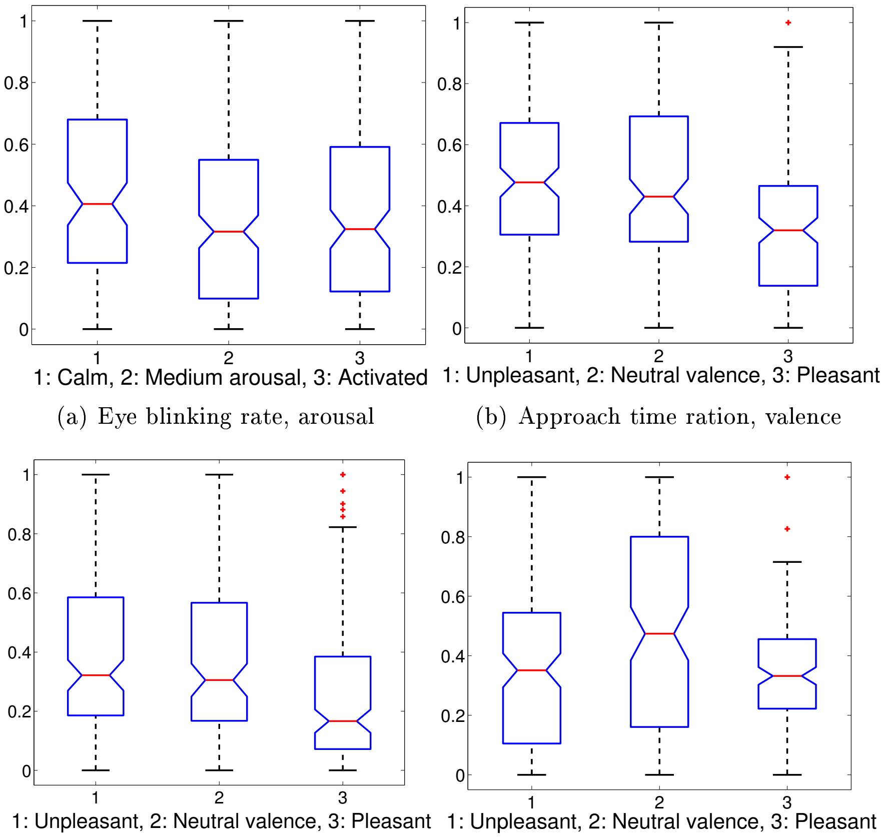
4.13. In average eye blinking rate was higher in calmer videos (see Fig. 4.13(a)). The amount of time participants spent getting closer to the screen is lower for the pleasant category. This shows that they had a tendency to seat more upright while watching more pleasant videos (see Fig. 4.13(b)). On the other hand, the maximum blink length or depth is higher for unpleasant videos. This is due to the fact that participants kept their eyes closed for some moments while watching unpleasant videos (see Fig. 4.13(c)). Pupillary response's standard deviation is also shown to be higher during neutral scenes (see Fig. 4.13(d)).

The results have shown that it is possible to accurately recognize emotions with a user-independent approach. The classification accuracy measures are summarized in Table 4.7. The traditional F-score which combines precision and recall by their harmonic mean was also computed for each emotion category to give an overall evaluation of classification performance (Equation 4.11). The F1 score varies between zero and one. The random level is 0.5 for binary classification and balanced classes; values closest to 1 indicate a better performance.

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{4.11}$$

Precision and recall can be only defined for one class; hence, the F1 scores were calculated from the results of one versus all classification schemes for each class separately. As a result, the expectation of F1 scores of a uniform random classifier are calculated and given in Table 4.7. The classification rates of both three class classifications are defined as the percentage of correctly classified samples.

For the SVM classifier, the size of the kernel, γ, was selected between $[0.01, 10]$, based on the average F1 score using a 20-fold cross validation on the training set. The C parameter that regulates the tradeoff between error minimization and margin maximization is empirically set to 1. Classifications were first performed with the goal of recovering the three classes with a leave-one-participant-out cross validation scheme. Regarding the single modality classification



(a) Eye blinking rate, arousal

(b) Approach time ration, valence

(c) Blink depth, valence

(d) SD of pupil diameter, valence

Figure 4.13: Box plots of four different gaze data features in three emotional conditions. (a) Eye blinking rate for arousal classification (b) Approach time ratio for valence classification (c) Blink depth, average blink time, for valence classification (d) STD of pupil diameter for valence classification. One way ANOVA results showed a significant difference between features mean of different classes ($p < 0.05$)

of arousal and valence in three classes, we obtained 62.1% and 50.5% accuracy from EEG signals and 71.1% and 66.6% accuracy from eye gaze data (see Table 4.7). Although EEG classification results are inferior to the eye gaze data, they are comparable to the state of the art classification rates considering the inter-participant classification scheme [127, 61].

The FLF did not improve the best single modality, gaze data, results. However, the modality fusion strategy using the DLF improved the best SVM classification rate for arousal up to 76.4%. The DLF did not underperform for valence classification. To test the statistical significance of the classification performance, a paired t-test was used to compare F1 scores of the DLF on one side and the self reports and the best single modality, eye gaze data, on the other side. The F1 scores from each participant's samples were compared and the improvement over arousal classification comparing to eye gaze data and self reports were found significant ($p < 0.01$). However, the difference between the eye gaze, DLF, and self reports F1 scores on valence classification was not found statistically significant. The confidence levels of the classification results from the two modalities were added to find the class with the highest summed confidence.

The confusion matrices for each modality show how they performed on each emotion category (Tables 4.8.a - 4.8.j). In these confusion matrices, the row represents the classified label, and each column represents the ground truth. Only for activated category EEG classification performed

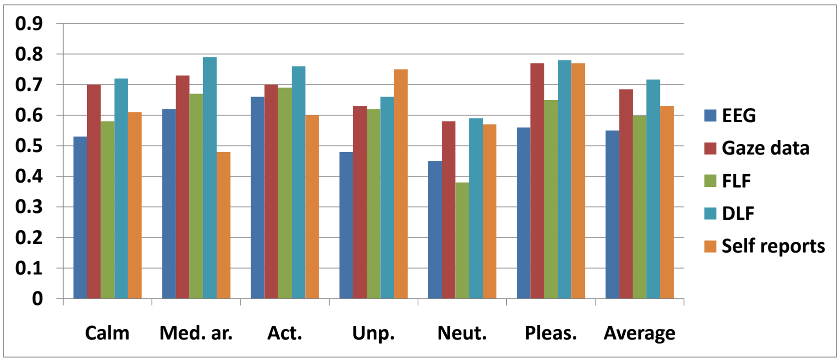


Figure 4.14: This bar chart shows the F1 score for classification results of each class from different modalities.

as well as gaze data modality. However, the fusion of both with the exceptions of neutral valence class outperformed gaze data results (see tables 4.8.a - 4.8.d and 4.8.f - 4.8.i). The DLF outperformed the feature level fusion for all categories except unpleasant (see tables 4.8.c, 4.8.d, 4.8.h, and 4.8.i).

The use of different stimuli and emotion classes make it difficult to directly compare the results to a similar work. Here, we compare the obtained accuracy with the most similar existing works. Kolodyazhniy et al. [64] used videos extracted from movies, in a user independent strategy, to classify three classes; namely, neutral, sadness and fear. They obtained 77.5% recognition rate from peripheral physiological responses while introducing random startles. Their results on three classes are almost at the same level of our arousal classification performance. In a gaming protocol, Chanel et al. [127] achieved the accuracy of 63% in a user-independent approach on the recognition of three classes; namely, boredom, engagement, anxiety. These three classes can be translated to our three arousal levels. Our results are inferior to the ones by Lisetti and Nasoz [60] on six classes. However, we did not only select the responses from the highly emotional moments in each scene or excerpt and our scene segmentation was independent of the emotional

Table 4.7: The classification rate and F1 scores of emotion recognition for different modalities.

Modality	**Classification rate**		**Average F1**	
dimension	**arousal**	**valence**	**arousal**	**valence**
EEG	62.1%	50.5%	0.60	0.50
Eye gaze	71.1%	66.6%	0.71	0.66
FLF	66.4%	58.4%	0.65	0.55
DLF	76.4%	68.5%	0.76	0.68
Self-reports with SAM manikins	55.7%	69.4%	0.57	0.70
Random level	33.3%	33.3%	0.36	0.40

Table 4.8: Confusion matrices of different classification schemes (row: classified label; column: ground truth). The numbers on the first row and the first column of tables a, b, c, d and e represents: 1. calm, 2. medium aroused, 3. activated and for tables f, g, h, i, and j represents: 1. unpleasant 2. neutral valence 3. pleasant. The confusion matrices relate to classification using (a, f) EEG signals (b, g) Eye gaze data (c, h) FLF (d, i) DLF (e, j) Self reports

Arousal

	1	2	3
1	44	15	11
2	22	111	39
3	30	60	135

(a) EEG

	1	2	3
1	60	7	9
2	10	136	40
3	26	43	136

(b) Eye gaze data

	1	2	3
1	49	14	10
2	15	117	31
3	32	55	144

(c) FLF

	1	2	3
1	62	6	8
2	8	146	28
3	26	34	149

(d) DLF

	1	2	3
1	63	35	11
2	24	88	65
3	9	63	109

(e) Self reports

Valence

	1	2	3
1	87	56	52
2	23	52	15
3	54	31	97

(f) EEG

	1	2	3
1	108	46	26
2	36	77	12
3	20	16	126

(g) Eye gaze data

	1	2	3
1	139	83	56
2	7	36	10
3	18	20	98

(h) FLF

	1	2	3
1	115	48	22
2	27	75	12
3	22	16	130

(i) DLF

	1	2	3
1	126	40	4
2	38	91	53
3	0	8	107

(j) Self reports

content.

The agreement between participants' self-reports and ground truth is shown in the confusion matrix given in Table 4.8.e and Table 4.8.j. The columns of this table represent how the videos of each class defined by ground truth were individually self-reported. For example, the first column of this table represent how many of the samples which were actually in class one were classified into different classes.

In order to measure the agreement between individual self-reports and the ground truth, the self-reported arousal and valence scores on nine point scale were translated into three levels. These levels were then treated like classified labels, and the classification rate was computed. This was done by considering that the goal of each participant is to label a video clip by the correct label, the most common tag. The classification rate for individually self-reported labels was 55.7% for arousal which is inferior to the worst classifier's result. Although, looking at the inter-annotation agreement, participants found that it easier to self-report pleasantness, the classification rate for valence is not significantly lower than the self-report rate. Therefore, the accuracy of obtained tags via classification is comparable to the individually reported labels.

Fig. 4.14 summarizes the comparison of different classification strategies showing the F1 scores for each category and on average. Looking at the bars on the most right side of the chart, only EEG results are inferior to the explicit self-report agreements using self assessment manikins.

4.4 Emotion Recognition in Response to Music Videos [4]

4.4.1 Overview

In this study, we present an emotion detection method for music videos using central and peripheral nervous system physiological signals. A set of 40 music clips eliciting a broad range of emotions were first selected. After extracting the one minute long, emotional highlight of each video, they were shown to 32 participants while their physiological responses were recorded. Participants self-reported their felt emotions after watching each clip by means of arousal, valence, dominance, and liking ratings. The physiological signals included electroencephalogram, galvanic skin response, respiration pattern, skin temperature, electromyograms and blood volume pulse using plethysmograph. Emotional features were extracted from the signals. The details on the experimental setup and protocol are given in detail in Section 3.2.2.3.

4.4.2 Emotional Features

The following peripheral nervous system signals were recorded: GSR, respiration amplitude, skin temperature, electrocardiogram, blood volume by plethysmograph, electromyograms of Zygomaticus and Trapezius muscles, and EOG. GSR provides a measure of the resistance of the skin by positioning two electrodes on the distal phalanges of the middle and index fingers. A plethysmograph measures blood volume in the participant's thumb. Skin temperature and respiration were recorded since they vary with different emotional states.

Regarding the EMG signals, the Trapezius muscle (neck) activity was recorded to investigate possible head movements during music listening. The activity of the Zygomaticus major was also monitored, since this muscle is activated when the participant laughs or smiles.

Most of the power in the spectrum of an EMG during muscle contraction is in the frequency range between 4 to 40 Hz. Thus, the muscle activity features were obtained from the energy of EMG signals in this frequency range for the different muscles. The rate of eye blinking is another feature, which is correlated with anxiety. Eye-blinking affects the EOG signal and results in easily detectable peaks in that signal.

All the physiological responses were recorded at a 512Hz sampling rate and later down-sampled to 256Hz to reduce processing time. The trend of the ECG and GSR signals was removed by subtracting the temporal, low frequency drift. The low frequency drift was computed by smoothing the signals on each ECG and GSR channels with a 256 points moving average.

In total 106 features were extracted from peripheral physiological responses based on the proposed features in the literature [61, 8, 107, 58, 10] (see also Table 4.9).

From the EEG signals, power spectral features were extracted. The logarithms of the spectral power from theta (4-8 Hz), slow alpha (8-10 Hz), alpha (8-12 Hz), beta (12-30Hz) and gamma (30+ Hz) bands were extracted from all 32 electrodes as features. In addition to power spectral

[4] This study was done in collaboration with, Sander Koelstra, Christian Mühl, Ashkan Yazdani, and Jong-Seok Lee in the context of Petamedia European network of excellence.

features the difference between the spectral power of all the symmetrical pairs of electrodes on the right and left hemisphere was extracted to measure the possible asymmetry in the brain activities due to emotional stimuli. The total number of EEG features of a trial for 32 electrodes is 216. Table 4.9 summarizes the list of features extracted from the physiological signals.

Table 4.9: Features extracted from EEG and physiological signals.

Signal	**Extracted features**
GSR	average skin resistance, average of derivative, average of derivative for negative values only (average decrease rate during decay time), proportion of negative samples in the derivative vs. all samples, number of local minima in the GSR signal, average rising time of the GSR signal, 10 spectral power in the [0-2.4]Hz bands, zero crossing rate of Skin conductance slow response (SCSR) [0-0.2]Hz, zero crossing rate of Skin conductance very slow response (SCVSR) [0-0.08]Hz, SCSR and SCVSR mean of peaks magnitude
BVP	Average and standard deviation of HR, HRV, and inter beat intervals, energy ratio between the frequency bands [0.04-0.15]Hz and [0.15-0.5]Hz, spectral power in the bands ([0.1-0.2]Hz, [0.2-0.3]Hz, [0.3-0.4]Hz), low frequency [0.01-0.08]Hz, medium frequency [0.08-0.15]Hz and high frequency [0.15-0.5]Hz components of HRV power spectrum.
Respiration pattern	band energy ratio (difference between the logarithm of energy between the lower (0.05-0.25Hz) and the higher (0.25-5Hz) bands), average respiration signal, mean of derivative (variation of the respiration signal), standard deviation, range or greatest breath, breathing rhythm (spectral centroid), breathing rate, 10 spectral power in the bands from 0 to 2.4Hz, average peak to peak time, median peak to peak time
Skin temperature	average, average of its derivative, spectral power in the bands ([0-0.1]Hz, [0.1-0.2]Hz)
EMG and EOG	eye blinking rate, energy of the signal, mean and variance of the signal
EEG	theta, slow alpha, alpha, beta, and gamma PSD for each electrode. The spectral power asymmetry between 14 pairs of electrodes in the four bands of alpha, beta, theta and gamma.

4.4.3 Emotion classification

In this section, we present the methodology and results of single-trial classification of the responses to videos. Two different modalities were used for classification; namely EEG signals, and peripheral physiological signals.

Three different binary classification problems were posed: the classification of low/high arousal, low/high valence and low/high liking. The participants' ratings during the experiment were used to generate the ground truth. The ratings for each of these scales are thresholded into two classes (low and high). On the nine points scale, the hard threshold was simply set to five which is in the middle of the minimum and maximum values. For some participants, this led to unbalanced classes. To show to what extent the classes were unbalanced the average and standard deviation of percentage of samples belonging to the high class were calculated. The mean and standard deviation (over participants) of the percentage of videos belonging to the high class per rating scale are: arousal ($M = 59\%,\ STD = 15\%$), valence ($M = 57\%,\ STD = 9\%$) and liking ($M = 67\%,\ STD = 12\%$).

In order to consider the unbalanced classes for evaluation, we reported the F1-score, which is commonly employed in information retrieval and takes the class balance into account, contrary to the mere classification rate. A naïve Bayes classifier was employed which is a simple and generalizable classifier and is able to deal with unbalanced classes in small training sets.

First, the features for the given modality were extracted from each trial or response to each video. Then, for each participant, the F1 scores were used to evaluate the performance of emotion classification in a leave-one-out cross validation scheme. At each step of the cross validation, one trial was used as the test-set and the of the trial served as the training set.

We used Fisher's linear discriminant J for feature selection:

$$J(f) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2} \tag{4.12}$$

where μ_i and σ_i are the mean and standard deviation for feature f from class i. We calculated this criterion for each feature and then applied a threshold to select the maximally discriminating ones. This threshold was empirically set to 0.3.

A Gaussian naïve Bayes classifier was used to classify the test-set as low/high arousal, valence or liking. The naïve Bayes classifier G assumes independence of the features and is given by:

$$G(f_1, .., f_n) = \underset{c}{\operatorname{argmax}}\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c) \tag{4.13}$$

where F is the set of features and C the classes. $p(F_i = f_i|C = c)$ is estimated by assuming Gaussian distributions of the features and modeling these from the training set.

4.4.4 Experimental results

Table 4.10 shows the average recognition rates and F1-scores (average F1-score for both classes) over participants for each modality and each rating scale. We compared the results to the expected values (analytically determined) of random classifiers, voting according to the majority class in the training data, and voting for each class with the probability of its occurrence in the training data. For determining the expected values of majority voting and class ratio voting, we used the class ratio of each participant's feedback during the experiment. These results are slightly over-estimated since, in reality, the class ratio would have to be estimated from the training set in each fold of the leave-one-out cross-validation.

Voting according to the class ratio gives an expected F1-score of 0.5 for each participant. An independent one-sample t-test was performed to test the significance of the difference between the F1 scores of all participants and random level F1 score, 0.5. According to the results reported in Table 4.10, 8 out of the 9 obtained F1-scores are significantly superior to the class ratio

Table 4.10: Average accuracies (ACC) and F1-scores (F1, average of score for each class) over participants. Stars indicate whether the F1-score distribution over subjects is significantly higher than random level, 0.5, according to an independent one-sample t-test ($** = p < .01$, $* = p < .05$). For comparison, expected results are given for classification based on random voting, voting according to the majority class and voting with the ratio of the classes.

	Arousal		**Valence**		**Liking**	
Modality	**ACC**	**F1**	**ACC**	**F1**	**ACC**	**F1**
EEG	0.62	0.58**	0.58	0.56**	0.55	0.50
Peripheral	0.57	0.53*	0.63	0.61**	0.59	0.54**
Random	0.50	0.48	0.50	0.49	0.50	0.48
Majority class	0.64	0.39	0.59	0.37	0.67	0.40
Class ratio	0.56	0.50	0.52	0.50	0.59	0.50

random baseline. The exception is the classification of liking using EEG signals ($p = 0.068$). However, looking at the F1 this voting scheme does not perform better. Overall, classification using EEG and peripheral are not significantly different ($p = 0.41$) (tested using a two-sided repeated samples t-test over the concatenated results from each rating scale and participant).

The modalities can be seen to perform moderately complementarily, where EEG scores best for arousal, peripheral for valence. The results on valence dimension classification performed best, followed by liking and then arousal. Although the classification results are significantly higher than random level, they are still far from being ideal for an emotional tagging application. This can be due to a noisy ground-truth which is based on self-reports and low intensity emotions. Signal noise, individual physiological differences and low number of samples are the other possible causes of the relatively poor classification accuracy.

4.5 Discussions

In this section, we discuss limitations of the conducted studies and provide the open issues. Physiological responses can vary due to non-emotional changes, such as circadian rhythms, ambient temperature, body posture and other psychophysiological factors such as attention, anticipation and mental effort [128]. Emotion recognition from bodily responses is therefore, limited by controlling different contextual factors. Moreover, like other similar works [64], the generalization of the results are limited by the videos shown to the participants.

The inter-annotation agreement for arousal self-reports in general is lower comparing to keyword based self-assessments. In a real-case scenario for an explicit tagging system, using words will be easier for an ordinary user and leads to higher between participant agreement in comparison to arousal and valence reported by Self Assessment Manikin (SAM). However, emotional keywords are difficult to translate and might not exist with the same meaning in different languages [35]. Emotion detection can overcome those difficulties with keeping the accuracy at the same level. In these studies, participants were asked to explicitly choose an emotional indicator to form the ground truth. In a real application, with the existence of a reliable user-independent emotion recognition method, the self-reporting phase can be eliminated.

In the first study (see Section 4.2.1), the results of a participant dependent emotion detection using regression (RVM) using peripheral physiological signals showed the feasibility of using these signals for video affect detection. These results were shown further useful by the results obtained by Kierkels et al. [129, 13] for implicit tagging using the regression outputs.

In the second study (see Section 4.3), the first study was expanded to include more participants and a participant independent approach was taken. Arousal labels were on average detected with higher accuracy using EEG signals comparing to valence labels. This might be due to higher visual and auditory variance of the arousal variant videos comparing to valence variant ones. Exciting scenes usually contain fast movements and loud noises which manifest themselves both in EEG signals and pupillary responses, whereas the difference between pleasant and unpleasant responses can be hidden in the semantics. The direct bodily responses to different stimuli can increase the variance in responses and improve the emotion recognition results. For example, faster changing video induces a different response in occipital cortex comparing to a more static video.

The DLF superior classification rate for arousal and its similar performance for valence classification shows that the proposed emotion classification can replace the self-reporting of single participants for detecting popular, emotional tags for this dataset. These popular, emotional tags are defined by the emotions felt by the majority of users watching the same video. After detecting emotional classes, they can be stored along other metadata attached to each video. Emotional labels can be converted to scores for arousal and valence for each video. The emotional scores can be then used, like in the image recommendation applications [12], to improve a video recommender's performance. In future works, the recognized emotional labels should be added as features to a video recommendation system to study the effect of introducing emotional labels on those systems. This effect can be determined by assessing users' satisfaction from a recommendation or retrieval system with and without emotional information. The emotion detection can be also used indirectly as a tool to detect topical relevance in information retrieval systems [89].

In the second experiment, the determination of ground truth is based on the participants' feedback in the online preliminary experiment. In order to measure the agreement between the popular responses of the two separate populations of the preliminary assessments and the experiments, we computed the median valence and arousal reported during the experiment and compared the labels based on the recorded participants' popular response. Only three arousal labels out of 40 labels were changed from activated to medium arousal or vice-versa. No valence label has changed when comparing two populations. This is due to valence dimension's higher inter-annotator agreement.

These studies still has open issues that need to be considered in the future. In a real case scenario, any new user will need to have few minutes of signals recorded to provide reliable values for feature normalization due to the utilized normalization. The estimation of pupillary response to illumination is an open issue which needs more investigation. Although we assumed that the lighting pupillary responses are similar between participants, there will be a large difference in case of introducing users from a different age group. Therefore, the parameters of a reliable model for pupillary reflex to lighting, such as [119], should be determined before introducing a new user to the system. Alternatively, the real time lighting effect similar to [71] can be employed to remove the lighting effect. A larger video set and a larger number of participant can be considered to increase the generalization of the developed emotion classifier in response to videos.

In the last experiment on music videos (see Section 4.4), the results were poor comparing to the second experiment. The inferior results can be caused by three factors. First, the emotions elicited by music videos were weaker comparing to movie scenes. Second, the classification was participant dependent, and the number of samples were smaller. Finally, eye gaze, which contributed the most in the superior results for the second experiment, was not recorded.

In these studies, the length of the experimental session limited number of videos we could show to each participant. The number of participants in these studies were large and diverse enough comparing to similar studies [8, 61]. However, the population only consisted of young students, which limits the trained algorithm to this particular group. These limitations might worsen the results in case of introducing a new genre of video which was not present in the current video set. To generalize and train such a system, the recruited participants should be as close as possible to the target audience of the video retrieval or recommendation systems.

Emotions can co-occur and/or last for very short moments. This puts using a single emotional label for a video clip under question. In order to address this issue, self-reporting should include the possibility to indicate different emotions and their degree of strength. This is possible by using questionnaires such as positive and negative affect schedule (PANAS) [130] or Geneva emotion wheel [15]. However, these emotional self-reporting methods are more complex and make the experiment longer. In future work, multiple or co-occurring emotions should be assessed using a more sophisticated self reporting tool.

4.6 Summary

In this chapter, the methodology, evaluation criteria and results of emotion detection for three different experiments were presented. First, a personalized regression based emotion detection was proposed, implemented and evaluated on eight participants. The second study was a participant independent multi-modal emotion recognition in response to videos. The third study was conducted to detect emotions in response to music videos in a participant dependent approach. In the next Chapter, the methods and results on the estimation of the emotion from the content itself will be given.

Chapter 5

Content Analysis for Affect Estimation

5.1 Overview

In this chapter, the developed methodology and the results concerning the evaluations of the methods estimating emotions from video content are presented. Although, emotions are personal and vary depending on the contextual factors, there exists a more common emotion which is induced by a video excerpt or a movie scene. The common emotion eliciting characteristic is an crucial factor for the popularity of commercial movies. Different genres of movies elicit certain emotions in the majority of audience, e.g., comedy, drama, horror. The present study is focused on movies because they represent one of the most common and popular types of multimedia content. An affective representation of scenes will be useful for tagging, indexing and highlighting of important parts in a movie.

5.1.1 Audio-Visual Features

Audio and video were first demultiplexed. Music videos were encoded into the Moving Picture Experts Group (MPEG)-1 format to extract motion vectors and I-frames for further feature extraction. The video stream of the music clips has been segmented at the shot level using the method proposed in [131].

Movie scenes have been segmented at the shot level using the OMT shot segmentation software [132]. Video clips were encoded into MPEG-1 format to extract motion vectors and I frames for further feature extraction. We used the OVAL library (Object-based Video Access Library) to capture video frames and extract motion vectors.

Sound has an significant impact on user's affect. For example, according to the findings of Picard [133], loudness of speech (energy) is related to evoked arousal, while rhythm and average pitch are related to valence. The audio channels of the movie scenes were extracted and encoded into monophonic information (MPEG layer 3 format) at a sampling rate of 48 kHz, and their amplitude range was normalized in [-1, 1]. All of the resulting audio signals were normalized to the same amplitude range before further processing. A total of 53 low-level audio features were determined for each of the audio signals. These features, listed in Table 5.1, are commonly used in audio and speech processing and audio classification [134, 135].

Wang et al [50] demonstrated the relationship between audio type's proportions and affect, where these proportions refer to the respective duration of music, speech, environment, and silence in the audio signal of a video clip. To determine the three principal audio types (music, speech, and environment), silence was first identified by comparing the audio signal energy of each sound sample with a pre-defined threshold empirically set at $5 \times 10^{-}7$. After removing silence, the remaining audio signals were classified into three classes using a SVM. We implemented a three class audio type classifier using support vector machines (SVM with polynomial kernel) operating on audio low-level features in a time window of one second. Despite some classes overlapping (e.g., presence of a musical background during a dialogue), the classifier was usually able to recognize the dominant audio type. The SVM was trained utilizing more than 3 hours of audio, extracted from movies and labeled manually. The classification results were used to form 4 bins (3 audio types and silence) normalized histogram; these histogram values were used as affective features for the affective representation. Mel Frequency Cepstral Coefficients (MFCC), Formants and the pitch of audio signals were extracted using the PRAAT software package [136].

From a movie director's point of view, lighting key [50, 137] and color variance [137] are important parameters to evoke emotions. We therefore, extracted lighting key from frames in the Hue, Saturation, Value (HSV) space by multiplying the average value (V in HSV) by the standard deviation of the values. Color variance was obtained in the CIE LUV color space by computing the determinant of the covariance matrix of L, U, and V.

The average shot change rate and shot length variance were extracted to characterize video rhythm. Hanjalic et al. [7] showed the relationship between video rhythm and affect. Fast object movements in successive frames are also an effective factor to evoke excitement. To measure this factor, the motion component was defined as the amount of motion in consecutive frames computed by accumulating magnitudes of motion vectors for all B and P frames.

Colors and their proportions have an effect to elicit emotions. In order to use colors in the list of video features, 20 bin color histograms of hue and lightness values in the HSV space were computed for each I frame and subsequently averaged over all frames. The resulting averages in the 20 bins were used as video content-based features. The median of L value in Hue, Saturation, Lightness (HSL) space was computed to obtain the median lightness of a frame. Shadow proportion or the proportion of the dark area in a video frame is another feature which relates to affect [50]. Shadow proportion is determined by comparing the lightness values in HSL color space with an empirical threshold. Pixels with lightness level below this threshold (0.18 [50]) are assumed to be dark and in the shadow in the frame.

Visual excitement is a measure of the average pixel's color change between two consecutive frames [50]. It is defined as the average change between the CIE Luv histograms of the 20×20 blocks of two consecutive frames. In our case, this visual excitement feature was implemented from the definition given in [50] for each key frame. Two visual cues were also implemented to characterize these key frames. The first one, called visual detail, is used as an indicator of the distance from the camera to the scene and differentiates between close-ups and long-shots. The visual detail was computed by the average gray level co-occurrence matrix (GLCM) [50]. The other visual cue is the grayness which was computed from the proportion of the pixels with saturation below 20%, which is the threshold determined for colors that are perceived as gray [50].

Table 5.1: Low-level features extracted from audio signals.

Feature category	**Extracted features**
MFCC	MFCC coefficients (13 features) [135], Derivative of MFCC (13 features), Autocorrelation of MFCC (13 features)
Energy	Average energy of audio signal [135]
Formants	Formants up to 5500Hz (female voice) (five features)
Time frequency	Spectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio, [135, 134]
Pitch	First pitch frequency
Zero crossing rate	Average, Standard deviation [135]
Silence ratio	Proportion of silence in a time window [138]

5.1.2 Textual features

Two textual features were also extracted from the subtitles track of the movies. According to [139] the semantic analysis of the textual information can improve affect classification. As the semantic analysis over the textual data is not the focus of our work, we extracted simple features from subtitles by tokenizing the text and counting the number of words. These statistics have been used with the timing of the subtitles to extract the talking rate feature which is the number of words that had been spoken per second on the subtitles show time. The other extracted feature is the number of spoken words in a scene divided by the length of the scene, which can represent the amount or existence of dialogues in a scene.

A method based on the bag of words of strategy and estimating the affect elicited by content using Whissel Dictionary [80] is also proposed (see Section 5.3).

5.1.3 Continuous characterization of multimedia

A method estimating dimensional emotions are proposed. The dataset consisting of 64 movie scenes (details in Section 3.1.2.1) was used and shown to 8 participants. Their continuous emotional self reports served as the ground truth for a continuous affect estimation method. The affect was estimated using regression and evaluated using a leave one out cross validation strategy. A similar method detecting affect from physiological responses in Section 4.2.1 is used to estimate what emotion is going to be elicited showing videos.

First, the correlations between content features and self assessments were determined. Table 5.2 shows, for each participant, the features which had the highest absolute correlations with that participant's self-assessments of valence and arousal. The large variation between participants regarding which multimedia features have the highest absolute correlation value with their self assessment indicates the variance in individual preferences to different audio or video features. For instance, an increase in motion component leads to higher arousal for participant 8. For the same feature, increase in motion component resulted in lower valence for participant 5, which means that the participant had a negative feeling for exciting scenes with a large amount of movement in objects or background.

Moreover, The correlation between physiological responses and content features was studied. Table 5.3 shows, for all participants, the correlation coefficients between four different pairs of

Table 5.2: Content features with the highest absolute correlation with self assessments for participants 1 to 8.

Part.	**Arousal**	ρ	**Valence**	ρ
1	6th MFCC coefficient	0.44	15th bin of the Hue histogram (purplish)	-0.47
2	19th bin of the Hue histogram (purplish)	-0.47	Shadow proportion standard deviation	-0.51
3	8th MFCC coefficient	0.45	Last autocorrelation MFCC coefficient (standard deviation)	0.53
4	First autocorrelation MFCC coefficient (standard deviation)	0.44	3rd autocorrelation MFCC coefficient (standard deviation)	0.39
5	4th Derivative MFCC	0.35	Motion component	-0.47
6	11th autocorrelation MFCC coefficient	-0.37	5th bin of Luminance histogram	-0.39
7	12th MFCC coefficient	0.43	Color variance standard deviation	0.48
8	Motion component	0.40	Visual cue, detail	0.52

physiological features and multimedia features. These eight features have been chosen from the features which have significant correlation with self assessments and thus are more importance for affect characterization. The correlations show that the indicated physiological responses are significantly correlated with changes in multimedia content. This is for instance the case with the positive correlation between EMG Zygomaticus energy and key lighting of the video content: lighter scenes have a direct, positive effect on the Zygomaticus activity.

The arousal and valence were estimated using regressors. If $y(j)$ is the emotion, arousal or valence, felt by viewers and $\hat{y}$ is our estimation. In order to determine the optimum $\hat{y}$, the weights in Equation 4.1 were computed via a linear RVM from the Tipping RVM toolbox [114]. This procedure was applied on the user self assessed valence-arousal, $y(j)$, and on the feature-estimated valence-arousal, $\hat{y}(j)$, over all movie scenes. This procedure is performed two times for optimizing the weights corresponding to:

– multimedia features when estimating valence;

– multimedia features when estimating arousal.

The mean absolute error (E_{MAE}) and Euclidean distance (E_{ED}) are used to evaluate the regression results in the same was as Section 4.2.1.4.

E_{ED} and E_{MAE} values are shown in Table 5.4. It can in particular be observed that the average Euclidean distance results are all below random level (which is around 0.5). The E_{MAE} represents the distance of the determined emotion from the self assessed emotion in the dimen-

Table 5.3: The linear correlation ρ values between multimedia features, and physiological features which are significantly correlated with self assessments (participants 1 to 8).

Part.	EMG Zygomaticus (Sum of absolute values) / Key lighting	GSR power spectral density 0-0.1 Hz band / Standard deviation (SD) of the first autocorrelation of MFCC	BVP spectral density 0.1-0.2 Hz band / Shot length variation	EMG Zygomaticus (Sum of absolute values) / Visual cue, details
1	-	0.73	0.54	-
2	0.71	0.72	0.84	0.53
3	0.35	0.71	0.89	0.33
4	0.51	0.50	0.78	0.43
5	0.39	0.63	0.88	0.36
6	0.41	0.63	0.91	0.30
7	0.46	0.76	0.86	0.40
8	0.64	0.55	0.82	0.56

sions of arousal or valence. E_{MAE} is thus useful to compare each dimension's results. The E_{MAE} of arousal and valence shows that valence determination was more precise than arousal determination. The superior valence results might have been caused by the easier valence assessment and therefore, a more precise self assessment on valence.

5.2 Movie Scene Classification Using A Bayesian Framework

Emotions that are elicited in response to a video scene contain valuable information for multimedia tagging and indexing. The novelty of the method proposed in this Section is to introduce a Bayesian classification framework for affective video tagging that allows taking contextual information into account. We think that using the existing online metadata can improve the affective representation and classification of movies. Such metadata, like movie genre, is available on the Internet (e.g., the internet movie database http://www.imdb.com). Movie genre can be exploited to improve an affect representation system's inference about the possible emotion which is going to be elicited in the audience. For example, the probability of a happy scene in a comedy certainly differs from that in a drama. Moreover, the temporal order of the evoked emotions, which can be modeled by the probability of emotion transition in consecutive scenes, is also expected to be useful for the improvement of an affective representation system.

It is shown here how to benefit from the proposed priors in a Bayesian classification framework. Affect classification was done for a three labels scene classification problem, where the

Table 5.4: Mean absolute error (EMAE), and Euclidean distance (EED) between estimated valence-arousal grades and self assessments (participants 1 to 8).

Part.	E_{MAE} **Arousal estimated from MCA**	E_{MAE} **Valence estimated from MCA**	E_{ED} **MCA**
1	0.18	0.13	0.24
2	0.17	0.15	0.23
3	0.15	0.04	0.21
4	0.15	0.13	0.21
5	0.15	0.15	0.24
6	0.15	0.12	0.22
7	0.12	0.12	0.18
8	0.13	0.07	0.16
Average	0.15	0.11	0.21
Random level	**~0.4**	**~0.4**	**~0.5**

labels are "calm", "positive excited", and "negative excited". Ground truth was obtained through manual annotation with a FEELTRACE-like [48] annotation tool with the self-assessments serving as the classification ground-truth. The usefulness of priors is shown by comparing classification results with or without using them.

In our proposed affective indexing and retrieval system, different modalities, such as video, audio, and textual data (subtitles) of a movie will be used for feature extraction. Fig. 5.1 shows the diagram of such a system. The feature extraction block extracts features from the three modalities and stores them in a database. Then, the affect representation system fuses the extracted features, the stored personal information, and the metadata to represent the evoked emotion. For a personalized retrieval, a personal profile of a user (with his/her gender, age, location, social network) will help the affective retrieval process.

5.2.1 Dataset and annotations

21 commercial and famous movies were first gathered and fully annotated. A list of the movies in the dataset and their corresponding genre is given in Table 5.5.

FEELTRACE is a self-assessment tool which was proposed by Cowie et al. [48] to assess emotion in the valence-arousal space. In this assessment tool, the coordinates of a pointer manipulated by the user are continuously recorded during the show time of the stimuli (video, image, or external source) and used as the affect indicators. Inspired by this tool designed for psychological studies, an affective self reporting tool has been implemented to assess emotion during the watching of a video. The emotion is recorded as the coordinates of the pointer on the

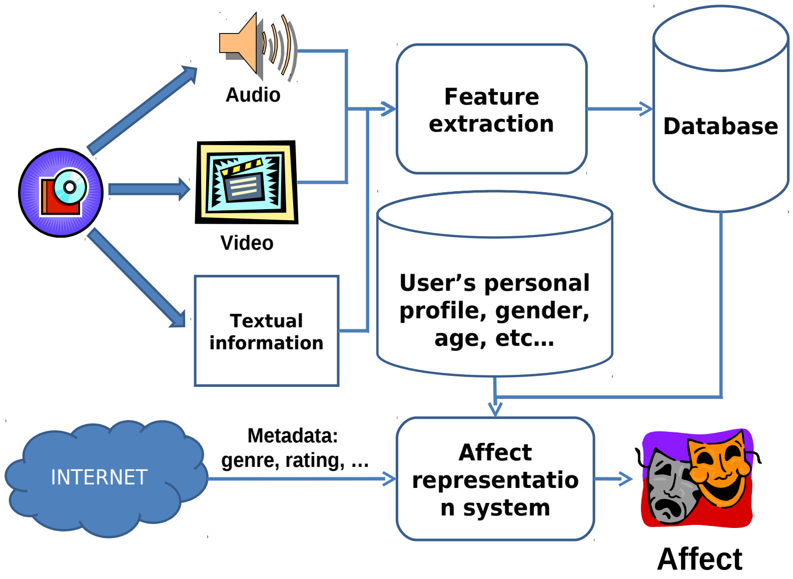

Audio
Video
Textual information
Feature extraction
Database
User's personal profile, gender, age, etc...
INTERNET
Metadata: genre, rating, ...
Affect representatio n system
Affect

Figure 5.1: A diagram of the proposed video affective representation.

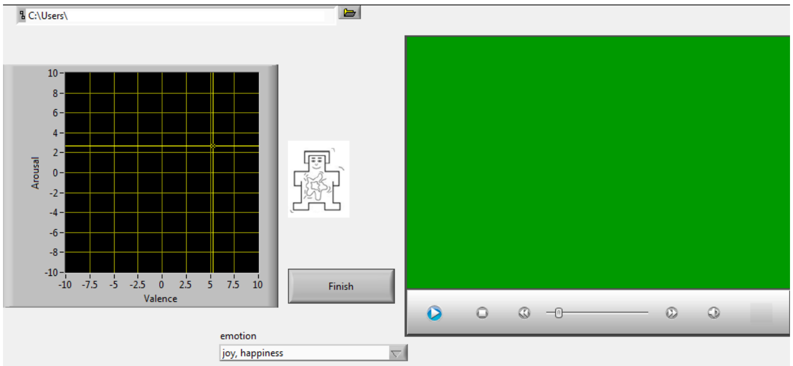
Table 5.5: List of movies in the dataset.

Drama movies	**Comedy movies**
The pianist, Blood diamond, Hotel Rwanda, Apocalypse now, American history X, Hannibal	Man on the moon, Mr. Bean's holiday, Love actually, Shaun of the dead, Shrek
Horror movies	**Action movies**
Silent hill, 28 days later, Rungu (Japanese), The shining	Man on Fire, Kill Bill Vol. 1, Kill Bill Vol. 2, Platoon, The thin red line, Gangs of New York

click event.

One participant annotated the movies so as to indicate at which times his felt emotion has changed. Thus, the valence and arousal values received from the participant should occur when there was a change in the participant's emotion. The participant was asked to indicate at least one point during each scene not to leave any scene without assessment. Continuous annotation of the movies is a time consuming process; hence the participant was asked to annotate at most two movies per day of different genres.

A set of SAM manikins [40] are generated for different combinations of arousal and valence to help the user understand the emotions related to the regions of valence-arousal space. E.g. the positive, excited manikin is generated by combining the positive manikin and the excited manikin. A preview of the annotation software is given in Fig. 5.2.


C:\Users\

emotion: joy, happiness

Figure 5.2: Snapshot of the affective annotation software which is implemented in LABVIEW. The positive excited manikin can be seen in the central part of the display.

5.2.2 Arousal estimation with regression on shots

Informative features for arousal estimation include loudness and energy of the audio signals, motion component, visual excitement and shot duration. Using a method similar to Hanjalic et al. [7] and to the one proposed in [10], the felt arousal from each shot is computed by a regression of the content features (see Section 5.1.1 for a detailed description). In order to find the best weights for arousal estimation using regression, a leave one movie out strategy on the whole dataset was used, and the linear weights were computed by means of a RVM from the RVM toolbox provided by Tipping [114]. The RVM is able to reject uninformative features during its training hence no further feature selection was used for arousal determination. Equation 5.1 shows how N_s audio and video based features z_i^k of the k-th shot are linearly combined by the weights w_i to compute the arousal $\hat{a}_k$ at the shot level.

$$\hat{a}_k = \sum_{i=1}^{N_s} w_i Z_i^k + w_0 \tag{5.1}$$

5.2.3 Scene classification

After computing arousal at the shot level, the average and maximum arousals of the shots of each scene are computed and used as arousal indicator features for the scene affective classification. During an exciting scene, the arousal related features do not all remain at their extreme level. In order to represent the highest arousal of each scene, the maximum of the shots' arousal was chosen to be used as a feature for scene classification.

The linear regression weights that were computed from our data set were used to determine the arousal of each movie's shots. This was done in such a way that all movies from the dataset except for the one to which the shot belonged to were used as the training set for the RVM. Any missing affective annotation for a shot was approximated using linear interpolation from the closest affective annotated time points in a movie.

It was observed that arousal has higher linear correlation with multimedia content-based features than valence. Valence estimation from regression is not as accurate as arousal estimation and therefore,valence estimation has not been performed at the shot level.

For the purpose of categorizing the valence-arousal space into three affect classes, the valence-arousal space was divided into the three areas shown in Fig. 5.3, each corresponding to one class. According to [57] emotions mapped to the lower arousal category are neither extreme pleasant nor unpleasant emotions and are difficult to differentiate. Emotional evaluations are shown to have a heart shaped distribution on valence-arousal space [57]. Hence, we categorized the lower half of the plane into one class. The points with an arousal of zero were counted in class 1, and the points with arousal greater than zero and valence equal to zero were considered in class 2. These classes were used as a simple representation for the emotion categories based on the previous literature on emotion assessment [61].

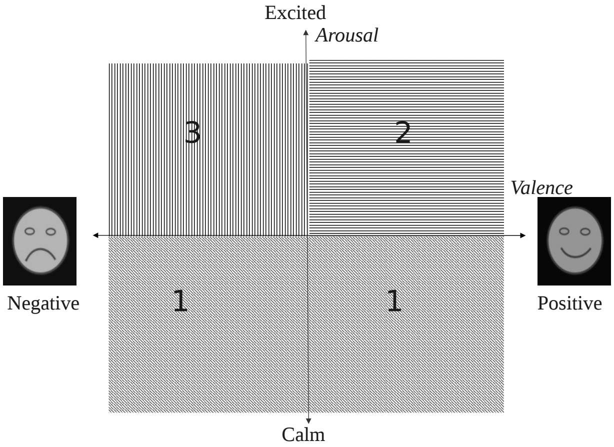


Figure 5.3: Three classes in the valence-arousal space are shown; namely calm (1), positive excited (2) and negative excited (3).

In order to characterize movie scenes into these affective categories, the average and maximum arousal of the shots of each scene and the low level extracted audio- and video- based features were used to form a feature vector. This feature vector, in turn, was used for the classification. The content features are listed in Section 5.1.1.

If the content feature vector of the j-th scene is x_j, the problem of finding the emotion class, $\hat{y}_j$, of this scene is formulated as estimating the $\hat{y}_j$ which maximizes the probability $p(y_j|x_j, theta)$ where θ is the prior information which can include the user's preferences and video clip's metadata. In this study one of the prior metadata (θ) we used is for instance the genre of the movie. Personal profile parameters can be also added to θ. Since in this study the whole affect representation is trained by the self report of one participant the model is assumed to be personalized for this participant. When the emotion of the previous scene is used as another prior the scene affect probability formula changes to $p(y_j|y_{j-1}, x_j, \theta)$. Assuming for simplification that the emotion of the previous scene is independent of the content features of the current scene this probability can be reformulated as:

$$p(y_j|y_{j-1}, x_j, \theta) = \frac{p(y_{j-1}|y_j, \theta)p(y_j|x_j, \theta)}{p(y_{j-1}|\theta)} \tag{5.2}$$

The classification problem is then simplified into the determination of the maximum value of the numerator of Equation 5.2, since the denominator will be the same for all different affect classes y_j. The priors are established based on the empirical probabilities obtained from the training data. For example, the occurrence probability of having a given emotion followed by any of the emotion categories was computed from the participant's self-assessments and for each genre. This allowed to obtain the $p(y_{j-1}|y_j, \theta)$. Different methods were evaluated to estimate the posterior probability $p(y_j|x_j)$. A naïve Bayesian approach which assumes the conditional probabilities are Gaussian was chosen as providing the best performance on the dataset; the superiority of this method can be attributed to its generalization abilities.

5.2.4 Results

5.2.4.1 Arousal estimation on shots

Fig. 5.4 shows a sample arousal curve from part of the film entitled "Silent Hill". Fig. 4 is a typical example of the obtained results on arousal estimation. The estimated affect curve, in the first half, fairly closely follows the self-assessment curve. This moreover shows the correlation between arousal related content features and participant's self-estimated affect. The participant's felt emotion was, however, not entirely in agreement with the estimated curve, as can for instance be observed in the second half of the plot. A possible cause for the discrepancy is the low temporal resolution of the self-assessment. Another possible cause is experimental weariness: after having had exciting stimuli for minutes, a participant's arousal might be decreasing despite strong movements in the video and loud audio. Finally, some emotional feelings might simply not be captured by low-level features; this would for instance be the case for a racist comment in a movie dialogue which evokes disgust for a participant. Without some form of semantic high level analysis of the movie script, the content features are unable to detect verbal behavior in movie scenes.

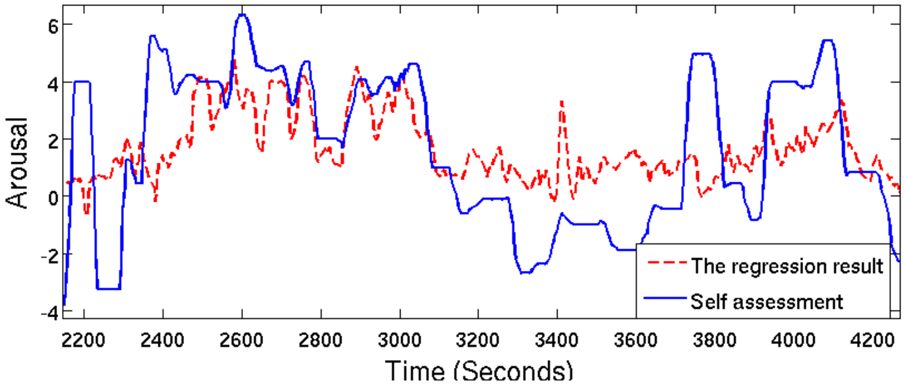


Figure 5.4: Five-points smoothed shot arousal curve (full line), and corresponding self-assessments (dashed line).

5.2.4.2 Scene classification results

The naïve Bayesian classifier results are shown in Table 5.6-a. A more complex SVM classifier was also evaluated on this dataset but did not demonstrate an improvement in terms of classification results. The superiority of the Bayesian classifier with priors over a more complex classifier such as the SVM shows that the improvement brought in through the use of adequate priors, can be higher than the one provided by a more complex classifier. Results achieved by such a support vector machine classifier with a linear kernel are reported in Table 5.6-e.

Table 5.6: Affective scene classification accuracies and F1 scores with different combinations of priors (on the left) and their confusion matrices (on the right). "1", "2", "3" correspond to the 3 classes "calm", "positive excited", and "negative excited". In the confusion matrices the rows are classified labels and the columns are ground truth.

(a) Naïve Bayesian				1	2	3
	F1	0.55	1	0.77	0.36	0.24
	Accuracy	0.56	2	0.17	0.36	0.22
			3	0.06	0.28	0.54

(b) Bayesian + time				1	2	3
	F1	0.56	1	0.76	0.35	0.25
	Accuracy	0.57	2	0.16	0.40	0.23
			3	0.08	0.25	0.52

(c) Bayesian + genre				1	2	3
	F1	0.60	1	0.87	0.44	0.36
	Accuracy	0.61	2	0.02	0.35	0.03
			3	0.11	0.21	0.61

(d) Bayesian + genre & time				1	2	3
	F1	0.63	1	0.83	0.31	0.35
	Accuracy	0.64	2	0.09	0.49	0.05
			3	0.08	0.20	0.60

(e) SVM linear kernel				1	2	3
	F1	0.56	1	0.77	0.32	0.21
	Accuracy	0.56	2	0.13	0.34	0.22
			3	0.11	0.33	0.57

The results obtained using the preceding emotion prior (temporal prior) are given in Table 5.6-b. The affect class of the previous scene, that is the preceding emotion, was obtained by using the same trained classifier to classify the preceding scene's feature vector. Using the temporal prior, the improvement in the F1 score and accuracy of the classifier showed that even in case of misclassification of the previous scenes the Bayesian classifier is robust enough to slightly improve the classification performance.

The genre prior was then included as the only prior; the classification results obtained are shown in Table 5.6-c. Finally, the best results were obtained using both the genre and temporal priors as can be seen in Table 5.6-d. The F1 score increased about 9 percent utilizing both priors in comparison to naïve Bayesian.

As with the temporal prior, the genre prior leads to better estimate of the emotion class. The classification accuracies of the first class "calm" and the third class, "negative excited" have been improved with this prior. Regarding the "calm" class, the reason for this improvement is that the genre has a clear impact on arousal, thus on "calm" vs. "aroused" classification (horror and action movies have higher arousal than drama and comedy). The determination of the second class, "positive excited", was only improved by utilizing the temporal prior and not by the genre prior. The reason is that positive excited emotions were spread among different genres in this training data. A sample of movie scenes classification along time is shown in Fig. 5.5. The evolution of classification results over consecutive scenes, when adding the time prior, shows that this prior allows correcting results for some samples that were misclassified using the genre prior only. For example on the 4th and 20th scene the classifier with time prior was able to find the correct class while the naïve Bayesian with only genre prior missed it. Moreover, adding the time prior did not change any correct classification of the naïve Bayesian classifier.

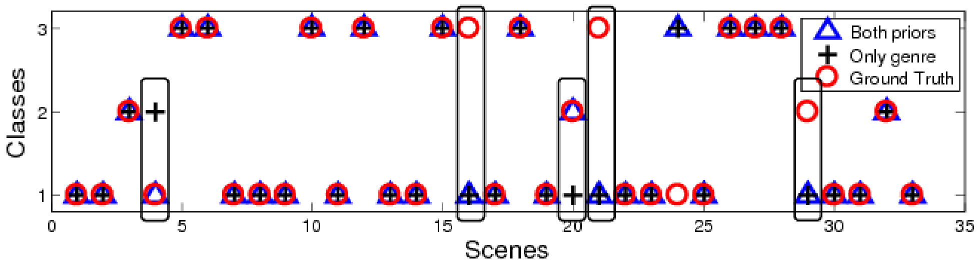


Figure 5.5: Classification results for consecutive scenes in a movie. The circle represents the target class, and the plus sign shows the results of the Naïve Bayesian classifier with genre prior and the triangle shows the results with both genre and time priors. The samples, which are misclassified by the Bayesian classifier with genre prior, are encircled.

One of the main drawbacks of the proposed approach is the low temporal resolution of affective annotations. It is impossible to guarantee a perfect continuous assessment and annotation of a movie without the user being distracted at times from the movie events. It is also non-realistic to expect an average user to be able to use psychological terms and consistent words to express his/her emotions. Using physiological signals or audio-visual recordings will help overcome these problems and facilitate this part of the work, by yielding continuous affective annotations without interrupting the user [10].

More prior information and semantic analysis of the movie's script (subtitles), as well as higher level features, are necessary to further improve affect representation. Priors can be personality related information that help in the definition of personal, affective profiles. An example of such pieces of information is a social network groups indicating people with the same taste, gender, ethnicity, age, etc. A larger dataset of movies with annotations from multiple participants with different backgrounds will therefore, enable us to examine more priors. It will also provide us with a better understanding of the feasibility of using group-wise profiles containing some affective characteristics that are shared between users.

w	Pleasantness	Activation	imagery
a	2.0000	1.3846	1.0
abandon	1.0000	2.3750	2.4
abandoned	1.1429	2.1000	3.0
abandonment	1.0000	2.0000	1.4
abated	1.6667	1.3333	1.2
abilities	2.5000	2.1111	2.2
ability	2.5714	2.5000	2.4
able	2.2000	1.6250	2.0
abnormal	1.0000	2.0000	2.4
.	.	.	.
.	.	.	.
.	.	.	.

Figure 5.6: The Whissel dictionary of affect in language. Each word had three ratings.

5.3 Video Characterization Using Text Analysis

Subtitles of movies carry valuable semantic information that can be used for affective understanding [139]. In this Section, the same dataset utilized in the previous Section, Section 5.2, was employed to study the usefulness of text analysis for affective characterization. Two different approaches were taken for affective understanding of movies based on subtitles. First, a shot level characterization based on Whissel Dictionary [80] and then a scene classification based on the bag of words strategy.

5.3.1 Shot level characterization using dictionary

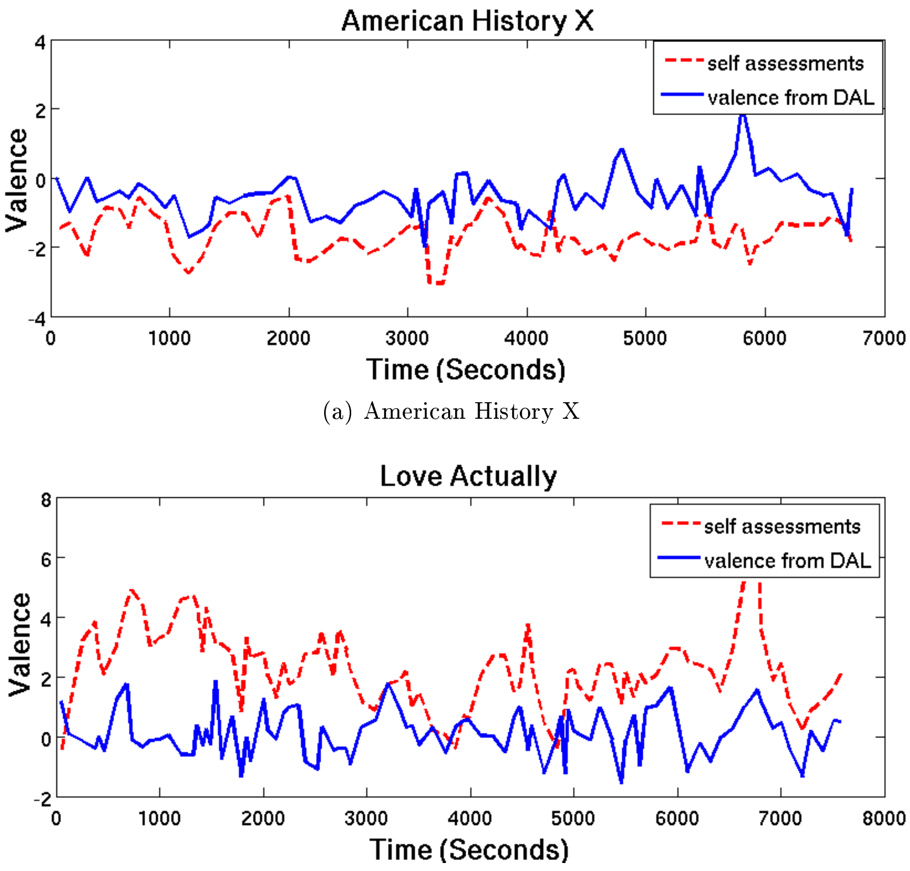
The dictionary of affect in language [80] contains 8743 words, which are rated by 200 volunteers in three dimensions; namely, pleasantness, activation and imagery which means whether a word gives a clear mental picture. Example of words and their ratings are shown in Fig. 5.3.1.

In order to characterize the pleasantness of movies in shot level, first the subtitles were tokenized, and then the words, which were spoken in each shot, were extracted. The pleasantness ratings of words were averaged which gave us the pleasantness in each shot. Two examples of obtained results with a self reported valence curve are shown in Fig. 5.7.

Although, in the first half of American History X (see Fig. 5.7(a)) the affect estimation follows the self assessments there are moments in which the estimated valence is not entirely in accordance with the self assessments, e.g., from seconds 5000 to 6000 in American History X and around 1000 seconds in Love actually (see Fig. 5.7).

5.3.2 Scene classification using text analysis

In this Section, a bag of words strategy is employed for classification of scenes into three classes of calm", "positive excited", and "negative excited" which is the same classification scheme that was used in Section 5.2. For this purpose, the subtitles were first tokenized, and the frequency of terms were calculated (after removing the stop words). In the proposed method, each affective class was dealt with like a document in text retrieval and every word spoken in a scene with



(a) American History X

(b) Love Actually

Figure 5.7: Valence estimation and self reports on shot level for two movies are shown.

unknown affect was presented as a query to a repository of indexed documents (classes). The similarity of the scene to the indexed documents was measured by first computing the normalized word frequencies in classes and comparing the normalized frequencies of the query in different classes. The class or document with the highest frequency of the query was identified as relevant and its class, e.g., “calm”, was identified as the class corresponding to that word. For example, the following sentence words were classified as:

“oh”	“you’re”	“crazy”	“i’m”	“just”	“bringing”
2	1	3	1	2	2

A majority vote over the query results decides the class of the whole scene, e.g., the above example is therefore, classified in class “2”. A ten fold cross validation was used for the classification of scenes based on a majority vote between the subtitle words. Three text processing schemes were applied on the dataset. First, simply by tokenizing and computing the frequencies. Second, after removing the English stop words and third after removing stop word and stemming. Te classification results are shown in Table 5.7. The results showed that removing stop words and stemming did not improve the results and only by using text and bag of words strategy it is possible to estimate the emotions with a rate which is significantly higher than a random guess. The text analysis combined with video and audio content features can further improve the MCA results.

5.4 Boredom Ranking

5.4.1 Dataset and features

The dataset selected for the developed corpus is Bill's Travel Project, a travelogue series called "My Name is Bill" created by the film maker Bill Bowles (http://www.mynameisbill.com/) (see Section 3.1.2.3). Each video is annotated by multiple annotators with boredom scores on nine point scale. The average boredom score given by participants of the preliminary study served as the ground truth for this benchmarking challenge. First 42 videos were released and used for training. The remaining 80 videos were served as the evaluation.

The dataset consists of information from different modalities; namely, visual information from video, speech transcripts, audio signals, titles and publication dates.

The low level content features were extracted from audio and video signals; namely, key lighting, color variance, motion component zero crossing rate, audio energy. A detailed description of the content features in Section 5.1.1 . Shot boundaries were detected using the method described in [131]. Video length, shot change rate and variation (standard deviation and skewness), number of shots and average shot length were extracted using the detected shot boundaries.

The speech transcripts were provided by a software implemented originally for speech recognition in meetings [140]. Using WordNet [141] the nouns and nouns which could describe a country, place or land were first extracted as location names. Each noun was checked to see if it has a Wikipedia page in English. The number of nouns indexed in Wikipedia was counted as information related feature. This was extracted as an indicator of the amount of information each transcript carries. The sum of the length of all nouns' Wikipedia pages was also extracted

Table 5.7: Affective scene classification accuracies and F1 scores with different combinations of priors (on the left) and their confusion matrices (on the right). "1", "2", "3" correspond to the three classes "calm", "positive excited", and "negative excited".

(a) Without preprocessing				1	2	3
	F1	0.48	1	0.35	0.21	0.23
	Accuracy	0.51	2	0.26	0.56	0.14
			3	0.39	0.23	0.62

(b) Stop word removal				1	2	3
	F1	0.47	1	0.37	0.24	0.22
	Accuracy	0.49	2	0.23	0.53	0.15
			3	0.40	0.23	0.63

(c) Stemming & stop word removal				1	2	3
	F1	0.45	1	0.35	0.18	0.23
	Accuracy	0.47	2	0.26	0.55	0.16
			3	0.39	0.26	0.60

to represent the information significance of the content. Location fame score was computed by averaging the Wikipedia page size of all the location nouns. The number of location nouns was another feature in this class. Based on the scores given to the development set, information transfer, fame score, video length, number of shots and shot change related features were formed the set of proposed features.

In order to determine the boredom score estimation, the regression weights were computed by means of a RVM from the Tipping RVM toolbox [114].

5.4.2 Boredom ranking results

First, the correlation between low level content features with the development set was studied. Then the features with significant Spearman correlation were chosen in the feature set. Finally, three different runs were generated by combining the new proposed features and the selected content features. The content features with highest ranking correlation with the training set are shown in Table 5.8.

Table 5.8: Content features with significant ranking correlation with boredom scores in the development set.

Feature	**Kendall's Tau correlation**
Average of the third Mel-Frequency cepstral coefficients (MFCC)	-0.24
Average of the 10th MFCC	0.21
Average of the third MFCC	0.24
Standard deviation of the third coefficient of the autocorrelation of MFCC	0.24
Video key lighting	0.23
Average of 16th bin of the Luminance histogram (out of 20 bins)	0.25
Average of 17th bin of the Luminance histogram	0.23

Table 5.9: Ranking evaluation results for all the 5 submitted runs and random level.

Run	**Kendall's Tau ranking correlation**		**Spearman ρ**		**Kendall's Tau ranking distance**	**Spearman footrule distance**
	r	p	ρ	p		
Random level	-	-	-	-	1660	27.4
1	0.13	0.07	0.19	0.08	1700	23.2
2	0.10	0.17	0.14	0.20	1650	25.2
3	0.10	0.18	0.14	0.19	1439	24.4

Four ranking distance metric were used to evaluate the boredom ranking results, which are given in Table 5.9. The Kendall's Tau ranking correlation, Kendall's Tau ranking distance, Spearman ρ and Spearman footrule distance. The details about these metrics are available in [142].

The first run used the proposed feature and not the content features. The second run used all the content features and proposed features together (217 features). Finally, the last run used the combination of the selected content features and proposed features.

The best results were obtained from regression using the proposed set of features and the combination of selected content features with proposed features. None of the generated ranked lists on the test set had significant ranking correlation with the ground truth $p < 0.05$).

A set of features for the ranking of felt boredom in videos is proposed. The boredom ranking results were evaluated using Kendal Tau's ranking correlation and ranking distances. The fame score, amount of information and shot segmentation information are proposed to be useful for boredom ranking.

5.5 Summary

Different content based affective representation systems using different modalities for estimating felt emotions at the scene level have been proposed. The method, which used a Bayesian classification framework that allows taking some form of context into account, obtained the best results. Results showed the advantage of using well chosen priors, such as temporal information provided by the previous scene emotion, and movie genre. The F1 classification score of 0.55 that was obtained on three emotional classes with a naïve Bayesian classifier was improved to 0.56 and 0.59 using only the time and genre prior. This F1 score finally improved to 0.63 after utilizing all the priors. Promising results were also obtained from text analysis. The text analysis can be further developed using different similarity measures and combining with other modalities. A set of features to detect the boredom elicited in the audience from the content was also proposed.

Chapter 6

Conclusions and Perspectives

The studies reported in this thesis were focused on automatically and unobtrusively detecting emotional tags for videos. These emotional tags are useful means for filling the semantic gap between users' and computers' understanding of multimedia content. This problem was attacked from two different perspectives. Automatic detection of users' emotions in response to videos was first attempted. Then automatic prediction of users' emotion based on low-level content features was also studied.

6.1 Lessons Learned

Many lessons were learned by the author which are reflected in this thesis. To summarize, I give the following crucial points:

- Emotional understanding of multimedia is a challenging task and extremely difficult to be solved with a universal solution. Systems based on affective computing can only work if they are able to take context and personal profiles into account.
- Self reports and especially dimensional self reports are not easy to provide. The ground truth definition is one of the main challenges in affective computing and emotion recognition [79]. Special attention is needed in gathering emotional self reports, such as asking different questions to check the consistency and possibly using a post experiment questionnaire.
- Physiological responses vary from person to person and time to time. An effective baselining is required for a robust emotion recognition system.
- To reduce the between participant and between session variance, the experiment environment in the laboratory should be kept isolated from the outside environment. The participants should not be able to see the examiners or hear the noise from the outside. The light and temperature should be controlled to avoid variation in physiological reactions due to non controlled parameters.
- Choosing the right stimuli material is an important factor in any affective study. They should be long enough to induce emotions and short enough to prevent boredom. Furthermore, to assure that variation in stimuli length does not introduce variance in the measurements between emotional and non-emotional stimuli, we suggest the stimuli durations to be equal. The mixture of contradicting emotions can make problems for self-assessments.
- Regarding multimedia content analysis, speech transcripts and subtitles are rich sources

of semantic information. Taking to account textual and contextual information enhances multimedia content-based affective understanding.

6.2 Conclusions

In Chapter 2, I gave a background on emotion theories, affective content analysis and emotion recognition in response to videos. First, emotional theories concerning the emotional responses to videos and movies and cognitive theory of emotions were explained. Then, different emotional presentation such as discrete, and dimensional were explained. A literature review on emotional self reporting methods,emotional understanding of videos using content analysis, emotion recognition using bodily responses and implicit tagging were given.

Next, in Chapter 3, existing emotional corpora were listed and presented. The developed emotional video corpora and analysis over users' emotional self reports were presented. The results on video corpora development showed the importance of context and variation of emotional response by time, mood and gender. The developed databases of emotional, bodily responses were also presented. The experimental protocol and participants' information were provided in detail.

In Chapter 4, methodology and results of emotion recognition methods employed to detect emotion in response to videos were presented. First, a regression based method to detect emotion in continuous space was presented, and correlates of emotional self assessments and physiological responses were shown. Second, a multi-modal participant independent study was presented. The second study showed the performance of an inter-participant emotion recognition tagging approach using participants' EEG signals, gaze distance and pupillary response as affective feedbacks. The feasibility of an approach to recognize emotion in response to videos was shown. The best classification accuracy of 68.5% for three labels of valence and 76.4% for three labels of arousal were obtained using a modality fusion strategy and a support vector machine. Although the results were based on a fairly small video dataset, due to experimental limitations, the promising accuracy can be scalable to more samples from a larger population. The improved performance using multi-modal fusion techniques leads to the conclusion that by adding other modalities, such as facial expressions, accuracy as well as robustness should further improve. The last Section of Chapter 4, presented the results and methods for emotion assessment in response to music videos. Although, a similar approach was taken the results were inferior to the studies with movie scenes. This can be due to the weaker emotional content and the absence or weaker presence of narrative plot in music videos.

In Chapter 5, content analysis methods to detect emotions that are more likely to be elicited by a given multimedia content were presented. Low level content features, which were used for affective understanding, were introduced. Again the regression method was used for affective understanding of videos and the correlation between content features, physiological responses and emotional self reports were studied. Content based multimedia features' correlations with both physiological features and users' self-assessment of valence-arousal were shown to be significant.

Next, an affective representation system for estimating felt emotions at the scene level has been proposed using a Bayesian classification framework. The classification accuracy of 56% that was obtained on three emotional classes with a naïve Bayesian classifier was improved to 64%

after utilizing temporal and genre the priors. Results showed the advantage of using well chosen priors, such as temporal information provided by the previous scene emotion, and movie genre. Finally, a set of features for the ranking of felt boredom in videos were proposed. The boredom ranking results were evaluated using Kendal Tau' ranking correlation and ranking distances. The fame score, amount of information and shot segmentation information were shown to be useful for boredom ranking.

6.3 Perspectives

Different open questions arose during the studies which were conducted during the work presented in thesis. The open questions can be distilled into the following items:

- There are a lot of useful information in the temporal dimension since emotional responses often do not have abrupt changes. Emotion recognition in continuous time, which takes into account the temporal dimension, is one of the key problems which needs to be addressed.
- Text, audio and visual features can be combined at different levels to improve the emotional understanding of videos. Using language models and natural language processing for semantic analysis of subtitles can also improve the results.
- In this thesis, only simple feature level fusions were attempted. More advanced fusion techniques on sensorial level considering the correlation between different responses recorded by different sensors is an interesting approach to be taken.
- Real life or ambulatory recording can provide a more general estimate of emotional responses, which manifest themselves in bodily signals. Real life emotional responses despite the environment noise can have higher intensities. Less obtrusive devises such as the Q-sensor by Affectiva[1] can be employed for real life recordings in places such as movie theaters.
- Users are more responsive in a social context. In future, experiments can be conducted in the presence of more than one participant to boost their expressions and emotional responses.
- The performance of a retrieval system using emotional tags detected by content analysis and emotion recognition systems can be studied and compared to a baseline system using only user generated tags or self reports.

[1] http://www.affectiva.com/q-sensor/

Appendix A: List of Publications

Journal articles

1. M. Soleymani, M. Pantic, and T. Pun. “Multi-Modal Emotion Recognition in Response to Videos”, *IEEE Trans. on Affective Computing*, under review.
2. M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. “A Multi-Modal Database for Affect Recognition and Implicit Tagging”, *IEEE Trans. on Affective Computing, Special Issue on Naturalistic Affect Resources for System Building and Evaluation*, in press, 2011.
3. S. Koelstra*, C. Muhl*, M. Soleymani*, A. Yazdani, J.S Lee, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. “DEAP: A Database for Emotion Analysis using Physiological Signals”, *IEEE Transactions on Affective Computing, Special Issue on Naturalistic Affect Resources for System Building and Evaluation*, in press, 2011. (* shared first authorship, the first three authors are sorted alphabetically)
4. M. Soleymani, G. Chanel, J.J.M. Kierkels, and T. Pun. “Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes”, *Int'l Journal of Semantic Computing*, vol 3, no. 2, 235-254, 2009.
5. G. Chanel, J.J.M. Kierkels, M. Soleymani, and T. Pun. “Short-term emotion assessment in a recall paradigm”, *Int'l Journal of Human-Computer Studies*, vol. 67 no. 8, 607-627, 2009.

Peer-reviewed international conferences and workshops

1. M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, G. Friedland, V. Murdock, R. Ordelman and G.J.F. Jones. “Automatic Tagging and Geo-Tagging in Video Collections and Communities”, *ACM Int'l conference on Multimedia Retrieval (ICMR '11)*, Trento, Italy, 2011. (oral presentation)
2. M. Soleymani, S. Koelstra, I. Patras, T. Pun, “Continuous Emotion Detection in Response to Music Videos”, *Int'l Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE (EmoSPACE)*, Santa Barbara, CA, US, March 2011. (oral presentation)
3. J.J.M. Kierkels, M. Soleymani, and T. Pun, “Identification of Narrative Peaks in Video Clips: Text Features Perform Best”, *Lecture Notes on Computer Science (LNCS)*, *Proceeding of the Cross language evaluation forum (CLEF 2009)*, Springer, 2010.
4. M. Soleymani, Martha Larson, “Crowdsourcing for Affective Annotation of Video: Development of a Viewer-reported Boredom Corpus”, *Workshop on crowdsourcing for search evaluations, ACM SIGIR 2010*, Geneva, Switzerland. (**runner up award from Bing**) (oral presentation)
5. S. Koelstra, A. Yazdani, M. Soleymani, C. Muhl, J-S Lee, T. Ebrahimi, T. Pun, A. Nijlhot, and I. Patras. “Single Trial Classification of EEG and Peripheral Physiological Signals for Recognition of Emotions Induced by Music Videos”, *Int'l Conference on Brain Informatics (BI 2010)*, Toronto, Canada, 2010. (oral presentation)

6. J.J.M. Kierkels, M. Soleymani, and T. Pun, “Queries and tags in affect-based multimedia retrieval”, *IEEE Int’l Conference on Multimedia and Expo (ICME’09)*, Special Session on Implicit Tagging (ICME2009), New York, United States, 2009. (oral presentation)
7. M. Soleymani, G. Chanel, J.J.M. Kierkels, and T. Pun, “A Bayesian Framework for Video Affective Representation”, *Int’l Conference on Affective Computing and Intelligent interaction*, Amsterdam, Netherlands, 2009. (oral presentation)
8. M. Soleymani, G. Chanel, J.J.M. Kierkels, and T. Pun, “Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User’s Physiological Emotional Responses”, *IEEE Int’l Symposium on Multimedia*, Berkeley, US, December 2008. (oral presentation)
9. M. Soleymani, G. Chanel, J.J.M. Kierkels, and T. Pun,“Affective Ranking of Movie Scenes Using Physiological Signals and Content Analysis”, *the second ACM Workshop on the Many Faces of Multimedia Semantics at ACM conference on Multimedia*, Vancouver, Canada, October 2008. (oral presentation)
10. M. Benovoy, A. Brouse, T.G. Corcoran, H. Drayson, C. Erkut, and J. Filatriau, M. Soleymani, et al. “Audiovisual Content Generation Controlled by Physiological Signals for Clinical and Artistic Applications”, *Proceedings of the eNTERFACE 2007 Workshop*, Istanbul, Turkey, 2007.

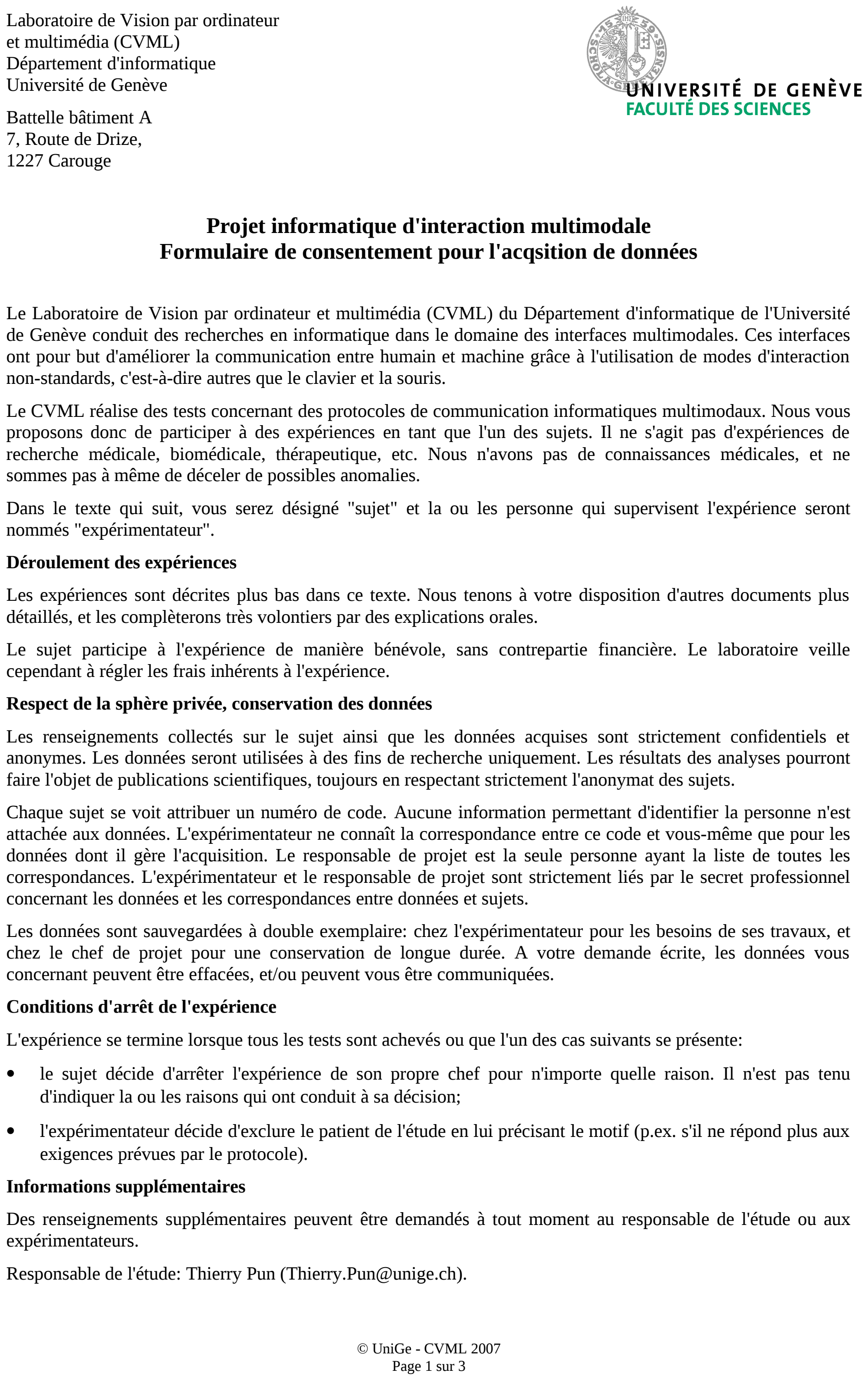
Posters and abstracts

1. G. Gninkoun, and M. Soleymani. “Automatic Violence Scenes Detection: A Multi-Modal Approach”, *Working notes proceedings of the MediaEval 2011 workshop*, Pisa, Italy, 2011.
2. C.-H. Demarty, C. Penet, G. Gravier, and M. Soleymani. “The MediaEval 2011 Affect Task: Violent Scenes Detection in Hollywood Movies”, *Working notes proceedings of the MediaEval 2011 workshop*, Pisa, Italy, 2011.
3. M. Soleymani. “Travelogue Boredom Detection with Content Features”, *Working notes proceedings of the MediaEval 2010 workshop*, Pisa, Italy, 2010.
4. M. Soleymani, Jeremy Davis, and T. Pun. “A collaborative Personalized Affective Video Retrieval System”, *Int’l Conference on Affective Computing and Intelligent interaction*, Amsterdam, Netherlands, 2009. (Demonstration)
5. M. Soleymani, G. Chanel, J, Kierkels, and T. Pun. “Valence-Arousal Representation of Movie Scenes Based on Multimedia Content Analysis and User’s Physiological Emotional Responses”, *5th Joint Workshop on Machine Learning and Multimodal Interaction (MLMI)*, Utrecht, Netherlands, 2008.
6. R. Lehembre, C. Muhl, U. Gundogdu, A. Sayin, M. Soleymani, and C. Erkut. “Auditory vs. visual feedback in an asynchronous brain-computer interface”, *BCI meets Robotics Workshop Leuven*, Belgium, 2007.

Proceeding and special session

1. M. Larson, M. Soleymani, P. Serdyukov, V. Murdock, and G.J.F. Jones. *Working notes proceedings of the MediaEval 2010 workshop*, Pisa, Italy, 2010.
2. M. Larson, M. Soleymani, and P. Serdyukov. *Special session on Automatic Tagging and Geo-Tagging in Video Collections and Communities at the ACM Int’l conference on Multimedia Retrieval (ICMR ’11)*, Trento, Italy, 2011.

Appendix B: Consent Form



Laboratoire de Vision par ordinateur
et multimédia (CVML)
Département d'informatique
Université de Genève

Battelle bâtiment A
7, Route de Drize,
1227 Carouge

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

Projet informatique d'interaction multimodale
Formulaire de consentement pour l'acqsition de données

Le Laboratoire de Vision par ordinateur et multimédia (CVML) du Département d'informatique de l'Université de Genève conduit des recherches en informatique dans le domaine des interfaces multimodales. Ces interfaces ont pour but d'améliorer la communication entre humain et machine grâce à l'utilisation de modes d'interaction non-standards, c'est-à-dire autres que le clavier et la souris.

Le CVML réalise des tests concernant des protocoles de communication informatiques multimodaux. Nous vous proposons donc de participer à des expériences en tant que l'un des sujets. Il ne s'agit pas d'expériences de recherche médicale, biomédicale, thérapeutique, etc. Nous n'avons pas de connaissances médicales, et ne sommes pas à même de déceler de possibles anomalies.

Dans le texte qui suit, vous serez désigné "sujet" et la ou les personne qui supervisent l'expérience seront nommés "expérimentateur".

Déroulement des expériences

Les expériences sont décrites plus bas dans ce texte. Nous tenons à votre disposition d'autres documents plus détaillés, et les complèterons très volontiers par des explications orales.

Le sujet participe à l'expérience de manière bénévole, sans contrepartie financière. Le laboratoire veille cependant à régler les frais inhérents à l'expérience.

Respect de la sphère privée, conservation des données

Les renseignements collectés sur le sujet ainsi que les données acquises sont strictement confidentiels et anonymes. Les données seront utilisées à des fins de recherche uniquement. Les résultats des analyses pourront faire l'objet de publications scientifiques, toujours en respectant strictement l'anonymat des sujets.

Chaque sujet se voit attribuer un numéro de code. Aucune information permettant d'identifier la personne n'est attachée aux données. L'expérimentateur ne connaît la correspondance entre ce code et vous-même que pour les données dont il gère l'acquisition. Le responsable de projet est la seule personne ayant la liste de toutes les correspondances. L'expérimentateur et le responsable de projet sont strictement liés par le secret professionnel concernant les données et les correspondances entre données et sujets.

Les données sont sauvegardées à double exemplaire: chez l'expérimentateur pour les besoins de ses travaux, et chez le chef de projet pour une conservation de longue durée. A votre demande écrite, les données vous concernant peuvent être effacées, et/ou peuvent vous être communiquées.

Conditions d'arrêt de l'expérience

L'expérience se termine lorsque tous les tests sont achevés ou que l'un des cas suivants se présente:

- le sujet décide d'arrêter l'expérience de son propre chef pour n'importe quelle raison. Il n'est pas tenu d'indiquer la ou les raisons qui ont conduit à sa décision;
- l'expérimentateur décide d'exclure le patient de l'étude en lui précisant le motif (p.ex. s'il ne répond plus aux exigences prévues par le protocole).

Informations supplémentaires

Des renseignements supplémentaires peuvent être demandés à tout moment au responsable de l'étude ou aux expérimentateurs.

Responsable de l'étude: Thierry Pun (Thierry.Pun@unige.ch).

© UniGe - CVML 2007
Page 1 sur 3

Expérimentateurs actuels: Guillaume Chanel, guillaume.chanel@unige.ch, Mohammad Soleymani, mohammad.soleymani@unige.ch.

Acquisition de signaux physiologiques

Nous utilisons pour enregistrer les signaux physiologiques le dispositif d'acquisisition Active II de la société Biosemi (http://www.biosemi.com). La caractéristique de ce système est qu'il utilise des électrodes dites actives, c'est-à-dire qu'une infime quantité de courant est diffusée par la surface de l'électrode. Selon l'information reçue et à notre connaissance, la quantité infime de courant injectée permet de supposer qu'il n'y a pas de risques pour la santé du sujet. De la même manière, nous ne sommes pas au courant de contre-indications ou risques associés à l'utilisation de cet équipement. Le système Biosemi Active II est utilisé par de nombreux laboratoires à travers le monde pour des expériences similaires. Il n'y a pas de risque d'électrocution car le dispositif à électrodes est isolé galvaniquement du reste du système d'acquisition (liaison par fibre optique) et est alimenté par batterie.

Acquisition de signaux électro-encéphalographiques (EEG)

Des signaux EEG - électroencéphalographiques sont utilisés pour le développement d'interfaces informatiques interactives et multimodales (faisant appel à plusieurs sens humains). Les signaux acquis permettent de localiser les régions du cerveau activées pour une tâche donnée. Dans le futur, et c'est là l'un des buts de ces recherches, ils devraient pouvoir aussi servir à contrôler directement une machine par "la pensée", chaque commande étant associée à un état mental précis ("tâche") de l'utilisateur. Dans le futur toujours, ils devraient également permettre de détecter de manière grossière l'état émotionnel de l'utilisateur.

Acquisition d'autres types de signaux physiologiques

D'autres types de signaux physiologiques peuvent également être enregistrés, pour étudier l'ensemble des réponses à certains stimulis. En fonction des capteurs à disposition, ces signaux peuvent être par exemple de type électro-cardiographique (ECG), électro-myographique (EMG), résistance de la peau (GSR - *Galvanic Skin Resistance*).

Acquisition vidéo

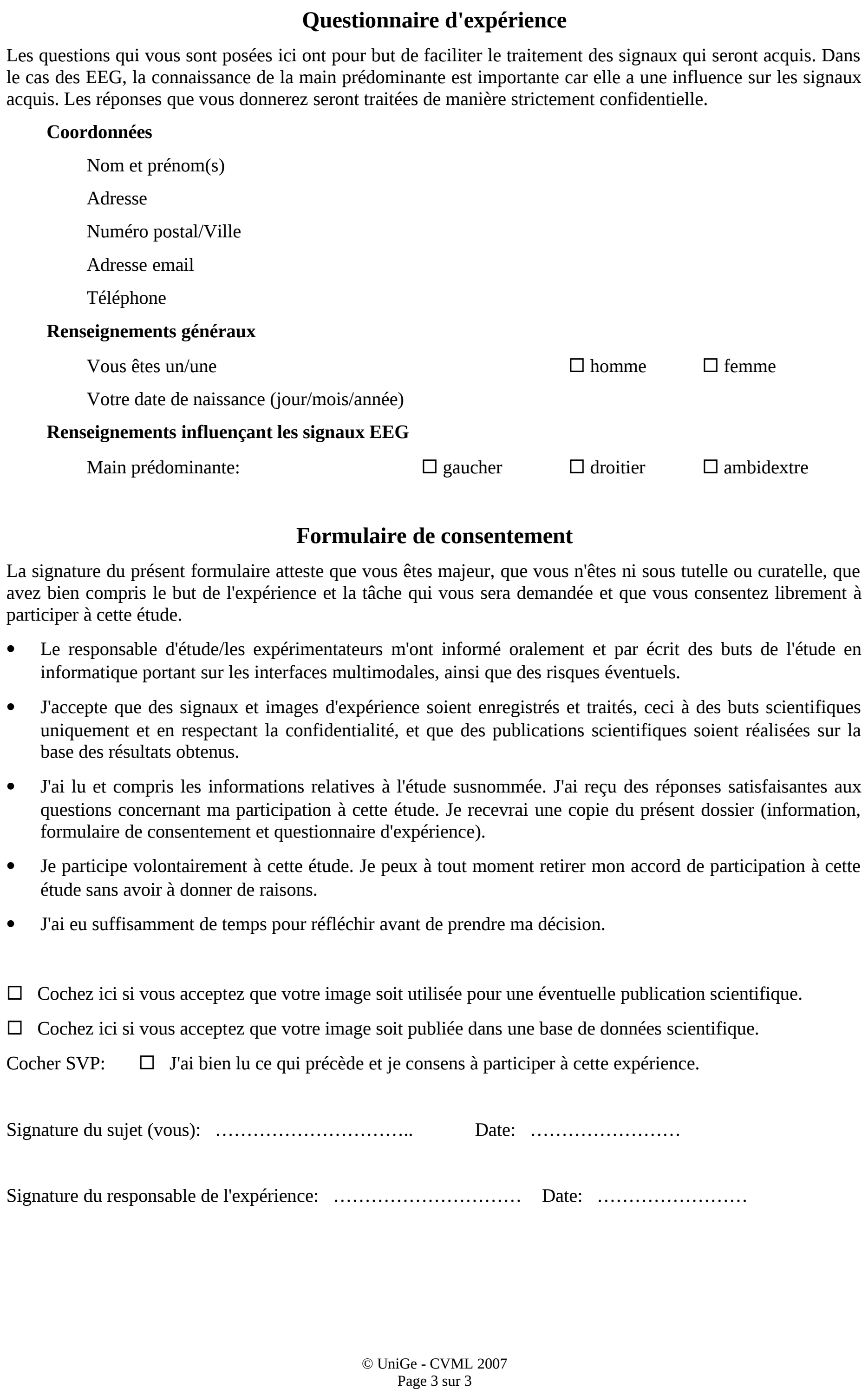
Il est possible qu'une vidéo de l'expérience, incluant votre visage, soie enregistrée dans un but de contrôle (pour vérifier que l'expérience c'est déroulée comme prévue) mais également pour étudier les aspects comportementaux.

Déroulement des expériences d'enregistrement des EEGs

Le sujet est assis sur une chaise et porte un casque à électrodes sur la tête (figure ci-contre). Ce casque est relié à un ordinateur via un dispositif d'acquisition qui stocke les données reçues en temps réel. Ces données sont les potentiels électriques mesurés par chaque électrode (maximum 64 électrodes). Selon le type de test, les expériences peuvent se dérouler de diverses manières.

Pour la problématique du contrôle direct d'une machine par la pensée, le sujet effectue plusieurs tâches mentales (p.ex. imagination de mouvement, calcul mental) et les données ainsi acquises sont traitées par l'ordinateur qui tente d'extraire des commandes pour une application informatique. L'expérience se déroule en deux phases. Durant la première phase dite d'apprentissage, le sujet se familiarise avec l'appareil et réalise cet apprentissage dont la durée dépend des capacités du sujet. Il est difficile de prévoir à l'avance la durée de l'apprentissage. Les cas reportés les plus longs mentionnent un maximum de 30 séances d'une heure réparties dans le temps. Dans nos propres expériences, cette durée est sensiblement inférieure, et l'un des buts de la recherche est de la réduire encore. La seconde phase consiste en l'utilisation de l'application informatique. La durée de cette phase dépend de l'expérience et est inférieure à la durée de l'apprentissage (au maximum 15 heures).

Pour la détection de l'état émotionnel, divers stimulus sont présentés (images, vidéos, sons) et les EEGs sont enregistrés. Dans ces expériences, il n'y a en principe pas de phase d'apprentissage, et leur durée est plus courte que pour la problématique du contrôle direct d'une machine par la pensée. **Attention certaines scènes des vidéos présentées peuvent heurter votre sensibilité**.



Questionnaire d'expérience

Les questions qui vous sont posées ici ont pour but de faciliter le traitement des signaux qui seront acquis. Dans le cas des EEG, la connaissance de la main prédominante est importante car elle a une influence sur les signaux acquis. Les réponses que vous donnerez seront traitées de manière strictement confidentielle.

Coordonnées

Nom et prénom(s)

Adresse

Numéro postal/Ville

Adresse email

Téléphone

Renseignements généraux

Vous êtes un/une □ homme □ femme

Votre date de naissance (jour/mois/année)

Renseignements influençant les signaux EEG

Main prédominante: □ gaucher □ droitier □ ambidextre

Formulaire de consentement

La signature du présent formulaire atteste que vous êtes majeur, que vous n'êtes ni sous tutelle ou curatelle, que avez bien compris le but de l'expérience et la tâche qui vous sera demandée et que vous consentez librement à participer à cette étude.

- Le responsable d'étude/les expérimentateurs m'ont informé oralement et par écrit des buts de l'étude en informatique portant sur les interfaces multimodales, ainsi que des risques éventuels.
- J'accepte que des signaux et images d'expérience soient enregistrés et traités, ceci à des buts scientifiques uniquement et en respectant la confidentialité, et que des publications scientifiques soient réalisées sur la base des résultats obtenus.
- J'ai lu et compris les informations relatives à l'étude susnommée. J'ai reçu des réponses satisfaisantes aux questions concernant ma participation à cette étude. Je recevrai une copie du présent dossier (information, formulaire de consentement et questionnaire d'expérience).
- Je participe volontairement à cette étude. Je peux à tout moment retirer mon accord de participation à cette étude sans avoir à donner de raisons.
- J'ai eu suffisamment de temps pour réfléchir avant de prendre ma décision.

□ Cochez ici si vous acceptez que votre image soit utilisée pour une éventuelle publication scientifique.

□ Cochez ici si vous acceptez que votre image soit publiée dans une base de données scientifique.

Cocher SVP: □ J'ai bien lu ce qui précède et je consens à participer à cette expérience.

Signature du sujet (vous): ………………………….. Date: ……………………

Signature du responsable de l'expérience: ………………………… Date: ……………………

Appendix C: Participant's Instructions

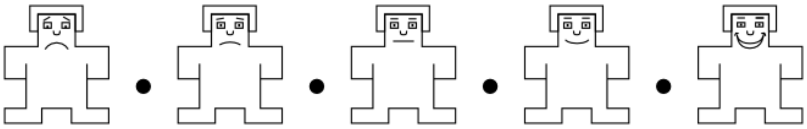
Participant instructions

In the experiment you are participating in, we study the relation between music videos and emotions. Your task is to watch the videos and afterwards to judge your own emotional state, that is your feelings elicited by the music video, and to rate the video.

There are 3 scales that you can use to indicate your feelings: pleasure, arousal, and dominance. All scales range from 1 to 9.

Valence scale

The scale ranges from big frown to big smile, which represents feelings that range from **unhappy** or **sad** to **happy** or **joyful**. The rating is done by choosing one among 9 options within the range, as shown below. The facial expressions of the mannequins help you find appropriate options.

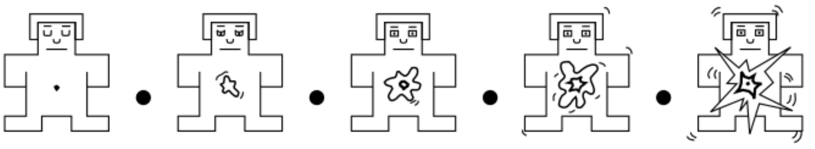


unhappy
sad

happy
joyful

Arousal scale

The scale ranges from **calm** or **bored** to **stimulated** or **excited**. Again, you are to choose one among 9 options. You can see that the facial expression of the mannequin remains the same, but the "explosion" or grumbling in the stomach of the mannequin indicates the degree of arousal.

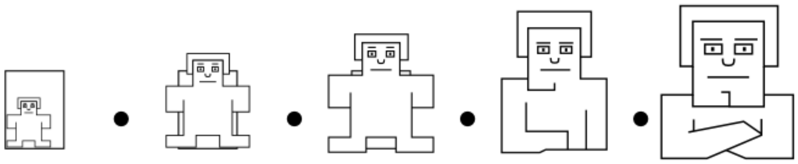


calm
bored

stimulated
excited

Dominance/Control

Sometimes you feel empowered (in control of everything) when listening to music or seeing a video. Other times, you feel rather helpless and weak (without control). The scale ranges from submissive (or "without control") to dominant (or "in control"). A small mannequin in the left side indicates that you feel unimportant or without control, bullied, like someone else is the leader or in charge, or like you can't handle the situation. A big mannequin in the right side indicates that you feel important, very big, or like you don't need anyone's help.



unimportant
without control

important
in control

General liking scale

A fourth scale asks for your personal rating of the video, that is how much you liked the music video. Please be careful not to confuse this with the “pleasure scale”. Here we are interested in your taste, not your feeling. For example you can still like a video that makes you feel bad, i.e. sad or angry.

dislike

like

If you have more questions please don’t hesitate to refer to the experimenter.

And now .. enjoy the experiment ☺

Appendix D: Video Clips

First Movie Database

To create the video dataset, we extracted video scenes from eight movies selected either according to similar studies, or from recent famous movies. The movies included four major genres: drama, horror, action, and comedy. Video clips used for this study are from the following: Saving Private Ryan (action), Kill Bill, Vol. 1 (action), Hotel Rwanda (drama), The Pianist (drama), Mr. Bean's Holiday (comedy), Love Actually (comedy), The Ring, Japanese version (horror) and 28 Days Later (horror). The extracted scenes, eight for each movie, had durations of approximately one to two minutes each and contained an emotional event. The following dataset has been presented and used in Sections 3.1.2.1, 4.2.1 and 5.1.3.

28 Days Later

The version of the movie we used starts by the "Fox searchlights pictures" Intro.

Scene number	Start time	Stop time	Description of the start of the scene
1	00:09:05	00:10:09	Jim (Cillian Murphy) walks in the deserted streets of London with plastic bag in his hand
2	00:13:39	00:15:15	Starts with a close shot on Jim's face in the church
3	00:42:33	00:44:55	The taxi starts ascending the trash dump in the tunnel
4	00:49:04	00:50:02	Inside the grocery store in the deserted gas station
5	01:08:29	01:09:24	Major Henry West (Christopher Eccleston) and Jim are walking in a corridor
6	01:34:58	01:35:46	Hannah (Megan Burns) is sitting in a red dress frightened in a room
7	01:36:09	01:37:58	A fight between the black soldier and a zombie
8	01:39:14	01:39:48	Jim opens the taxi's door facing Major Henry West

Hotel Rwanda

The version of the movie we used starts by the "METROPOLITAN FILMS" Intro.

Scene number	Start time	Stop time	Description of the start of the scene
1	00:07:09	00:07:50	General Bizimungu (Fana Mokoena) and Colonel Oliver(Nick Nolte) are talking in the hotel's garden (while tourists are in the pool in the background)
2	00:09:37	00:10:58	In the Paul's (Don Cheadle) house, their son run into the living room frightened
3	00:23:04	00:24:21	Discussion between Paul and the Rwandan officer and he askes for Paul's ID
4	00:51:04	00:54:33	French soldiers are checking tourists' passports
5	01:07:45	01:09:16	The hotel's van is passing by is passing by burning houses at night
6	01:11:41	01:12:25	The hotel's van is on the road in a foggy dawn
7	01:27:03	01:28:40	Rebells are dancing around the road waiting for the UN trucks
8	01:28:15	01:29:39	Rebells are hitting the refugees in the truck

Kill Bill VOL 1

The version of the movie we used starts by the "MIRAMAX" Intro.

Scene number	Start time	Stop time	Description of the start of the scene
1	00:05:46	00:06:46	Uma thrman is fighting with Vernita Green (Vivica A. Fox)
2	00:58:50	01:01:00	The Japanese gangs are sitting around a black table
3	01:10:13	01:11:40	Japanese gangs are drinking in a room
4	01:15:02	01:17:07	Gogo Yubari (Chiaki Kuriyama) starts fighting with Uma Thurman
5	01:18:24	01:22:50	The fight scene of Uma Thurman and Japanese fighters in black suits
6	01:24:43	01:25:40	The fight scene of Uma Thurman and the Japanese bald fighter (Kenji Ohba)
7	01:28:48	01:30:53	The final fight scene between Uma Thurman and O-Ren Ishii(Lucy Lin)
8	01:02:10	01:03:35	Motorbikes are escorting a Mercedes in Tokyo streets

Love Actually

The version of the movie we used starts by the “UNIVERSAL” Intro

Scene number	Start time	Stop time	Description of the start of the scene
1	00:12:19	00:13:57	Colin Frissell (Kriss Marchall) serves at the party
2	00:47:37	00:49:09	Aurelia (Lúcia Moniz) is bringing a cup of coffee for Jamie Bennett (Colin Firth) in the garden
3	01:18:28	01:20:33	The jewellery salesman (Rowan Atkinson) is rapping a necklace
4	01:23:45	01:27:04	Colin Frissell arrives at Milwaukee
5	01:43:19	01:44:57	The old lady opens the door and surprises by seeing the prime minister (Hugh Grant)
6	01:51:38	01:54:14	School’s Christmas concert starts
7	01:58:15	02:05:00	Jamie Bennett arrived at Portugal
8	00:28:26	00:29:52	Daniel (Liam Neeson) and Sam (Thomas Sangster) are sitting on a bench

Mr. Bean’s Holiday

The version of the movie we used starts by the “UNIVERSAL” Intro.

Scene number	Start time	Stop time	Description of the start of the scene
1	00:05:39	00:07:22	Mr. Bean (Rowan Atkinson) takes a taxi at the train station
2	00:10:41	00:12:59	Mr. Bean is being served in a French restaurant
3	00:31:52	00:33:50	Mr. Bean is trying to raise money by dancing and imitating singer’s acts
4	00:37:33	00:39:19	Mr. Bean rides a bike on a road
5	00:40:37	00:42:35	Mr. Bean tries to hitchhike
6	00:45:45	00:47:15	Mr. Bean wakes up in a middle of the shooting of a commercial
7	01:08:04	01:09:02	Dressed as a woman tries to get into the theater with a fake ID
8	01:11:35	01:13:22	Mr. Bean is changing the projecting movie to his webcam videos

Ringu (Japanese version)

The version of the movie we used starts by the “STUDIO CANAL” Intro.

Scene number	Start time	Stop time	Description of the start of the scene
1	00:05:32	00:08:15	School girls are frightened by hearing the ring tone
2	00:27:39	00:29:38	Reiko (Nanako Matsushima) watches the video alone in an empty room
3	00:59:39	01:01:52	Reiko sees the past in black and white
4	01:19:10	01:21:48	Reiko is descending into the well
5	01:25:24	01:27:54	Ryuji (Hiroyuki Sanada) is writing at home and he notices that the TV is on and showing the terrifying video
6	01:29:56	01:31:55	Reiko seats on the sofa in her house
7	01:12:05	01:14:48	Reiko and Ryuji are pushing the well’s lead
8	00:48:20	01:49:42	Reiko is sleeping in her father’s house

Saving private Ryan

The version of the movie we used starts by the “Paramount pictures” Intro.

Scene number	Start time	Stop time	Description of the start of the scene
1	00:04:29	00:06:08	Start scene of the approaching of boats with these words apear “June 6, 1944”
2	00:06:09	00:08:15	Landing on the Omaha beach
3	00:09:33	00:12:56	Combat scene on the beach
4	02:13:38	02:14:23	The sniper is praying when he is on a tower
5	00:18:51	00:21:38	The commander is looking into a mirror to see the source of the gunfire in a combat scene
6	02:26:05	02:27:21	The combat scene where Capt. John H. Miller (Tom Hanks) was shot
7	00:56:04	00:57:25	While they are looking for private Ryan, by accident a wall collapses and they face a group of German soldiers on the other side of the destroyed wall
8	01:20:00	01:20:45	Group of soldiers are walking on a green field

The Pianist

The version of the movie we used starts by the “BAC films” Intro.

Scene number	Start time	Stop time	Description of the start of the scene
1	00:00:24	00:02:09	Warsaw in 1939 (black and white shots)
2	00:21:10	00:22:34	Szpilman (Adrien Brody) is playing in a restaurant
3	00:24:28	00:25:54	Szpilman walks in the streets of Warsaw
4	00:32:57	00:34:11	A crazy man and children on the street
5	01:56:12	01:58:13	Szpilman (with long hair and beards) tries to open a can
6	00:44:34	00:47:07	Jewish families are waiting to be sent to concentration camps
7	00:58:01	01:59:27	Szpilman in a construction site
8	01:50:38	01:51:52	German soldiers are burning everything with flamethrower

The Stimuli Videos for MAHNOB-HCI Database, Second Movie Database

Video fragments which were shown as stimuli in the affective tagging experiments, MAHNOB-HCI.

			source		
Cut#	**file name**	**emotion**	**movie name**	**start time**	**end time**
1	69.avi	disgust	*Hannibal*	1:44:50.7	1:45:49.9
2	55.avi	anger/sadness	*The pianist*	0:54:33.3	0:55:50.4
3	58.avi	amusement	*Mr Bean's Holiday*	1:17:19	1:18:18
4	earworm_f.avi	disgust	http://blip.tv/file/1335283/		
5	53.avi	amusement	*Kill Bill VOL I*	1:12:12.2	1:13:57.2
6	80.avi	joy	*Love actually*	0:09:45.76	0:11:22.96
7	52.avi	amusement	*Mr Bean's Holiday*	1:05:53.2	1:07:30.6
8	79.avi	joy	*The thin red line*	0:07:37.96	0:08:21.68
9	73.avi	fear	*The shining*	2:16:42.3	2:17:55.2
10	90.avi	joy	*Love actually*	0:33:59.6	0:35:25.8
11	107.avi	fear	*The shining*	2:07:02.8	2:07:38.2
12	146.avi	sadness	*Gangs of New York*	2:34:41.1	2:36:10
13	30.avi	fear	*Silent Hill*	1:22:27.6	1:23:39.5
14	138.avi	sadness	*The thin red line*	1:06:32	1:08:29.8
15	newyork_f.avi	neutral	http://accuweather.com/ n.a. (please refer to audio ch. 1)		
16	111.avi	sadness	*American History X*	1:52:05.9	1:54:00
17	detroit_f.avi	neutral	http://accuweather.com/ n.a. (please refer to audio ch. 1)		
18	cats_f.avi	joy	http://www.youtube.com/watch?v=E6h1KsWNU-A		
19	dallas_f.avi	neutral	http://accuweather.com/ n.a. (please refer to audio ch. 1)		
20	funny_f.avi	joy	http://blip.tv/file/1854578/		

Bibliography

[1] D. Sander, D. Grandjean, and K. R. Scherer, “A systems approach to appraisal mechanisms in emotion,” *Neural Netw.*, vol. 18, no. 4, pp. 317–352, May 2005.

[2] M. Soleymani, S. Koelstra, I. Patras, and T. Pun, “Continuous emotion detection in response to music videos,” in *1st International Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE (EmoSPACE)*, Santa Barbara, US, March 2011.

[3] S. Koelstra, C. Mühl, M. Soleymani, A. Yazdani, J.-S. Lee, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, “DEAP: A Database for Emotion Analysis using Physiological Signals,” *IEEE Trans. Affective Computing, Special Issue on Naturalistic Affect Resources for System Building and Evaluation*, 2011, in Press.

[4] (2010) Great scott! over 35 hours of video uploaded every minute to youtube. Blog post, Visited on 10 august 2011. [Online]. Available: http://youtube-global.blogspot.com/2010/11/great-scott-over-35-hours-of-video.html

[5] M. Pantic and A. Vinciarelli, “Implicit human-centered tagging,” *IEEE Signal Processing Magazine*, vol. 26, no. 6, pp. 173–180, November 2009.

[6] A. F. Smeaton, P. Over, and W. Kraaij, “High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements,” in *Multimedia Content Analysis, Theory and Applications*, A. Divakaran, Ed. Berlin: Springer Verlag, 2009, pp. 151–174.

[7] A. Hanjalic and L.-Q. Xu, “Affective video content representation and modeling,” *Multimedia, IEEE Transactions on*, vol. 7, no. 1, pp. 143–154, 2005.

[8] J. Kim and E. André, “Emotion recognition based on physiological changes in music listening,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 30, no. 12, pp. 2067–2083, 2008.

[9] J. Rottenberg, R. D. Ray, and J. J. Gross, *Emotion elicitation using films*, ser. Series in affective science. Oxford University Press, 2007, pp. 9–28.

[10] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, “Affective Characterization of Movie Scenes Based on Content Analysis and Physiological Changes,” *International Journal of Semantic Computing*, vol. 3, no. 2, pp. 235–254, June 2009.

[11] M. K. Shan, F. F. Kuo, M. F. Chiang, and S. Y. Lee, “Emotion-based music recommendation by affinity discovery from film music,” *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7666–7674, September 2009.

[12] M. Tkalčič, U. Burnik, and A. Košir, “Using affective parameters in a content-based recommender system for images,” *User Modeling and User-Adapted Interaction*, vol. 20, no. 4, pp. 279–311, September 2010.

[13] J. J. M. Kierkels, M. Soleymani, and T. Pun, “Queries and tags in affect-based multimedia retrieval,” in *ICME’09: Proceedings of the 2009 IEEE international conference on Multimedia and Expo*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 1436–1439.

[14] R. W. Picard and S. B. Daily, “Evaluating Affective Interactions: Alternatives to Asking What Users Feel,” in *CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches*, 2005.

[15] K. R. Scherer, “What are emotions? And how can they be measured?” *Social Science Information*, vol. 44, no. 4, pp. 695–729, December 2005.

[16] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J.-S. Lee, A. Nijholt, T. Pun, T. Ebrahimi, and I. Patras, “Single Trial Classification of EEG and Peripheral Physiological Signals for Recognition of Emotions Induced by Music Videos,” in *Brain Informatics*, ser. Lecture Notes in Computer Science, Y. Yao, R. Sun, T. Poggio, J. Liu, N. Zhong, and J. Huang, Eds. Berlin, Heidelberg: Springer, 2010, vol. 6334, ch. 9, pp. 89–100.

[17] M. Soleymani, G. Chanel, J. J. M. Kierkels, and T. Pun, “Affective ranking of movie scenes using physiological signals and content analysis,” in *MS ’08: Proceeding of the 2nd ACM workshop on Multimedia semantics.* New York, NY, USA: ACM, 2008, pp. 32–39.

[18] M. Soleymani, J. Davis, and T. Pun, “A collaborative personalized affective video retrieval system,” in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, sep 2009.

[19] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun, “A bayesian framework for video affective representation,” in *Proceedings of the International Conference on Affective Computing and Intelligent interaction (ACII 2009)*, September 2009, pp. 1–7.

[20] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A Multi-Modal Affective Database for Affect Recognition and Implicit Tagging,” *IEEE Trans. Affective Computing, Special Issue on Naturalistic Affect Resources for System Building and Evaluation*, under review.

[21] M. Soleymani, G. Chanel, J. Kierkels, and T. Pun, “Affective Characterization of Movie Scenes Based on Multimedia Content Analysis and User’s Physiological Emotional Responses,” in *IEEE International Symposium on Multimedia (ISM’08)*, Berkeley, US, December 2008.

[22] 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, SIGIR 2010*, Geneva, Switzerland, July 2010.

[23] W. Wirth and H. Schramm, “Media and Emotions,” *Communication research trends*, vol. 24, no. 3, pp. 3–39, 2005.

[24] P. Ekman, *Basic Emotions.* John Wiley & Sons, Ltd, 2005, pp. 45–60.

[25] W. James, “What is an emotion?” *Mind*, vol. 9, pp. 17+, 1884.

[26] K. R. Scherer, “Studying the emotion-antecedent appraisal process: An expert system approach,” *Cognition & Emotion*, vol. 7, no. 3, pp. 325–355, 1993.

[27] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions.* Cambridge University Press, July 1988.

[28] K. R. Scherer, *The component process model: Architecture for a comprehensive computational model of emergent emotion.* Oxford, UK: Oxford University Press, 2010, ch. 2.1, pp. 105–130.

[29] D. Zillmann, *The psychology of suspense in dramatic exposition.* Lawrence Erlbaum Associates, Inc, 1996, pp. 199–231.

[30] A. I. Nathanson, *Rethinking Empathy.* Lawrence Erlbaum Associates, Inc, 2003, pp. 107–130.

[31] D. Zillmann, *Empathy: Affect from bearing witness to the emotions of others.* Lawrence Erlbaum Associates, Inc, 1991, pp. 135–168.

[32] R. Plutchik, *A general psychoevolutionary theory of emotion.* New York: Academic press, 1980, pp. 3–33.

[33] W. Wundt, *Grundzüge der physiologischen Psychologie.* Leipzig: Engelmann, 1905.

[34] S. Marsella, J. Gratch, and P. Petta, *Computational models of emotion.* Oxford, UK: Oxford University Press, 2010, ch. 1.2, pp. 21–41.

[35] J. A. Russell, "Culture and the Categorization of Emotions," *Psychological Bulletin*, vol. 110, no. 3, pp. 426–450, 1991.

[36] J. A. Russell and A. Mehrabian, "Evidence for a three-factor theory of emotions," *Journal of Research in Personality*, vol. 11, no. 3, pp. 273–294, September 1977.

[37] J. Fontaine, K. Scherer, E. Roesch, and P. Ellsworth, "The World of Emotions is not Two-Dimensional," *Psychological Science*, vol. 18, no. 12, pp. 1050–1057, 2007.

[38] P. Desmet, *Measuring emotion: development and application of an instrument to measure emotional responses to products.* Norwell, MA, USA: Kluwer Academic Publishers, 2003, ch. 9, pp. 111–123.

[39] P. Winoto and T. Y. Tang, "The role of user mood in movie recommendations," *Expert Systems with Applications*, vol. 37, no. 8, pp. 6086–6092, 2010.

[40] M. M. Bradley and P. J. Lang, "Measuring emotion: the Self-Assessment Manikin and the Semantic Differential." *J Behav Ther Exp Psychiatry*, vol. 25, no. 1, pp. 49–59, March 1994.

[41] J. A. Russell, A. Weiss, and G. A. Mendelsohn, "Affect Grid: A single-item scale of pleasure and arousal," *Journal of Personality and Social Psychology*, vol. 57, no. 3, pp. 493–502, September 1989.

[42] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, "Universals and cultural differences in the judgments of facial expressions of emotion." *Journal of personality and social psychology*, vol. 53, no. 4, pp. 712–717, October 1987.

[43] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, December 1980.

[44] A. Schaefer, F. Nils, X. Sanchez, and P. Philippot, "Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers," *Cognition & Emotion*, vol. 24, no. 7, pp. 1153–1172, 2010.

[45] D. Watson and L. A. Clark, "The PANAS-X: Manual for the Positive and Negative Affect Schedule Expanded Form," 1994.

[46] S. Zoghbi, D. Kulic, E. Croft, and M. Van der Loos, "Evaluation of affective state estimations using an on-line reporting device during human-robot interactions," in *Proceedings of the 2009 IEEE/RSJ international conference on Intelligent robots and systems*, ser. IROS'09. Piscataway, NJ, USA: IEEE Press, 2009, pp. 3742–3749.

[47] O. Villon, "Modeling affective evaluation of multimedia contents: user models to Associate subjective experience, physiological expression and contents description," Ph.D. dissertation, Université de Nice - Sophia Antipolis, Nice, France, October 2007.

[48] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. Mcmahon, M. Sawey, and M. Schröder. (2000) 'feeltrace': an instrument for recording perceived emotion in real time.

[49] E. Douglas-cowie, R. Cowie, and M. Schröder, "A New Emotion Database: Considerations, Sources and Scope," in *In*, 2000, pp. 39–44.

[50] H. L. Wang and L.-F. Cheong, "Affective understanding in film," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 16, no. 6, pp. 689–704, jun 2006.

[51] G. Irie, T. Satou, A. Kojima, T. Yamasaki, and K. Aizawa, "Affective Audio-Visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification," *Multimedia, IEEE Transactions on*, vol. 12, no. 6, pp. 523–535, October 2010.

[52] M. Xu, J. S. Jin, S. Luo, and L. Duan, "Hierarchical movie affective content analysis based on arousal and valence features," in *Proceeding of the 16th ACM international conference on Multimedia*, ser. MM '08. New York, NY, USA: ACM, 2008, pp. 677–680.

[53] S. Zhang, Q. Tian, S. Jiang, Q. Huang, and W. Gao, "Affective mtv analysis based on arousal and valence features," in *Multimedia and Expo, 2008 IEEE International Conference on*, april 2008, pp. 1369–1372.

[54] H.-B. Kang, "Affective content detection using hmms," in *Proceedings of the eleventh ACM international conference on Multimedia*, ser. MULTIMEDIA '03. New York, NY, USA: ACM, 2003, pp. 259–262.

[55] S. Arifin and P. Cheung, "Affective Level Video Segmentation by Utilizing the Pleasure-Arousal-Dominance Information," *Multimedia, IEEE Transactions on*, vol. 10, no. 7, pp. 1325–1341, 2008.

[56] P. Lang, M. Bradley, and B. Cuthbert, "International affective picture system (iaps): Affective ratings of pictures and instruction manual," University of Florida, Gainesville, Florida, US, Tech. Rep. A-8, 2005.

[57] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[58] J. Wang and Y. Gong, "Recognition of multiple drivers' emotional state," in *In ICPR'08: The 19th International Conference on Pattern Recognition*, 2008.

[59] J. A. Healey, "Wearable and automotive systems for affect recognition from physiology," Ph.D. dissertation, MIT, 2000.

[60] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP J. Appl. Signal Process.*, vol. 2004, no. 1, pp. 1672–1687, January 2004.

[61] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, August 2009.

[62] K. Takahashi, "Remarks on Emotion Recognition from BioPotential Signals," in *in 2nd Int. Conf. on Autonomous Robots and Agents, 2004*, 2005.

[63] J. Bailenson, E. Pontikakis, I. Mauss, J. Gross, M. Jabon, C. Hutcherson, C. Nass, and O. John, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *International Journal of Human-Computer Studies*, vol. 66, no. 5, pp. 303–317, may 2008.

[64] V. Kolodyazhniy, S. D. Kreibig, J. J. Gross, W. T. Roth, and F. H. Wilhelm, "An affective computing approach to physiological emotion specificity: Toward subject-independent and stimulus-independent classification of film-induced emotions," *Psychophysiology*, p. In Press, 2011.

[65] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *Intelligent Transportation Systems, IEEE Transactions on*, vol. 6, no. 2, pp. 156–166, 2005.

[66] R. W. Picard, E. Vyzas, and J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, Oct. 2001.

[67] C. D. Katsis, N. Katertsidis, G. Ganiatsas, and D. I. Fotiadis, "Toward Emotion Recognition in Car-Racing Drivers: A Biosignal Processing Approach," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. 38, no. 3, pp. 502–512, may 2008.

[68] S. D. Kreibig, F. H. Wilhelm, W. T. Roth, and J. J. Gross, “Cardiovascular, electrodermal, and respiratory response patterns to fear and sadness-inducing films,” *Psychophysiology*, vol. 44, pp. 787–806, 2007.

[69] Bradley, M. Margaret, Miccoli, Laura, Escrig, A. Miguel, Lang, and J. Peter, “The pupil as a measure of emotional arousal and autonomic activation,” *Psychophysiology*, vol. 45, no. 4, pp. 602–607, July 2008.

[70] T. Partala and V. Surakka, “Pupil size variation as an indication of affective processing,” *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 185–198, 2003.

[71] Y. Gao, A. Barreto, and M. Adjouadi, “Monitoring and processing of the pupil diameter signal for affective assessment of a computer user,” in *Proceedings of the 13th International Conference on Human-Computer Interaction. Part I: New Trends.* Berlin, Heidelberg: Springer-Verlag, 2009, pp. 49–58.

[72] 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*, vol. 31, no. 1, pp. 39–58, March 2009.

[73] J. Russell and J. Fernandez-Dols, Eds., *The psychology of facial expression.* New York, US: Cambridge University Press, 1997.

[74] Keltner, D. and Ekman, P., *Facial Expression Of Emotion*, 2nd ed. New York, US: Guilford Publications, 2000, pp. 236–249.

[75] M. Pantic and I. Patras, “Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences,” *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, no. 2, pp. 433–449, Mar. 2006.

[76] P. R. De Silva and N. Bianchi-Berthouze, “Modeling human affective postures: an information theoretic characterization of posture features,” *Comp. Anim. Virtual Worlds*, vol. 15, no. 3-4, pp. 269–276, 2004.

[77] M. Nicolaou, H. Gunes, and M. Pantic, “Output-associative rvm regression for dimensional and continuous emotion prediction,” in *Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*, march 2011, pp. 16–23.

[78] A. Kleinsmith, Silva, and N. Bianchi-Berthouze, “Cross-cultural differences in recognizing affect from body posture,” *Interacting with Computers*, vol. 18, no. 6, pp. 1371–1389, 2006.

[79] H. Gunes and M. Pantic, “Automatic, dimensional and continuous emotion recognition,” *Int’l Journal of Synthetic Emotion*, vol. 1, no. 1, pp. 68–99, 2010.

[80] C. M. Whissell, *The Dictionary of Affect in Language.* Academic Press, 1989, vol. 4, pp. 113–131.

[81] C. M. Lee and S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 2, pp. 293–303, 2005.

[82] M. Nicolaou, H. Gunes, and M. Pantic, “Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space,” *IEEE Transactions on Affective Computing*, vol. 2, pp. 92–105, 2011.

[83] R. A. Calvo and S. D’Mello, “Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications,” *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan. 2010.

[84] H. Joho, J. Staiano, N. Sebe, and J. Jose, “Looking at the viewer: analysing facial activity to detect personal highlights of multimedia contents,” *Multimedia Tools and Applications*, vol. 51, no. 2, pp. 505–523, October 2010.

[85] H. Joho, J. M. Jose, R. Valenti, and N. Sebe, “Exploiting facial expressions for affective video summarisation,” in *Proceeding of the ACM International Conference on Image and Video Retrieval*, ser. CIVR ’09. New York, NY, USA: ACM, 2009.

[86] A. Yazdani, J.-S. Lee, and T. Ebrahimi, "Implicit emotional tagging of multimedia using EEG signals and brain computer interface," in *Proc. SIGMM Workshop on Social media*, 2009, pp. 81–88.

[87] A. Hanjalic, "Adaptive Extraction of Highlights From a Sport Video Based on Excitement Modeling," *Multimedia, IEEE Transactions on*, vol. 7, no. 6, pp. 1114–1122, 2005.

[88] I. Arapakis, Y. Moshfeghi, H. Joho, R. Ren, D. Hannah, and J. M. Jose, "Integrating facial expressions into user profiling for the improvement of a multimodal recommender system," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, July 2009, pp. 1440–1443.

[89] I. Arapakis, I. Konstas, and J. M. Jose, "Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance," in *Proceedings of the seventeen ACM international conference on Multimedia*, ser. MM '09. New York, NY, USA: ACM, 2009, pp. 461–470.

[90] R. B. Dietz and A. Lang, "Aefective agents: Effects of agent affect on arousal, attention, liking and learning," in *Cognitive Technology Conference*, 1999.

[91] J. D. Laird, *Feelings: The Perception of Self*, 1st ed. USA: Oxford University Press, Jan. 2007.

[92] A. Kittur, E. H. Chi, and B. Suh, "Crowdsourcing user studies with Mechanical Turk," in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 453–456.

[93] M. Quirin, M. Kazén, and J. Kuhl, "When Nonsense Sounds Happy or Helpless: The Implicit Positive and Negative Affect Test (IPANAT)," *Journal of Personality and Social Psychology*, vol. 97, no. 3, pp. 500–516, 2009.

[94] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*. IEEE, 2005, pp. 317–321.

[95] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner, N. Amir, and K. Karpouzis, "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data," in *Affective Computing and Intelligent Interaction*, ser. Lecture Notes in Computer Science, A. Paiva, R. Prada, and R. Picard, Eds. Berlin, Heidelberg: Springer Berlin/Heidelberg, 2007, vol. 4738, ch. 43, pp. 488–500.

[96] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *Multimedia and Expo, 2008 IEEE International Conference on*, april 2008, pp. 865–868.

[97] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, "The SEMAINE corpus of emotionally coloured character interactions," in *Multimedia and Expo (ICME), 2010 IEEE International Conference on*, July 2010, pp. 1079–1084.

[98] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, october 2010.

[99] M. F. Valstar and M. Pantic, "Induced Disgust, Happiness and Surprise: an Addition to the MMI Facial Expression Database," in *Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION*, Malta, May 2010, pp. 65–70.

[100] A. Savran, K. Ciftci, G. Chanel, J. C. Mota, L. H. Viet, B. Sankur, L. Akarun, A. Caplier, and M. Rombaut, "Emotion detection in the loop from brain signals and facial images," in *Proceedings of the eNTERFACE 2006 Workshop*, Dubrovnik, Croatia, July 2006.

[101] J. Lichtenauer, M. Valstar, J. Shen, and M. Pantic, “Cost-effective solution to synchronized audio-visual capture using multiple sensors,” in *AVSS ’09: Proceedings of the 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance.* Washington, DC, USA: IEEE Computer Society, 2009, pp. 324–329.

[102] T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on.* IEEE Press, 2000, pp. 46–53.

[103] Pantic, M. and Rothkrantz, L. J. M., “Toward an affect-sensitive multimodal human-computer interaction,” *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1370–1390, 2003.

[104] S. Petridis and M. Pantic, “Is this joke really funny? judging the mirth by audiovisual laughter analysis,” in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, 2009, pp. 1444 –1447.

[105] P. Ekman, *Commentaries: Duchenne and facial expression of emotion*, ser. Studies in Emotion and Social Interaction. Cambridge University Press, Nov. 2006, pp. 270–284.

[106] M. E. Dawson, A. M. Schell, and D. L. Filion, *The electrodermal system*, 2nd ed. New York, NY, US: Cambridge University Press, 2000, pp. 200–223.

[107] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, “Basic emotions are associated with distinct patterns of cardiorespiratory activity.” *International journal of psychophysiology*, vol. 61, no. 1, pp. 5–18, July 2006.

[108] R. McCraty, M. Atkinson, W. A. Tiller, G. Rein, and A. D. Watkins, “The effects of emotions on short-term power spectrum analysis of heart rate variability,” *The American Journal of Cardiology*, vol. 76, no. 14, pp. 1089–1093, 1995.

[109] R. A. McFarland, “Relationship of skin temperature changes to the emotions accompanying music,” *Applied Psychophysiology and Biofeedback*, vol. 10, pp. 255–267, 1985.

[110] S. D. Kreibig, “Autonomic nervous system activity in emotion: A review,” *Biological Psychology*, vol. 84, no. 3, pp. 394–421, july 2010.

[111] R. Adolphs, D. Tranel, and A. R. Damasio, “Dissociable neural systems for recognizing emotions,” *Brain and Cognition*, vol. 52, no. 1, pp. 61–69, June 2003.

[112] A. R. Damasio, T. J. Grabowski, A. Bechara, H. Damasio, L. L. B. Ponto, J. Parvizi, and R. D. Hichwa, “Subcortical and cortical brain activity during the feeling of self-generated emotions,” *Nature Neuroscience*, vol. 3, no. 10, pp. 1049–1056, October 2000.

[113] F. H. Kanfer, “Verbal rate, eyeblink, and content in structured psychiatric interviews,” *Journal of Abnormal and Social Psychology*, vol. 61, no. 3, pp. 341–347, 1960.

[114] M. E. Tipping, “Sparse Bayesian Learning and the Relevance Vector Machine,” *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[115] R. J. Davidson, “Affective neuroscience and psychophysiology: toward a synthesis.” *Psychophysiology*, vol. 40, no. 5, pp. 655–665, September 2003.

[116] L. I. Aftanas, N. V. Reva, A. A. Varlamov, S. V. Pavlov, and V. P. Makhnev, “Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics.” *Neuroscience and behavioral physiology*, vol. 34, no. 8, pp. 859–867, October 2004.

[117] P. D. Welch, “The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms,” *IEEE Transactions on Audio and Electroacoustics*, vol. 15, pp. 70–73, 1967.

[118] S. K. Sutton and R. J. Davidson, “Prefrontal Brain Asymmetry: A Biological Substrate of the Behavioral Approach and Inhibition Systems,” *Psychological Science*, vol. 8, no. 3, pp. 204–210, 1997.

[119] V. F. Pamplona, M. M. Oliveira, and G. V. G. Baranoski, "Photorealistic models for pupil light reflex and iridal pattern deformation," *ACM Trans. Graph.*, vol. 28, no. 4, pp. 1–12, 2009.

[120] A. Longtin and J. Milton, "Modelling autonomous oscillations in the human pupil light reflex using non-linear delay-differential equations," *Bulletin of Mathematical Biology*, vol. 51, no. 5, pp. 605–624, September 1989.

[121] H. Bouma and L. C. J. Baghuis, "Hippus of the pupil: Periods of slow oscillations of unknown origin," *Vision Research*, vol. 11, no. 11, pp. 1345–1351, 1971.

[122] D. Ruta and B. Gabrys. (2000) An Overview of Classifier Fusion Methods.

[123] L. I. Kuncheva, *Combining Pattern Classifiers: Methods and Algorithms.* Wiley-Interscience, july 2004.

[124] J. C. Platt, *Probabilities for SV Machines.* MIT Press, 2000, pp. 61–74.

[125] T. F. Wu, C. J. Lin, and R. C. Weng, "Probability Estimates for Multi-class Classification by Pairwise Coupling," *J. Mach. Learn. Res.*, vol. 5, pp. 975–1005, 2004.

[126] C. Chang and C. Lin, "LIBSVM: a Library for Support Vector Machines," 2001.

[127] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on*, vol. PP, no. PP, p. In Press, 2011.

[128] S. D. Kreibig, G. Schaefer, and T. Brosch, *Psychophysiological response patterning in emotion: Implications for affective computing.* Oxford, UK: Oxford University Press, 2010, ch. 2.4, pp. 105–130.

[129] J. J. M. Kierkels and T. Pun, "Simultaneous exploitation of explicit and implicit tags in affect-based multimedia retrieval," in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*, September 2009, pp. 1–6.

[130] D. Watson, L. A. Clark, and A. Tellegen, "Development and validation of brief measures of positive and negative affect: the PANAS scales." *Journal of personality and social psychology*, vol. 54, no. 6, pp. 1063–1070, June 1988.

[131] P. Kelm, S. Schmiedeke, and T. Sikora, "Feature-based video key frame extraction for low quality video sequences," in *Proc. Int. Workshop on Image Analysis for Multimedia Interactive Services*, May 2009, pp. 25 –28.

[132] B. Janvier, E. Bruno, T. Pun, and S. Marchand-Maillet, "Information-theoretic temporal segmentation of video and applications: multiscale keyframes selection and shot boundaries detection," *Multimedia Tools and Applications*, vol. 30, no. 3, pp. 273–288, Sep. 2006.

[133] R. W. Picard, *Affective Computing.* MIT Press, Sep. 1997.

[134] L. Lu, H. Jiang, and H. Zhang, "A robust audio classification and segmentation method," in *Proc. ACM Int. Conf. Multimedia*, Ottawa, Canada, 2001, pp. 203–211.

[135] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee, "Classification of general audio data for content-based retrieval," *Pattern Recognition Letters*, vol. 22, no. 5, pp. 533–544, 2001.

[136] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[137] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 52–64, 2005.

[138] L. Chen, S. Gunduz, and M. Ozsu, "Mixed type audio classification with support vector machine," in *Proc. Int. Conf. Multimedia and Expo*, Toronto, Canada, July 2006, pp. 781 –784.

[139] M. Xu, L.-T. Chia, H. Yi, and D. Rajan, "Affective content detection in sitcom using subtitle and audio," in *Multi-Media Modelling Conference Proceedings, 2006 12th International*, 2006.

[140] A. Stolcke, X. Anguera, K. Boakye, O. Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "Multimodal technologies for perception of humans," R. Stiefelhagen, R. Bowers, and J. Fiscus, Eds. Berlin, Heidelberg: Springer-Verlag, 2008, ch. The SRI-ICSI Spring 2007 Meeting and Lecture Recognition System, pp. 450–463.

[141] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995.

[142] R. Fagin, R. Kumar, and D. Sivakumar, "Comparing top k lists," in *SODA '03: Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms.* Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 2003, pp. 28–36.