Queen Mary
University of London

School of Electronic Engineering and Computer Science

Affective and Implicit Tagging using Facial Expressions and Electroencephalography

Thesis submitted to the University of London in partial fulfillment
of the requirements for the degree of Doctor of Philosophy.

Author:
Reinder Alexander Lambertus (Sander) Koelstra

Supervisor:
Dr. Ioannis Patras

London, March 21, 2012

Declaration

I hereby declare that this dissertation is entirely the result of my own work, it arises out of my own research, and I have made full acknowledgements of the work and ideas of any other people who are cited in the thesis, or who have contributed to it.

Sander Koelstra

Acknowledgements

First, I would like to thank my supervisor Ioannis Patras, whose support and advice were invaluable in the production of this thesis. I'd like to thank my colleagues (and ex-colleagues) in the MMV group for allowing me to bug them with questions and the many lunches and half-moon outings we shared together. Thanks also go out to my colleagues in SpudTV for the great collaboration over the years.

I'd like to thank my lovely girlfriend Zoi for her support and for putting up with me in the last months as I avoided most human contact. Thanks go out to my flatmate and friend Brais for many discussions regarding my thesis and for pushing me to finish it already.

Lastly, I would like to thank all the friends I made since coming to London, all my friends back home (de lappen) and of course my family for their support in the last 4 years.

Abstract

Recent years have seen an explosion of user-generated, untagged multimedia data, generating a need for efficient search and retrieval of this data. The predominant method for content-based tagging is through manual annotation. Consequently, automatic tagging is currently the subject of intensive research. However, it is clear that the process will not be fully automated in the foreseeable future. We propose to involve the user and investigate methods for implicit tagging, wherein users' responses to the multimedia content are analysed in order to generate descriptive tags. We approach this problem through the modalities of facial expressions and EEG signals.

We investigate tag validation and affective tagging using EEG signals. The former relies on the detection of event-related potentials triggered in response to the presentation of invalid tags alongside multimedia material. We demonstrate significant differences in users' EEG responses for valid versus invalid tags, and present results towards single-trial classification. For affective tagging, we propose methodologies to map EEG signals onto the valence-arousal space and perform both binary classification as well as regression into this space. We apply these methods in a real-time affective recommendation system.

We also investigate the analysis of facial expressions for implicit tagging. This relies on a dynamic texture representation using non-rigid registration that we first evaluate on the problem of facial action unit recognition. We present results on well-known datasets (with both posed and spontaneous expressions) comparable to the state of the art in the field.

Finally, we present a multi-modal approach that fuses both modalities for affective tagging. We perform classification in the valence-arousal space based on these modalities and present results for both feature-level and decision-level fusion. We demonstrate improvement in the results when using both modalities, suggesting the modalities contain complementary information.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Contents

1.1 Motivation

Given the enormous amounts of unannotated video data available nowadays, the need for automatic categorisation and labelling of video material to enable efficient searching and retrieval is evident. To get a feel for the amount of generated video data, consider the quote of the co-founder of Youtube, stating that 13 hours of video content are uploaded every minute[1]. Up to now, the mainly used method for labelling video data is through manual annotation. This is a slow, laborious process that cannot possibly keep up with the growth of available data. To overcome this problem, research is performed to automate the annotation of video. So far, under restrictive conditions and for specific domains it has become possible to perform automatic content-based annotation with some accuracy. However, it is slowly becoming clear that we will not be able to fully automate the process for general video annotation in the foreseeable future. Therefore, we propose to keep the user in the loop by using methods for implicit tagging. More specifically, we aim to record and analyse users' passive physiological responses as they watch video clips. We approach the challenge of observing the user and gaining useful information on their implicit reactions to the videos through two different modalities; facial expressions and EEG signals. Facial expressions are probably the most information-rich modality for gaining information about a participant

[1] http://googleblog.blogspot.com/2008/09/future-of-online-video.html

passively (i.e. without conscious communication such as via speech). On the other hand, EEG signals may give us clues as to participants' internal states that are not or barely perceptible via other means. We intend to fuse the two modalities to take advantage of both properties. We show that the two modalities are, at least partly, complementary and that we can gain better insight by combining them rather than analysing either one separately.

This chapter first introduces the concept of implicit tagging in more detail in Section 1.2. Next, we introduce the used modalities for implicit tagging, namely EEG signals (Section 1.3) and facial expressions (Section 1.4). Section 1.5 summarizes the contributions of this work. Finally, Section 1.6 gives an overview of the organization of the thesis.

1.2 Implicit tagging

Implicit tagging concerns the automated annotation of multimedia data by analysis of users' behaviour. The advantage over explicit tagging is that it is done passively and requires little or no active involvement from users (other than their usual interaction with the multimedia data). Where explicit tagging can be a slow and labour-intensive process, implicit tagging is done in the background, can be performed each time a multimedia item is viewed and can deliver a wealth of annotation that can be used in search and indexing of multimedia data and can be combined with any tags derived explicitly (e.g. through crowd-sourcing of tags). In addition, implicit tagging can be used to assess the validity of existing tags, as well as for user profiling (storing particular preferences of a user based on his reactions to content)[155].

So far, tagging has mostly been done explicitly and manually by humans, or automatically using computer vision algorithms. Both types of existing tags suffer from various drawbacks. Manual tagging has seen a great increase with the rise of social media websites, allowing users to attach tags to uploaded multimedia items. However, as mentioned in [155], users do not typically tag with the intent of enriching the data for automated search and indexing. Instead users tag based on their own personal and social needs, bringing the value of these tags into question. Explicit tagging can also be done professionally. For instance, trained Facial Action Coding System (FACS) annotators can be hired to provide detailed annotations of facial expressions in image sequences. Recent years have seen an increase also in the use of untrained annotators, for instance through the use of Amazon's Mechanical Turk[2] platform. In this scenario, short annotation tasks are made available online and can be performed for a

[2] Amazon Mechanical Turk: https://www.mturk.com/

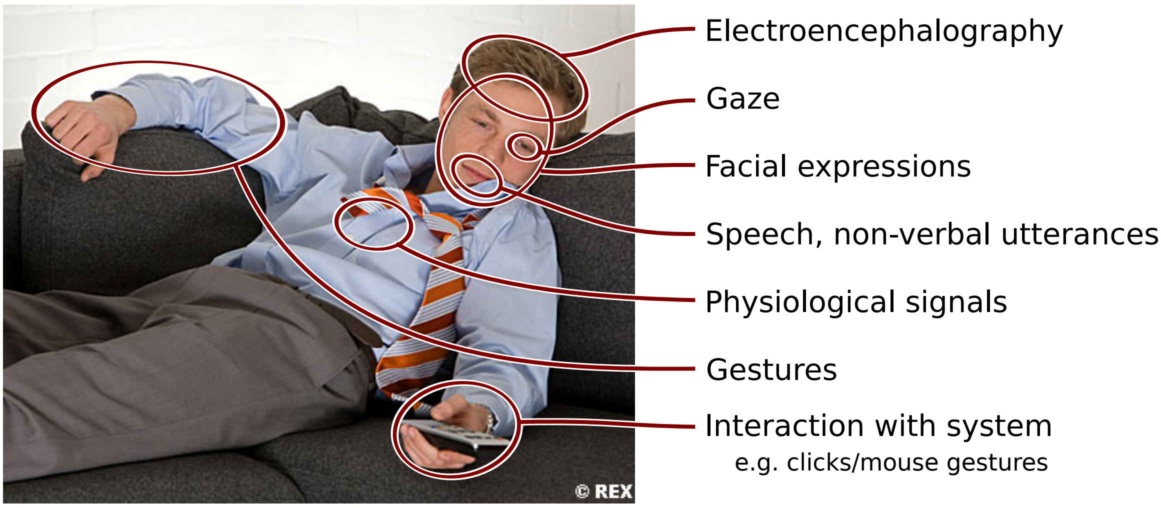


Figure 1.1: Illustration of the different modalities that can be used in observation-based implicit tagging.

small payment by any Mechanical Turk user. Paid annotation is in many cases cost-prohibitive, and when using cheaper solutions such as Mechanical Turk, reliability of the annotators is often in question.

Machine-based tagging suffers from the issue of the semantic gap, where tags may include recognized objects/locations/faces, the detected amount of motion, shot transition speed, etc. but tagging algorithms still have great difficulty assigning semantic meaning to multimedia items, such as detecting plot keywords or affective content. Implicit tagging may be able to alleviate some of these shortcomings.

Broadly speaking, there exist two approaches to the problem of implicit tagging: game-based tagging and observation-based tagging. In the former, the tagging of the data is a by-product of playing a game. The most well-known of these approaches is probably the ESP game[156]. In the game, two users are paired and are given the task of assigning tags to an image. Points are awarded when the same tag is given by both users. The users do not know each other and can not communicate, so their only way to score points is to assign straightforward tags to the image in the hope that their counterparts will assign the same ones. Since the publication of this work, several authors have expanded and refined this concept for a diverse set of tag types such as object localization[157], music metadata[100] and moods[76].

In observation-based tagging, users' responses as they view the media are recorded and analysed in order to extract tags describing the media. User responses can be recorded from a variety of modalities (see Fig. 1.1).

In this work, we focus on the modalities of facial expressions and EEG signals. Of all the modalities, facial expressions probably are the most informative, while

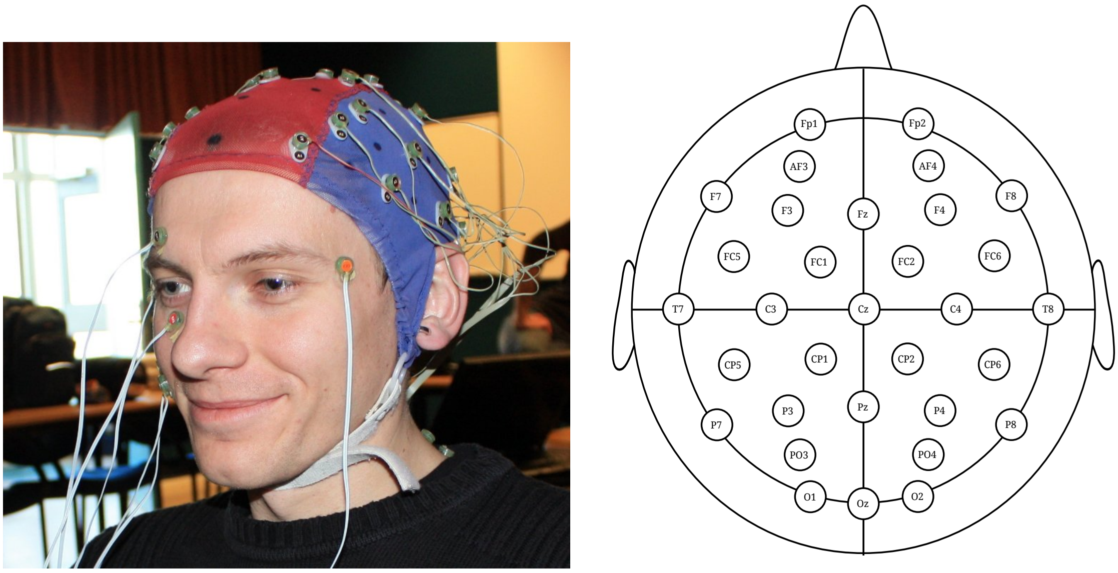


Figure 1.2: Left: Participant with cap and 32 EEG electrodes. Right: the 10-20 system for electrode placement.

EEG signals may reveal some otherwise unobservable affective states and may well complement the former modality.

We consider two uses for observation-based implicit tagging. Firstly, we investigate the possibility of implicitly validating previously assigned tags using EEG signals. Secondly, we explore the possibilities for implicit affective tagging. That is, we aim to classify the user's affective response to multimedia items using the modalities of EEG signals and facial expressions.

1.3 Electroencephalography

The first modality we exploit is the use of Electroencephalography (EEG) signals to assess a participant's reactions. EEG is a non-invasive technique for measuring a participant's brainwaves, recorded through electrodes placed on the scalp. A cap is placed on the participant's head, electrode gel is applied to ensure good conductivity and then electrodes are attached to the cap. An international standard known as the 10-20 system determines the location of the electrodes on the scalp (see Fig. 1.2).

EEG systems offer advantages for implicit tagging over other brain imaging systems (such as fMRI), namely the portability and relatively low cost of the system, as well as a much higher temporal resolution (in the order of milliseconds). The disadvantage of EEG is a much weaker signal compared to other techniques as well as a relatively poor spatial resolution. The weak signal is due to attenuation of the

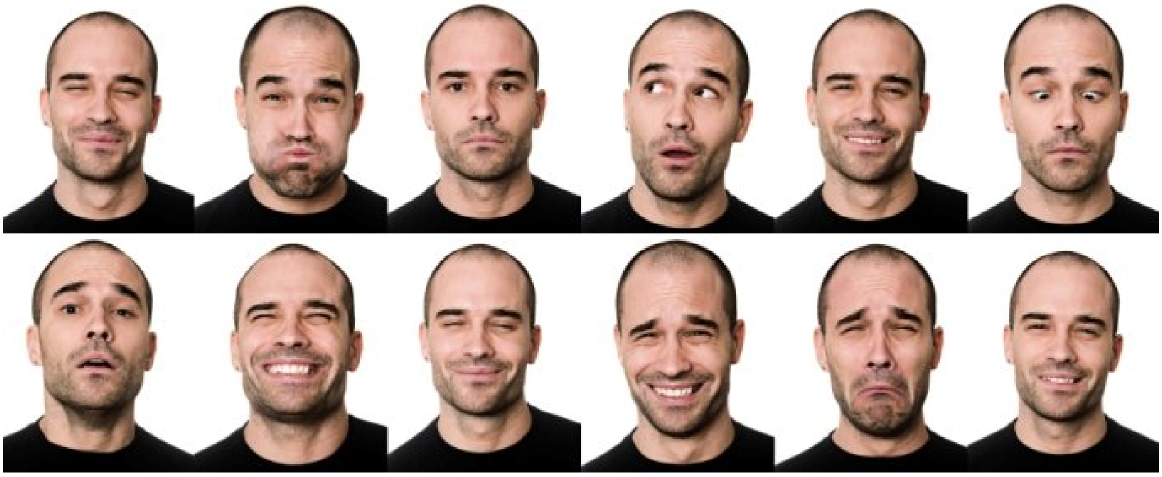


Figure 1.3: Example of the variety in facial expressions.

electrical activity by the skull and interference from other electrical sources (mains line, participant movement). In addition, the electrical activity measured at a single electrode is a mixture emanating from many different neurons and it is difficult to exactly localise the source. Nevertheless, in recent years EEG has begun to see various applications, such as controlling a mouse cursor[165], spelling by focusing on the correct letters on a virtual keyboard[80] and controlling a motorized wheelchair using imagined hand movements[125]. While at the moment the recording of EEG measurements is still quite a cumbersome process, recent improvements in the development of dry electrodes and the emergence of commercial EEG equipment intended for personal use may simplify the use of this modality and make it usable outside of the laboratory environment.

In Chapter 3, we describe a method for implicit tag validation based on EEG signal analysis. Chapter 4 describes our work on assessing participants' emotions as they watch a set of music videos.

1.4 Facial Expressions

Facial expressions form one of the most information-rich modalities for assessing a participant's internal state implicitly. They are one of the most cogent, naturally pre-eminent means for human beings to communicate emotions, 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[42]. Automatic analysis of facial expressions forms, therefore, the essence of numerous next-generation-computing tools including affective computing technologies (i.e. proactive and affective user interfaces, implicit tagging), learner-adaptive tutoring systems, patient-profiled personal wellness technologies, etc.[107].

Automated analysis of facial behaviour, has attracted increasing attention in computer vision, pattern recognition, and human-computer interaction. Facial expression analysis is a challenging problem due to the thousands of possible facial expressions (see for instance Fig. 1.3) and the often subtle differences between them. Despite decades of research, a foolproof solution to the problem still does not exist.

Chapter 5 explains in detail the work we performed on automatic facial Action Unit recognition. This work culminated in a system capable of recognizing Action Units (facial muscle action descriptors) with accuracy comparable to the state of the art. In Chapter 6 this work is combined with the work done on EEG signal analysis to perform multi-modal affective implicit tagging. Our hypothesis is that the two modalities are (partially) complementary, i.e. some emotional states are easier to read in facial expressions (smiling when something is funny), whereas the face may not always show any visual clues for other states (for example interest level).

1.5 Contributions

The primary aim of this work is the evaluation of facial expressions and EEG signals as possible modalities for affective implicit tagging. Within this framework, several contributions are identified:

- In Chapter 3, we present the case for the use of EEG signal analysis for detection of event-related potentials caused by the display of non-matching tags displayed in conjunction with video clips. We demonstrate significant differences in EEG responses to the display of matching versus the display of non-matching tags. We propose a method for single-trial classification that is shown to perform above chance level, indicating that it can be used for implicit tag-validation, especially when aggregating the tag validation over multiple participants. The dataset recorded during this experiment is made publicly available.

- In Chapter 4 we present methods for automatically detecting the emotional state (in terms of arousal and valence) of participants presented with music videos by analysis of their EEG signals. A series of experiments was performed in order to test the developed methods. Single-trial classification and regression results are significantly higher than chance level.

- In Chapter 5 we propose a method for frame-by-frame action unit classification in face videos based on dynamic textures (DT). We combine the use of Free-form Deformations, a non-uniform decomposition of the facial area based on quadtrees for feature extraction, a frame-based GentleBoost classifier and a dynamic, generative HMM model. This is one of the few DT-based methods for AU recognition proposed and is shown to outperform the state of the art.

- In Chapter 6, we propose methods for multi-modal affect estimation by fusing facial expressions and EEG signals. We show that feature-level fusion consistently outperforms the single modalities, and decision-level fusion does so in many cases. We demonstrate that the predicted affective state tags, when aggregated over multiple participants, performs comparably to the agreement level between humans. To the best of our knowledge, this is the first work concerning the fusion of the EEG and facial expression modalities for affect estimation.

- In Appendix A, we describe the implementation of a real-time affective music video recommendation system . To the best of our knowledge, this is the first such system ever to be implemented.

- In Appendix B, we present a publicly available dataset of EEG, facial video and peripheral physiological signal recordings for affective state assessment. A novel semi-automatic stimuli selection method is introduced. To the best of our knowledge, this is the most extensive dataset of its kind in terms of the number of participants and the variety of recorded modalities. As of the writing of this thesis, researchers from 36 different institutions have requested access to the dataset.

The mentioned contributions were published in two articles in IEEE Transactions journals[3, 1] and four articles in international conferences[4, 2, 6, 5]. See also the Publications section for the full listing.

1.6 Thesis Overview

The remainder of this thesis is organized as follows: Chapter 2 gives an overview of the literature in the fields of emotion modelling, EEG signal analysis and facial expression analysis methods. Chapter 3 describes in detail our research in using EEG signal analysis methods for the implicit validation of tags displayed in conjunction with video content. Chapter 4 details our work on methods for affective implicit tagging using EEG signal analysis. Chapter 5 explains our research into methods for automatic recognition of facial action units. Chapter 6 describes our approach to fusion of the EEG and facial expression modalities for affective state estimation and implicit tagging. Chapter 7 concludes the thesis.

Appendix A describes a real-time affective music video recommendation demonstration system implemented using the methods described in Chapter 4. Appendix B describes in more detail the publicly available DEAP dataset, which is composed of the data that was gathered for experiments described in Chapter 4.

Chapter 2

State of the art

Contents

In this Chapter, we describe the state of the art for the various aspects of our research. In order to study affective implicit tagging, a model is needed for representing human emotions. Section 2.1 describes various proposed models of human affect. An overview of the research in EEG signal analysis and its applicability to implicit tagging is given in Section 2.2. Finally, Section 2.3 describes the literature in automated facial expression recognition. Section 2.4 concludes the chapter.

2.1 Models of human emotion

2.1.1 Discrete models of emotion

Various discrete categorizations of emotions have been proposed, such as the six universal basic emotions (surprise, anger, happiness, sadness, fear and disgust, see Fig.

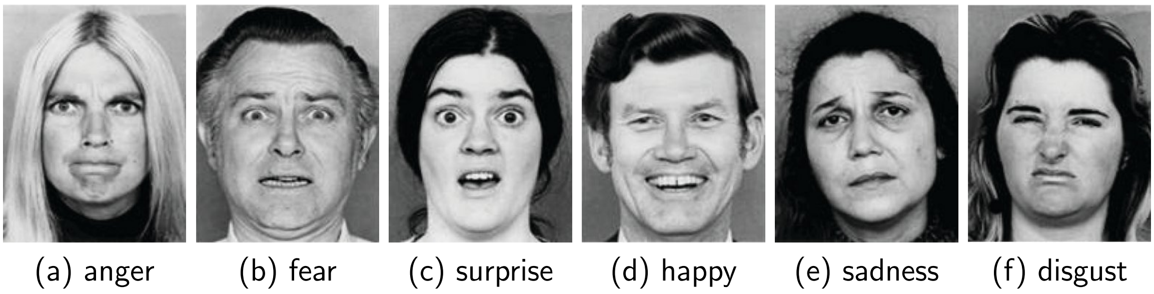

(a) anger (b) fear (c) surprise (d) happy (e) sadness (f) disgust

Figure 2.1: The six universal basic emotions[41]

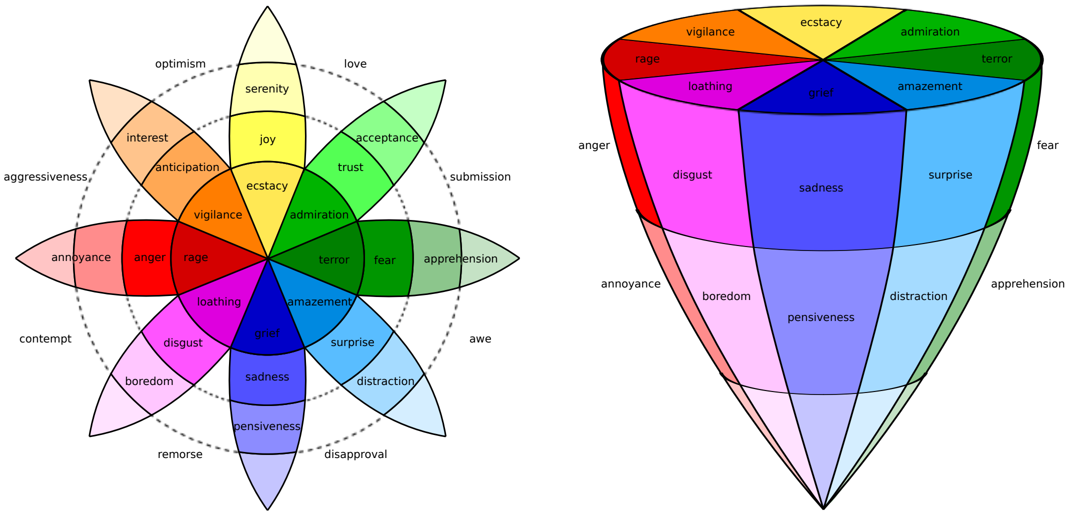


Figure 2.2: Plutchik's emotion wheel[119]. Left: 2D projected version, right: 3D conical version.

2.1) as proposed by Ekman and Friesen[41]. These expressions were shown to be displayed and recognised through facial expressions by participants from different cultures in the same fashion . Ekman later expanded the list of basic emotions to include amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure and shame[39].

W.G. Parrot proposed a tree structure of primary, secondary and tertiary emotions[114] with the primary emotions being love, joy, surprise, anger, sadness and fear.

In the approach of Plutchik[119], emotions are also distinguished into primary, secondary and tertiary emotions, with the primary emotions being anger, fear, sadness, disgust, surprise, anticipation, trust and joy. Additionally, he arranges these emotions in a conical shape, indicating the relations between them (see Fig. 2.2).

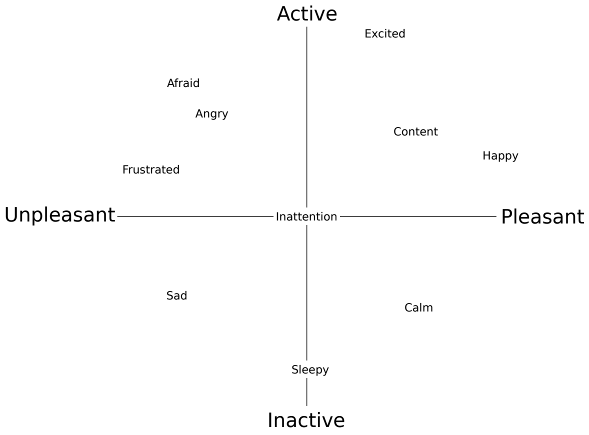


Figure 2.3: An illustration of the arousal-valence scale[127].

2.1.2 Continuous models of emotion

Continuous models of emotion have also been proposed. The arousal-valence scale, first proposed by Russell[127] is a much used scale in research on affect. The concept is that each emotional state can be placed on a two-dimensional plane with arousal and valence as the axes. Arousal can range here from inactive (e.g. uninterested, bored) to active (e.g. alert, excited), whereas valence ranges from unpleasant (e.g. sad, stressed) to pleasant (e.g. happy, elated). Figure 2.3 illustrates the concept.

While arousal and valence explain most of the variation in emotional states, it has been argued that a third dimension of control or dominance and even a fourth dimension of unpredictability should be added to the model[48]. Dominance ranges from a helpless and weak feeling (without control) to an empowered feeling (in control of everything). Unpredictability is related to the novelty or spontaneity of an emotion.

2.1.2.1 Self-assessment of emotions

Several tools have been proposed to facilitate self-assessment on the continuous scales.

One such tool is the Self-Assessment Manikin (SAM)[23, 103] as shown in Figure 2.4. For each of the valence, arousal and dominance dimensions there is a series of manikins visualizing the different values along the axes. Experiment participants can then select for each dimension the manikin that most closely expresses their felt emotion.

Another tool is the Geneva emotion wheel (GEW)[132]. The tool visualizes the valence-arousal space by relating it to a series of emotional keywords that represent the extremes in the space. Each keyword is in turn represented by a set of different-sized circles, where the size of the circle represents the strength of the felt emotion

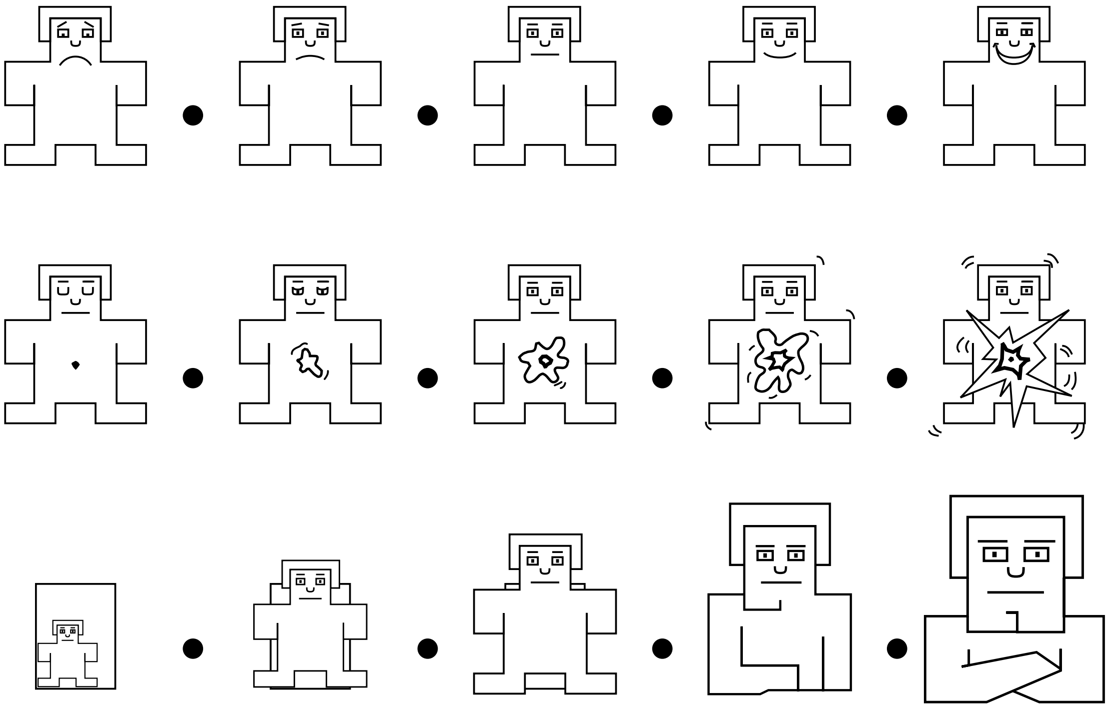


Figure 2.4: Self-Assessment Manikins[23, 103]. From top: valence, arousal, dominance.

(see Fig. 2.5). Participants can then select a keyword and a strength by clicking on one of the circles. The choice of keyword/strength has been shown to correlate with the location in the valence-arousal space[10].

2.1.3 Perceived and felt emotion

Perceived emotion can be described as the emotional characteristics a person can distinguish in a stimulus (e.g. that song is very sad). Felt emotion, on the other hand, concerns a person's emotional response to a stimulus (e.g. the song makes me feel sad). Several studies [50, 67] have shown that there are substantial differences between perceived and felt emotions. One can also imagine a difference in emotional rating for the stimulus in general, versus the specific elicited response for a particular viewing of the stimulus. For these reasons, it is important to give correct and complete instructions to participants in order to avoid confusion.

In this work, unless otherwise specified, participants were always specifically asked for felt emotion, specific to the particular instance of viewing the stimulus. Thus, we are not interested in ratings of the stimulus, but solely in ratings of persons' emotional

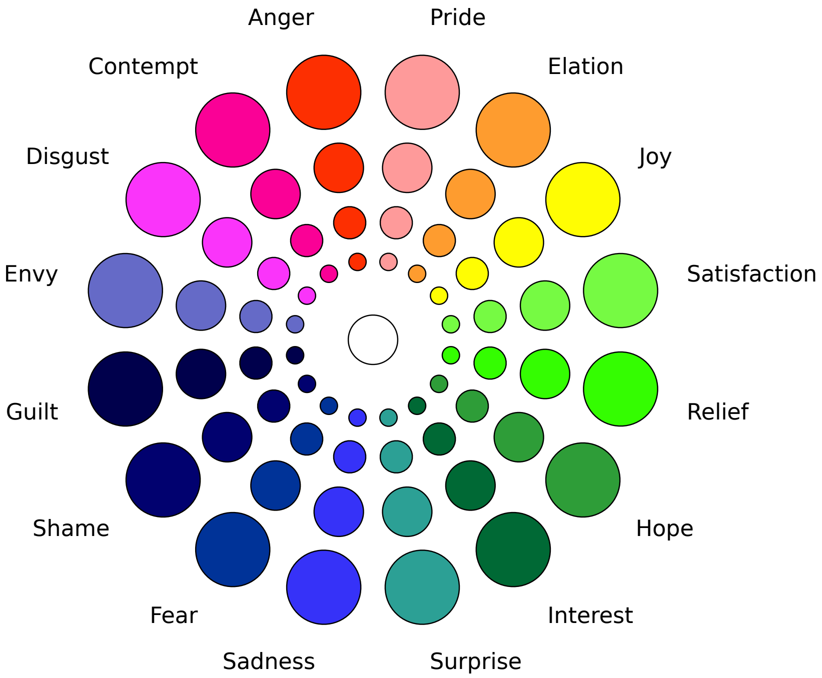


Figure 2.5: Geneva emotion wheel[132].

responses. The main object in this work is to analyse and classify felt emotions of participants, regardless of the specific stimulus. In other words, we aim to find classification functions that map from the raw (EEG) signal recordings directly to a person's felt emotion in single trials.

2.2 EEG Signal analysis

Traditionally, Electroencephalography (EEG) measurements have been used for clinical purposes, for instance to diagnose epilepsy and locate the origin of epileptic seizures. Patients that suffer from locked-in syndrome are almost completely paralysed and incapable of communicating with the outside world, in spite of retaining functionality in the upper brain and full consciousness and awareness of their surroundings. For these cases, several applications have been developed that use EEG to restore some communicative abilities, for instance a system for spelling words[47] and a system to move a mouse cursor on a screen[165]. EEG is also frequently used in research for understanding cognitive processes[104]

More recently, research has begun in tapping the potential of using EEG as a brain computer interface (BCI). In this paradigm, EEG can also be used by healthy people as an additional method of human machine interaction. One emerging application of BCIs are their use in games. In Brainball[63], the ratio between participants' alpha and beta brainwaves is measured and used to control the movement of a ball. The

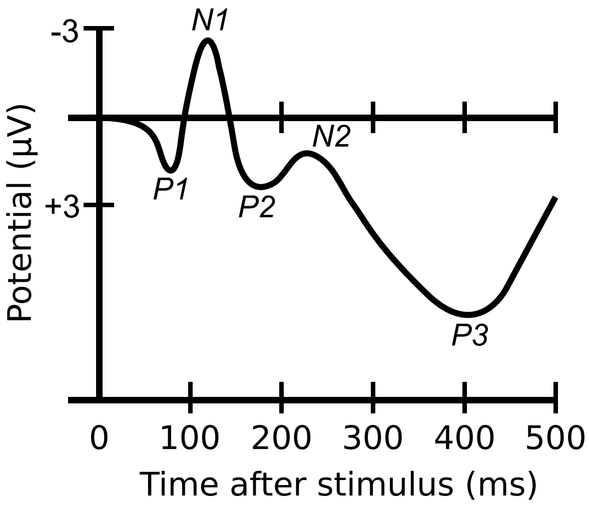


Figure 2.6: Illustration of Event-related potentials[1].

object of the game is to relax more than your opponent and thereby move the ball to the opponent's side of the table. The reader is referred to [118] for a review of other gaming approaches. Another application of EEG signal analysis is as a tool for multimedia tagging[51, 70, 71, 32, 74, 169], which is the main focus in this thesis and further described in Section 2.2.3.

2.2.1 EEG features

The analysis of EEG signals for brain-computer interfaces mainly focuses on event-related potentials (ERPs) or Spectral-power features. Other possible features, such as different kinds of evoked potentials, can also be used for games and control mechanisms. However, for implicit tagging they are not suitable as they require the direct stimulation of the senses, contradicting the passive nature of implicit tagging, and as such they are not discussed here.

2.2.1.1 Event-Related potentials (ERPs)

Many of the EEG applications rely on detecting event related potentials (ERPs), such as the P300, the N400 and the N170. Fig. 2.6 illustrates several of these ERPs. ERPs are time-locked responses (i.e. changes in EEG voltage) related to external events. ERPs represent electrical fields associated with populations of synchronously active neurons. As it is not possible to measure the electrical activity of single neurons at the scalp, the ERP must be a result of the combined activity of many neurons. For the activity to be measurable at the scalp, the activity of the neurons in the

[1] Image source: http://en.wikipedia.org/wiki/File:ComponentsofERP.svg

population must be synchronous and the fields they generate must be geometrically aligned in such a way as to accumulate and form a dipolar field. Neural structures meeting this alignment constraint are called ”open fields”. Furthermore, It is generally believed that ERPs are the result of postsynaptic (dendritic) potentials, as compared to presynaptic (axonal action) potentials, these potentials are slower, more likely to synchronize, and thus more likely to form fields measurable at the scalp[99, 44].

The voltage changes comprising an ERP are generally on the order of microvolts, where as the EEG waveform overall contains amplitudes of tens of microvolts. This means the ERP is not readily visible in the EEG and signal processing techniques are required to discriminate between the ERP and the background EEG. The predominant technique to increase this discrimination is the averaging of (synchronized) samples recorded in response to several repetitions of the same stimulus. The ERP is assumed to be time-locked to the external stimulus, in contrast to the background EEG, which is assumed to vary randomly from trial to trial. Therefore, the averaging should reduce the background EEG, whilst leaving the ERP mostly intact.

The method of averaging over multiple trials, while generally effective, poses several problems. It can not be guaranteed that the average waveform is an accurate reflection of the ERP as elicited by separate events. For instance, if the latency or the amplitude of the response varies between trials according to some distribution, the average ERP may differ substantially form the ERPs of individual events. Also, from the perspective of brain-computer interfaces, one can not use this kind of analysis in a real-time environment where the intention is to use the ERP as a control signal. It is often not feasible or desirable to display the same stimulus repeatedly to achieve a high enough level of averaging.

Therefore, research has begun into so-called single-trial analysis, where the goal is to detect the ERP in a single (or a few) trials. Early methods for ERP detection include template matching[166, 45] and peak-picking[115]. Lately, methods for ERP detection have grown more sophisticated. Techniques used for feature extraction include Principal Component Analysis (PCA)[148], Fisher Linear Discriminant analysis[51] and wavelets[21].

For the remainder of this section, we will list some well known ERPs, their properties and possible applications. The P300 ERP consists of a positive variation in the EEG signal occurring 300ms after stimulus presentation. It mainly appears when the user is shown unexpected or deviating stimuli. It is usually solicited in experiments using the ’oddball’ paradigm, where participants are shown a series of stimuli where one stimulus stands out in some way. An example application is the spelling technique described first in [47]. In this system, a 6x6 grid containing letters and commands is presented and rows and columns are highlighted alternately. The letter the participant is focussing on stands out from the rest and a P300 is observed when it's row or column are highlighted. Using this system, participants were able to spell text at 7.8 characters per minute. In [51, 113] the use of EEG in rapid image search is investigated. Participants are rapidly shown a series of standard, low-variation images for 100ms each with some images standing out (in this case, images of a forest with some containing a walking human). The images that stand out elicit a P300 response which is detected with a similar accuracy to a parallel experiment where users are asked to find the images with humans manually by clicking mouse buttons.

The N400 ERP presents itself as a negative voltage deflection in the EEG signal 400ms after presentation of the stimulus. It is usually elicited by semantic mismatches in the stimulus presentation. An example is the presentation of a sentence with a semantically incongruous ending[84]. This response has also been shown to occur with mismatches across different modalities, for instance the sound of barking and the display of the word 'dog'[106].

The N170 ERP (a negative voltage variation, 170ms after stimulus presentation) can be elicited reliably and more strongly when participants are shown images of human faces[17]. This may be a very useful characteristic when designing computer vision applications that use EEG signals.

2.2.1.2 Spectral-power features

Another approach in using EEG data is by analysing the spectral power of certain frequency bands in the signals. This technique is mainly used in more continuous experiments, where an ERP is not evoked at any specific time period. It has shown success mainly in detecting and distinguishing imagined movements. The most commonly used frequency bands are delta (0-4Hz), theta (4-7Hz), alpha (7-12Hz), beta (12-30Hz) and gamma (30-100+Hz) waves[104].

This approach has been used for instance in [80], where the workload of participants is monitored as they perform multiple tasks (including driving a car). The spectral power in frequency bands in the 3-15 Hz range, taken over 10-30s windows, is analysed. A Linear Discriminant Analysis (LDA) classifier is then trained to recognize the different workloads. Another example is the work described in [165], where participants train themselves to move a cursor on a computer screen by learning to vary the amplitude of their beta rhythms (for vertical movement) and alpha rhythms (for horizontal movement). In [90], participants imagine moving their hands or feet, triggering sensorimotor rhythms in the motor cortex of the brain, which are then used as a control mechanism to rotate an avatar in a virtual world. Many works use this paradigm of training the user to elicit certain brainwave patterns or use mental motor-imagery. However, we can not use such an approach in our case as we do not expect the user to be actively involved in the process but merely passively watching video sequences. So we must make do with any spectral power shifts brought on as a response to seeing the videos.

Techniques used for processing of spectral power features include common spatial patterns (CSP)[121, 36, 92, 19], independent component analysis (ICA)[62], adaptive auto-regressive (AAR) models[116] and discriminative spatial patters (DSP)[92].

2.2.2 Classification

For classification of the derived features, many researchers use linear discriminant analysis[121, 148, 36, 19], logistic regression[51, 146] or support vector machines (SVM)[92, 168, 62, 70].

2.2.3 EEG for multimedia tagging

The use of EEG in annotating multimedia data is a very new research direction and so far only a few works have investigated this area. In [51], an oddball paradigm is used in which images of a forest environment were shown to participants for 100 ms each. The goal was to detect a small subset of target images that contained pedestrians. The target images elicit a P300 event-related potential which was then classified using Fisher linear discriminant analysis. Another test was run without the EEG modality, where participants pressed a button upon seeing the target images. The results showed no significant differences in target image detection accuracy between the use of the EEG modality and the use of buttons.

In [70, 71], categories of images are classified based on EEG measurements recorded as the images were presented. The used categories were faces, animals and inanimate objects. This was based on the notion that the human visual system responds very differently to these categories of images. The authors propose a vision-based algorithm that uses pyramid match kernels to initially classify the images. The EEG data is then combined with the vision-based features using a kernel-alignment method. The combination of the two modalities outperforms the individual methods.

In [32] the Revolutionary Advanced Processing Image Detection (RAPID) system is proposed. The authors use ERP analysis in combination with eye tracking to assist intelligence analysts in rapidly reviewing and categorizing satellite imagery. The analyst is assigned a target category to look for in the images. When participants see an image in the target category, an ERP occurs in the EEG data which is then classified. Eye tracking is used to determine points of interest within the images.

Kierkels et al.[74] proposed a method for personalized affective tagging of multimedia using peripheral physiological signals. Valence and arousal levels of participants' emotions when watching videos were computed from physiological responses using linear regression[137]. 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.

Yazdani et al.[169] proposed using a brain computer interface (BCI) based on P300 evoked potentials to emotionally tag videos with one of the six Ekman basic emotions[41]. 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 behavioural and psycho-physiological responses.

There has been a large number of published works in the domain of emotion recognition from physiological signals[86, 75, 160, 60, 93, 25]. Of these studies, only a few achieved notable results using video stimuli. Lisetti and Nasoz used physiological responses to recognize emotions in response to movie scenes[93]. The movie scenes were selected to elicit 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.

2.3 Facial expression analysis

2.3.1 Facial features

Two main streams in the current research on automatic analysis of facial expressions consider facial affect (emotion) detection and facial muscle action (action unit) detection[110, 107, 171]. The most commonly used facial expression descriptors in facial affect detection approaches are the six basic emotions (fear, sadness, happiness, anger, disgust, surprise), proposed by Ekman and discrete emotion theorists, who suggest that these emotions are universally displayed and recognized from facial expressions. The most commonly used facial muscle action descriptors are the Action Units (AUs) defined in the Facial Action Coding System (FACS; [40]).

This categorization in terms of six basic emotions used in facial affect detection

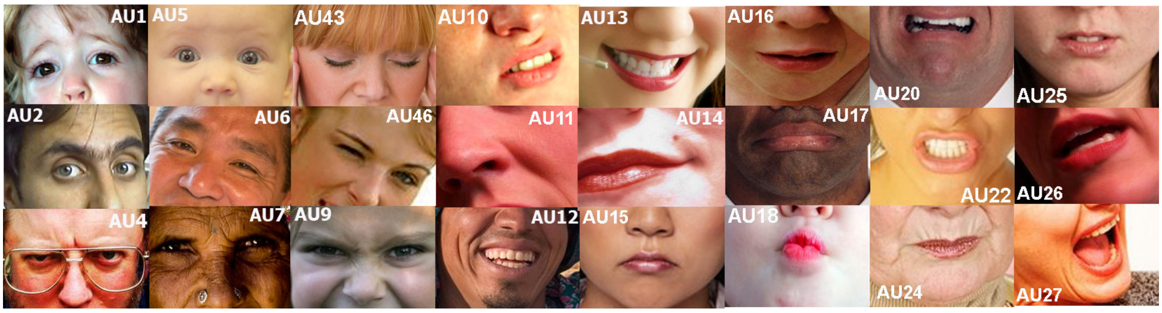


Figure 2.7: Apex phases of 22 AUs of the FACS system.

approaches, though quite intuitive, has some important downsides. The basic emotion categories form only a subset of the total range of possible facial displays and categorization of facial expressions can therefore be forced and unnatural. Boredom and interest, for instance, do not seem to fit well in any of the basic emotion categories. Moreover, in everyday life, these prototypic expressions occur relatively rarely; usually, emotions are displayed by subtle changes in discrete facial features, such as raising of the eyebrows in surprise. To detect such subtlety of human emotions automatic recognition of atomic facial signals, such as the AUs of the FACS system, is needed.

FACS was proposed by Ekman and Friesen in 1978 and revised in 2002[40]. FACS classifies atomic facial signals into Action Units (AUs) according to the facial muscles that cause them. It defines 9 upper face AUs and 18 lower face AUs, which are considered to be the smallest visually discernible facial movements. It also defines 20 Action Descriptors for eye and head position. FACS provides the rules both for AU intensity scoring and for recognition of temporal segments (onset, apex and offset) of AUs in a face video.

Most of the research on automatic AU recognition has been based on analysis of static images (e.g. [111]) or individual frames of an image sequence (e.g. [15, 16, 94, 82]). Some research efforts toward using dynamic textures (DT) for facial expression recognition (e.g. [153, 173]) and toward explicit coding of AU dynamics (e.g. with respect to AUs' temporal segments, like in [109, 152], or with respect to temporal correlation of different AUs like in [147]) have been proposed as well. However, most of these previously proposed systems recognise either the six basic emotions (e.g. [173]) or only subsets of the 27 defined AUs. Except for geometric-feature-based methods proposed in [108, 109, 152], none of the existing systems attains automatic recognition of AUs' temporal segments. Also, except for the method based on Motion History Images proposed in [153], none of the past works attempted automatic AU recognition using of a DT-based approach.

The existing approaches to facial expression analysis, can be divided into geometric and appearance-based approaches. Dynamic texture recognition can be seen as a generalisation of appearance-based approaches. Geometric features include shapes and positions of face components, as well as the location of facial feature points (such as the corners of the mouth). Often, the position and shape of these components and/or fiducial points are detected in the first frame and then tracked throughout the sequence. On the other hand, appearance-based methods rely on skin motion and texture changes (deformations of the skin) such as wrinkles, bulges and furrows. Both approaches have advantages and disadvantages. Geometric features only consider the motion of a number of points, so one ignores much information present in the skin texture changes. On the other hand, appearance-based methods may be more susceptible to changes in illumination and differences between individuals.

Bartlett et al.[14] reported that appearance-based systems outperform those based on geometric features. However, recent work by Pantic and Patras[109] has shown that this is not always the case. Using hybrid features has been reported to result in the best performance[144]. Hybrid features have been used by Tian et al.[142, 143], Zhang and Ji[172], Wen and Huang[161], Tong et al.[147], and others.

2.3.1.1 Geometric Features

Approaches that use only geometric features mostly rely on detecting sets of fiducial facial points (e.g. [111, 109, 152, 158]), a connected face mesh or active shape model (e.g. [52, 29, 26, 167, 141, 82]), or face component shape parametrization (e.g. [142]). Next, the points or shapes are tracked throughout the video and the utilized features are their relative and absolute position, mutual spatial position, speed, acceleration, etc. A geometric approach that attempts to automatically detect temporal segments of AUs is the work of Pantic and colleagues[108, 109, 152]. They locate and track a number of facial fiducial points and extract a set of spatio-temporal features from the trajectories. In [108] and [109], they use a rule-based approach to detect AUs and their temporal segments, while in [152] they use a combination of SVMs and HMMs to do so. Using only the movement of a number of feature points makes it difficult to detect certain AUs, such as AU 11 (nasolabial furrow deepener), 14 (mouth corner dimpler), 17 (chin raiser), 28 (inward sucking of the lips) (see also Fig. 2.7), the activation of which is not apparent from movements of facial points but rather from changes in skin texture. Yet, these AUs are typical for facial expressions of emotions such as sadness (see EMFACS[40]), and for expressions of more complex mental states including puzzlement and disagreement[42], which are of immense importance if the

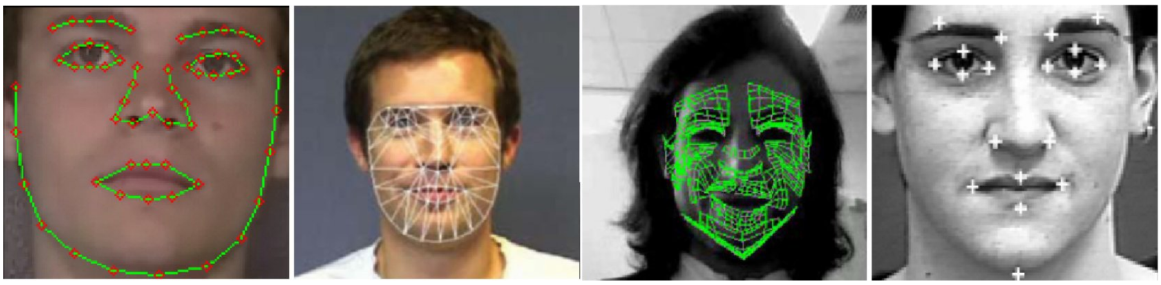


Figure 2.8: Examples of geometric feature extraction.

Left	Active appearance model used by Xiao et al. [167].
Center	3D meshes used by Tao and Huang [141].
Right	Facial feature points used by Vukadinovic and Pantic [158].

goal is to realize human-centred, adaptive interfaces. On the contrary, our appearance-based approach is capable of detecting the furrows and wrinkles associated with these AUs and is therefore better equipped to recognize them.

2.3.1.2 Appearance-based features

Systems using only appearance-based features have been proposed by for example Essa and Pentland[43], Bartlett et al.[14, 94, 15, 16], Guo and Dyer[57], Anderson and McOwan[7], Lucey et al.[98], and Valstar et al.[153].

The system developed by Essa and Pentland[43] fits a 3D dynamic physics-based face mesh on the face in the input video. Next, they use an optical flow method to find motion in the face region. This motion field is mapped to the face model. Then, they extract a motion energy image (depicting the amount of motion as grey level intensity) from the motion field, based on which prototypic facial expressions of emotions are further recognised. Anderson and McOwan[7] also used optical flow to create a motion field, extracting average motion at several regions in the face. The resulting motion signatures are classified using SVMs as either non-expressive or one of the prototypic facial expressions. Lucey et al.[98] fit active appearance models (AAMs, 2D triangulated meshes) to the face using a gradient descent algorithm. Then, they derive appearance and shape features from the AAM, and use SVMs to classify sequences as showing one of 15 AUs.

Several others have used Gabor wavelet coefficients as features (e.g. [57, 172, 161]). Bartlett et al.[14, 15, 94, 16] have tried different methods such as optical flow, explicit feature measurement (i.e. length of wrinkles, degree of eye opening), ICA (a generalization of PCA) and the use of Gabor wavelets. They report that the use of Gabor wavelets renders the best results[94]. Tian, Cohn and Kanade[142, 143,

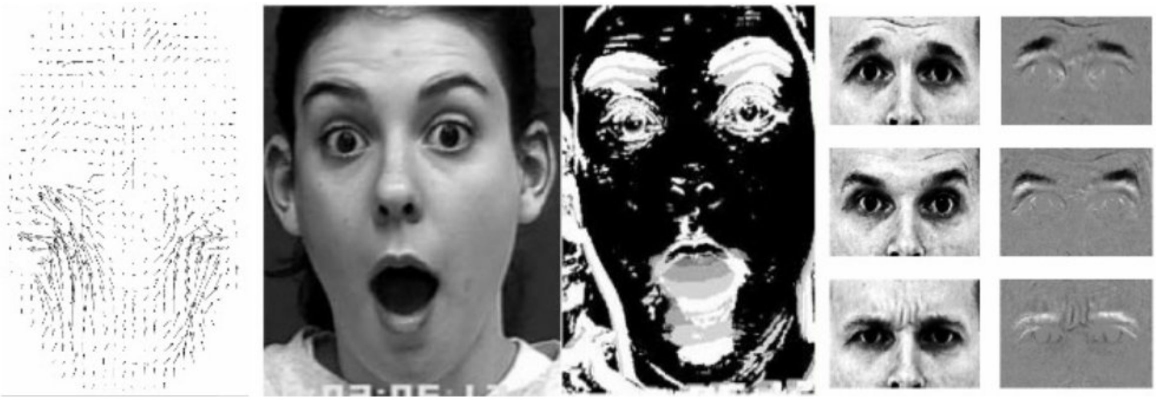


Figure 2.9: Examples of appearance-based feature extraction.
Left Optical flow of a smile used by Essa and Pentland[43].
Center Frame from a facial image sequence and generated MHI
as used by Valstar et al.[153].
Right Gabor features for different expressions used by Bartlett et al.[13].

144] use a combination of geometric (parametric descriptions of facial features) and appearance-based features (Gabor wavelets). They claim that the geometric features outperform the appearance-based ones, yet using both yields the best result. Some examples of appearance-based features are shown in Fig. 2.9.

2.3.1.3 Dynamic textures

An emerging new method of appearance-based activity recognition is Dynamic Texture recognition. Chetverikov and Péteri[28] define a Dynamic Texture (DT) as a "spatially repetitive, time-varying visual pattern that forms an image sequence with certain temporal stationarity". Typical examples of DTs are smoke, fire, sea waves and talking faces. There is an increasing body of work concerned with modelling and recognition of these textures. Most of the existing approaches to recognition of DTs are based on the computation of optical flow. An early influential approach to DT recognition by Nelson and Polana[120] uses the normal flow to compute features describing the motion in the DT. Similar to our approach, they base their classification on the use of orientation histograms, flow magnitude, divergence and curl. Of these features, the magnitude, divergence and curl are rotation-, but not scale-invariant. In our approach, due to normalisation of the face prior to extraction of these features, scale invariance is easily achieved. Another related study based on optical flow was performed by Lu et al.[97], who use complete rather than normal flow vectors, and similar to our approach calculate acceleration vectors, but use multi-resolution histograms to achieve scale-invariance. A different approach is used in [128]. Instead of using optical flow, they use system identification techniques to learn generative models. These models are then used to recognise 50 different DTs with encouraging results. Recently, Chetverikov and Péteri[28] published a more extensive overview of DT approaches.

The techniques applied to the DT recognition problem can also be used to tackle the problem of facial expression recognition. Valstar et al.[153] encoded face motion into Motion History Images. This representation shows a sequence of motion energy images superimposed in a single image, detailing recent motion in the face. Zhao and Pietikäinen[173] use volume local binary patterns (LBP), a temporal extension of local binary patterns often used in 2D texture analysis. The face is divided into overlapping blocks and the extracted LBP features in each block are concatenated into a single feature vector. SVMs are used for classification. The approach shows promising results, although only the six prototypic emotions are recognised and no temporal segmentation is performed. To the best of our knowledge, the method proposed here is the only other DT-based method for facial expression analysis proposed so far.

2.3.1.4 Related works

The methodologically most related works to the one presented here is the work of Valstar et al.[153] on MHIs and the work of Zhao and Pietikäinen[173] on Local Binary Patterns. In the MHI implementation of [153], a video was temporally segmented by manually selecting the start and endpoints of an AU activation and a single MHI was created from 6 frames that were distributed equidistantly between these points. Then, a k-NN-based classifier was applied to classify the MHI representation of the input video into one or more AU categories (each of which was represented by a template MHI of the activation of the AU in question). In the work presented here, no manual annotation is required. Also, while their method uses a multi-class classifier, we train separate binary classifiers for each AU and therefore we can detect any combination of AUs. The recent work of Zhao and Pietikäinen[173] is related since it also uses a DT-based approach, in this case for recognising the prototypic emotions. Similar to us, they first normalise the face using the eye position in the first frame, but they ignore any movement during the sequence. In addition, instead of our learned quadtree placement method for feature extraction regions, they use fixed overlapping blocks distributed evenly over the face. Using the quadtree method, we extract class(AU)-specific representations, for example representations at higher granularity in the eye areas for the recognition of AUs that involve eye motions. Finally, they do not attempt a temporal analysis of the emotions.

As far as the target application is concerned, the most similar previous works are those of Pantic and colleagues[108, 151, 152]. They detect up to 23 AUs and their temporal segments (we do so for all 27 AUs), but use a geometric-based approach. As suggested in the literature, appearance-based approaches and geometric-based approaches are often (partially) complementary[144, 107], therefore it seems valuable to develop an appearance-based approach to temporal segment detection of AUs. As we will present in the results section, we achieve results that are similar in terms of overall accuracy, but complementary (some AUs perform much better and some worse), to those presented in [151].

2.3.2 Expression recognition

Recognition is the problem of how to use the extracted feature vectors to classify the input samples into a particular facial expression and/or set of AUs. Often the extracted feature vectors are prohibitively large, making it impractical to train a classifier using all the features. Therefore many authors use either feature selection (e.g. using boosting algorithms[15, 94, 152]), where the most important features are selected and the rest is discarded, or feature reduction (e.g. using principal component analysis[144]), where the feature vector is transformed to a lower-dimensional space whilst retaining the most important information.

Many different classification techniques have been used by authors in this field, including neural networks[142, 143], boosting algorithms[16], linear discriminant analysis[15], Hidden Markov Models[152], Gaussian Mixture Models[161], Sparse Networks of Winnows[153], k-nearest neighbor algorithm[153], a normalised dot product similarity metric[43], dynamic Bayesian networks[29, 172, 147], rule-based expert systems[110, 111, 109], template matching[142] and a technique based on linear programming[57]. The most often used technique in the field appears to be Support Vector Machines[16, 15, 94, 7, 52, 150, 173, 152].

2.3.3 Facial expression analysis for tagging

The use of facial expressions for multimedia tagging is a very new concept and consequently only few works are available. In [64], facial expressions are utilised for implicit feedback to determine the relevance of search results. An excerpt of the result document is shown to the user and the result is then classified as relevant or irrelevant for the query based on the user's facial expressions. A similar methodology is applied in [8].

Jiao and Pantic[66] performed an experiment on implicit tag validation using facial expressions, very similar to our work on tag validation using EEG signals[2]. They use

Emotion	Action Units
From EMFACS[40], FACSAID	
Anger	4, 5, 7, 22, 23, 24
Contempt	12, 14
Disgust	9, 10, 25, 26
Fear	1, 2, 4, 5, 7, 20, 25, 26
Happiness	6, 12
Joy	6, 12
Sadness	1, 4, 15, 17
Surprise	1, 2, 5, 25, 26
From Williams et al.[163]	
Sadness	1, 4, 6, 11, 15, 25, 26
Fear	1, 2, 4, 5, 20, 25, 26, 27
Anger	4, 5, 7, 10, 17, 22, 23, 24, 25, 26
(Acute) pain	4, 6, 7, 9, 10, 12, 20, 25, 26, 27, 43

Table 2.1: Example associations found between AUs and categorical emotions in previous research.

geometric features obtained from a particle filter tracker with a Hidden Markov Model to assess the correctness of tags displayed alongside images from the participants' facial expressions. On average, 54% of the trails were correctly classified.

2.3.4 Connection of Action Units to Emotions

While Action Units are arguably the most effective method for objective annotation of facial expressions. However, AUs do not encode the semantic or affective meaning of the expression, which, for many applications, is the main focus. Thus, methods are needed to map the occurrence of AUs to the presence of higher-level affective states.

For discrete emotions, the EMFACS [40] method (for basic emotions) and the FACSAID[1] (for various affective states) methods provide rules to map AU (co-)occurrences to discrete emotions. FACS has also been used for determination of other complex psychological states such as depression[42] or pain[163]. In [122], AUs are detected and used to detect the high-level states of agreeing, concentrating, disagreeing, interested, thinking. However, no specific associations between occurences of AUs and these states are given. An example of AU-emotion associations is found in Table 2.1.

When it comes to associations between AUs and dimensional models of emotion, such as the valence-arousal model, less research is available[56]. We only found

[1] http://www.face-and-emotion.com/dataface/facsaid/description.jsp

one work that tries to estimate emotions in a dimensional model by first detecting AUs[102]. They first detect the AUs present in the video and subsequently classify the valence value. However, they do not list associations found of particular AUs with valence. The original work proposing the valence-arousal model[127] does however give information on the relation of valence and arousal to discrete emotions. Thus, some information can possibly be gleaned by first mapping the AUs to a discrete emotion using for instance FACSAID, and then finding the approximate location in valence-arousal space from [127] (see also Figure 2.3).

2.4 Conclusions

Research in the use of EEG signal analysis outside the traditional clinical setting is a relatively young field. Nevertheless, several types of features such as event-related potentials and spectral power features have begun to be used for applications such as gaming and multimedia tagging. There has also been some success in the field of emotion assessment. The main focus in this work is on the use of EEG signal analysis for implicit tag validation and for implicit affective tagging. In the field of implicit tag validation, we found no other work using EEG signal analysis. Several works exist based on the P300 event-related potential that allow users to tag odd-one-out images in large sets. However, no work is known that uses video for this purpose, or that does implicit tagging. The tagging is always the main and conscious goal of the work. In contrast, in implicit tagging the goal is to assign tags based on the users' responses automatically, without them making any conscious effort to do so. Whereas emotion assessment is concerned, similar concerns exist. Most approaches require users to actively imagine an emotion and only a few approaches use video as the stimuli data. Given the above, we see a gap in the current literature concerning video-based implicit tagging (or tag validation) using EEG signal analysis.

Compared to the research in EEG outside the clinical setting, the field of facial expression analysis has seen much more interest over the last decades. Early approaches mainly focused on the classification of images. In recent years, the emphasis has shifted to facial expression analysis in video sequences. Most of these works classify a whole sequence as containing a facial expression or manually indicate the frames of interest in the video. The number of works classifying the facial expressions on a frame-by-frame basis is still limited. Another area in which approaches differ is the categorisation of facial expressions. Many works use the six basic emotions, although in recent works the use of Action Units is becoming more popular. The use of continuous models of emotion in facial expression analysis such as the valence-arousal model however remains limited. A third consideration is the use of posed versus spontaneous emotions. It is well known that spontaneous emotions are generally harder to recognize, due to more subtle differences between expressions and the more complex temporal dynamics. Most works however concentrate on the recognition of posed facial expressions. In this work, we present a method for frame-by-frame classification of Action Units and their temporal dynamics. We test this method on both posed and spontaneous expression datasets. Furthermore, we use the same method to also detect valence and arousal in spontaneous video sequences. Finally, to the best of our knowledge, this is the first work attempting fusion of the EEG and facial expression modalities.

Chapter 3

EEG for implicit tag validation

Contents

3.1 Introduction

In this chapter, we describe our research in the use of EEG signals for validation of tags displayed in conjunction with video content. The use of EEG is interesting mainly because it offers the possibility of passive, implicit tagging. This means that tags can be validated by analysing the EEG signals from participants recorded as they watch the video, without active involvement or conscious effort on their part. Many social media websites rely for a large part on user-generated tags to index their content. However, tags given by users may be unreliable or too specific to apply to a large audience[155]. A method for implicitly validating the tags by other users would help in keeping the index clean of spurious tags. In this chapter, we describe a method for validating these tags by observing a user's reaction as the tags are displayed. This relies on the occurrence of an event-related potential (ERP) in the user's EEG after display of an invalid tag. We perform an experiment and demonstrate significant differences in the EEG signals of participants that are shown a valid tag versus the signals of those that are shown an invalid tag.

It has been shown that in cases of two semantically incongruent categories an N400 event-related potential occurs at around 400 ms after the second stimulus is presented (or better: after the mismatch becomes obvious to the viewer). This N400 has been observed even when the stimuli originate from different modalities (e.g. audio and text or images and text)[154, 106, 79, 21]. We aim to show here that the ERP can also be observed when we combine the modalities of video and text by priming the participant by the display of video content, followed by the display of a semantically mismatched tag. To the best of our knowledge the research presented in this chapter is the first work combining the video and text (tag) modalities in N400 ERP analysis. While the analysis of ERPs has been used previously for tagging (for instance in [51]), this is to the best of our knowledge the first work utilizing EEG signal analysis for implicit tag validation.

We describe the experimental setup and the applied preprocessing steps in Section 3.2. Section 3.3 describes the significant correlations found between participants' EEG signals and the validity of tags. In Section 3.4 we explain our methods and results for single-trial classification of the EEG signals. Section 3.5 concludes this chapter. This work has been published in an earlier form in the 2009 Workshop on Affective Brain-Computer Interfaces[2].

3.2 Experimental setup

In the experiment, a participant is shown a video followed by a tag, and from the EEG signals recorded during tag display, we aim to discern whether the tag applies to the video content or not. Our hypothesis is that if the shown tag does not match the video content an N400 ERP will occur.

3.2.1 Data acquisition

We collected a large dataset with 17 participants. Each participant was shown a set of 49 videos depicting 7 event categories. Each video was shown twice, once with a valid tag and once with an invalid tag (selected randomly from one of the other categories). 12 Participants were male, 5 female. Ages ranged from 19 to 31, with a mean age of 25. All but two participants were right-handed and all but three participants viewed the tags in their native language.

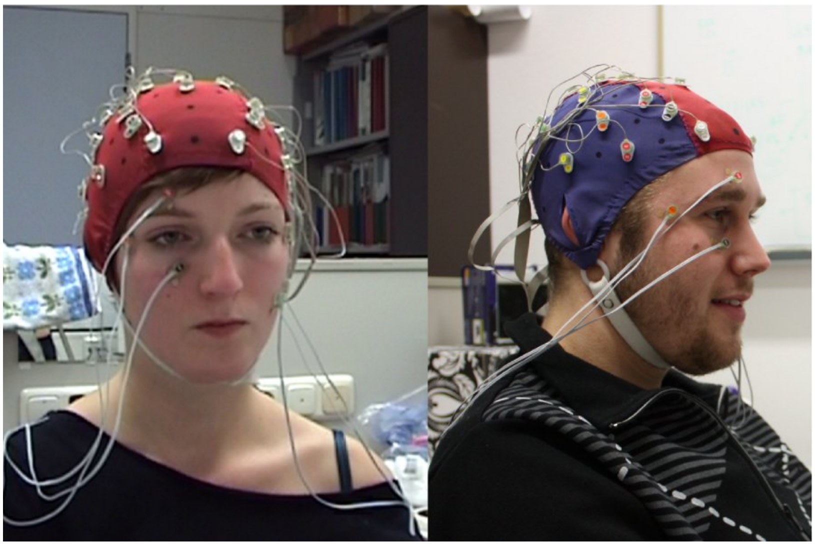


Figure 3.1: Participants performing the experiment.

EEG was recorded using a BioSemi ActiveTwo system[1] on a dedicated recording PC (P4, 3.2 GHz) using the BioSemi Actiview recording software. Stimuli were presented on a dedicated stimulus PC (P4, 3.2GHz) that sent synchronization markers directly to the recording PC. For presentation of the stimuli the Presentation software[2] was used. Participants were seated in a comfortable chair, approximately 70 cm from the presentation monitor (a 20 inch Samsung Syncmaster 203B). In order to minimise eye movements, the video stimuli were all shown with a width of 640 pixels, filling approximately a quarter of the screen. Each participant signed an informed consent form and filled in a short questionnaire. They were then instructed to try to restrict any movement to the periods between trials to minimize movement artefacts in the EEG signal. Participants were told they would be shown videos followed by tags, but were not given any further specific instructions as to the nature of the experiment. 32 active AgCl electrodes were used (placed according to the international 10-20 system) with a sampling rate of 512Hz. Fig. 3.1 shows two participants as they perform the experiment.

Each trial consisted of the following steps:

1. A fixation cross is displayed for 1000 ms (to minimise eye movements).

2. The video is displayed (ranging in duration from 6-10 seconds).

3. A fixation cross is displayed for 500 ms.

[1] BioSemi instrumentation (http://www.biosemi.com)

[2] Presentation by Neurobehavioral systems (http://www.neurobs.com)

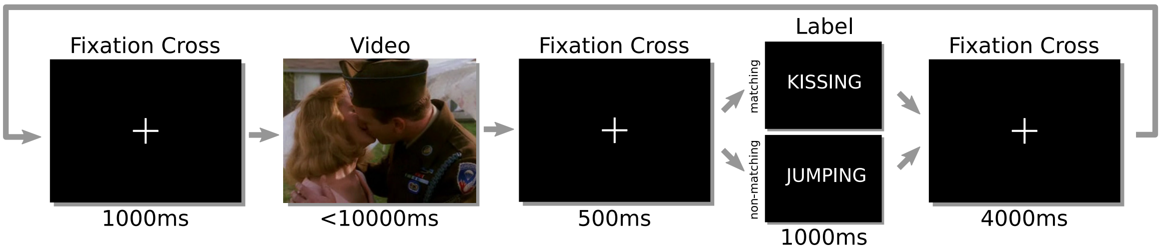


Figure 3.2: Order and timing of the experiment.

4. The tag is displayed for 1000 ms.

5. A fixation cross is displayed for 4000 ms before the start of the next trial.

The stimuli were presented in 3 blocks of 32-33 trials. In between the blocks, participants were given breaks and could move freely, reseat themselves or have a drink of water in order to avoid any muscle straining or fatigue. Fig. 3.2 illustrates the order and timing of the experiment.

Category/Label	Source
Airplane take off	Plane spotter home-videos [3]
People kissing	Hollywood movies dataset[87]
People getting out of cars	Hollywood movies dataset[87]
Mice drinking water	Mouse behaviour dataset[35]
Cats opening doors	Pet home-videos [4]
Jawdrop (posed facial expression)	MMI expression database[112]
Laughing people (spontaneous facial expression)	AMI meeting corpus[24]

Table 3.1: The different video event categories used in the experiment and their sources.

49 Videos from seven different categories were used as stimuli, with 7 videos in each of the 7 categories. Each video has a duration of ten seconds or less and was shown twice, once followed by a valid tag and once followed by an incorrect tag. Table 3.1 gives an overview of the different video categories and their sources. We selected categories with human faces, animals and inanimate objects, following [70], who indicate that these categories can be separated reasonably well by analysing the EEG signals from participants watching the videos.

[3] Plane spotter videos taken from `http://www.flightlevel350.com/`

[4] Pet home-videos taken from `http://www.youtube.com`

The collected dataset, including the stimuli videos, was made publicly available to the research community[5].

3.2.2 Data preprocessing

As a preprocessing step, the baseline of 0.5 seconds before each trial was subtracted from the data and the data was referenced to the common average (CAR). Also, the data was bandpass-filtered between 0.5 and 35Hz to remove DC drifts and suppress the 50Hz power line interference. We use independent component analysis (ICA) to remove eye blinks and other artefacts in the data. Components found to contain mostly noise were manually removed. See Section 3.2.2.1 for examples of artefact components and ERP components. We extracted epochs of each trial for further analysis ranging from 500 ms before tag display to 1000 ms after.

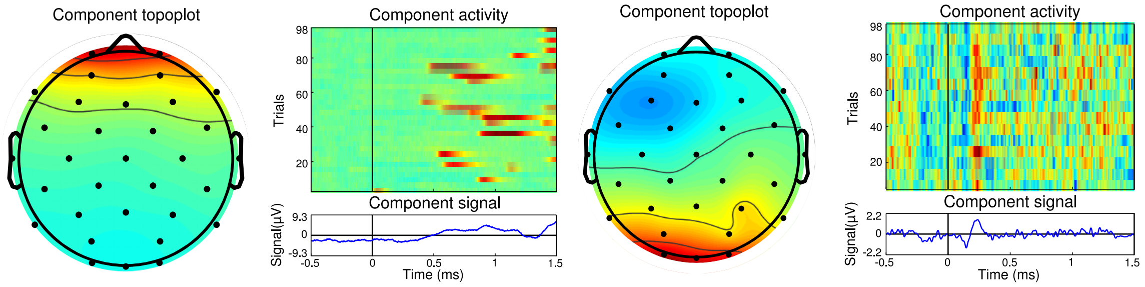
3.2.2.1 Independent component analysis

Independent component Analysis (ICA) is an extension of principal component analysis (PCA) and a form of blind source separation (BSS). In principal component analysis, the objective is to separate the signal into orthogonal components of decreasing variance. In independent component analysis however, we try to find statistically independent components which may not be orthogonal. ICA decomposes the source signal into a linear combination of maximally independent components. The assumption is that measured electrical activity is a linear mixture of underlying sources in the brain. Here, we use the FastICA algorithm, originally developed by Hyvärinen and Oja[65].

The data is first centered and sphered using singular value decomposition. Then, a fixed point iteration scheme is used that iteratively maximises the non-Gaussianity of the components. For details of the algorithm, see [65]. In this work, we mainly use ICA as an artefact removal technique to remove components correlated with artefacts such as blinks, eye motion and main line electrical interference. ICA has been used before in EEG data analysis with good results (e.g. [68]).

An example from the data gathered for this experiment is given in Fig. 3.3. Fig. 3.3(a) is an example of a component that is strongly correlated with eye blinks. This is evident because the activation occurs in isolated periods (blinks) that are not correlated across trials. Also, the component is mostly active in the frontal electrodes. Such components are removed by first applying the ICA transform, removing the components, and then reconstructing the data without them. Fig. 3.3(b) shows an

[5] QMUL-UT dataset: http://www.eecs.qmul.ac.uk/mmv/datasets/qmul_ut/



(a) An independent component that is strongly correlated with blinks. The component activity is concentrated in the frontal area and there is no correlation between trials.

(b) An independent component that is correlated with ERPs in the occipital cortex related with early visual processes. We can primarily see the activation here of the N100 and P200 ERP.

Figure 3.3: Visualisation of two independent components. In each of the sub-figures: On the left is a topoplot of the component activation. In the top right the component activation is shown for the 98 trials of one participant. In the lower right the average component signal is displayed.

example component correlated with the N100 and P200 ERP. The activation is concentrated in the occipital lobe (which is concerned with vision tasks), the component shows a resemblance to a typical ERP curve and there is a strong correlation between trials.

3.3 Statistical analysis

After preprocessing, we performed a paired t-test to determine whether significant differences occur in the recorded EEG signal means between the cases of valid and invalid tags. For this purpose, we only consider the period of 300-500 ms after tag display, during which the strongest N400 response can be expected. Responses for each condition are first averaged over trials and subsequently over the mentioned period. Then a paired t-test between the two conditions over subjects is performed.

Table 3.2 shows the results of the paired t-test. Since a large number (32) hypothesis tests are performed here, we will need to compensate for the multiple comparison problem. That is, as more tests are performed, it becomes more likely for differences to appear significant. To compensate for this, we adjust the α-level considered significant using the Bonferroni method. For an α-level of 0.05 over the whole study, the α-level considered significant per electrode becomes 0.003125. Three electrodes (Pz,T7, and CP2) show a significant difference at this level. It should be noted though, that the Bonferroni correction may be too conservative, especially since the signals

Electrode	***t*-statistic**	***p*-value**	**MSD (μV)**
Pz	5.04205	0.00012†	0.819
T7	-4.59825	0.00030†	-0.758
CP2	4.50180	0.00036†	0.776
CP1	3.69159	0.00198	0.535
FC5	-3.60269	0.00239	-0.640
F7	-3.42631	0.00346	-0.948
C4	3.29918	0.00453	0.364
Cz	2.75625	0.01405	0.480
F3	-2.74236	0.01446	-0.482
P4	2.60730	0.01906	0.478
PO3	2.57479	0.02036	0.616
AF3	-2.53853	0.02191	-0.557
Fp1	-2.39662	0.02911	-0.429
F8	-1.95282	0.06856	-0.455
P7	-1.92596	0.07207	-0.443
FC2	1.74296	0.10052	0.260
Fp2	-1.42725	0.17273	-0.417
CP6	1.39007	0.18355	0.236
Fz	-1.30000	0.21202	-0.261
AF4	-1.29933	0.21224	-0.250
P3	1.14787	0.26790	0.196
C3	0.95051	0.35599	0.094
P8	0.92562	0.36839	0.209
O2	0.83638	0.41526	0.182
FC6	0.81261	0.42837	0.203
O1	-0.79352	0.43908	-0.173
PO4	0.79206	0.43991	0.180
T8	0.37502	0.71257	0.053
F4	0.24187	0.81195	0.055
Oz	0.23799	0.81491	0.056
CP5	-0.14194	0.88889	-0.027
FC1	0.00445	0.99650	0.001

Table 3.2: paired t-test results per electrode, sorted by p-value. MSD stands for Mean Signal Difference. †: Significant differences after Bonferroni correction.

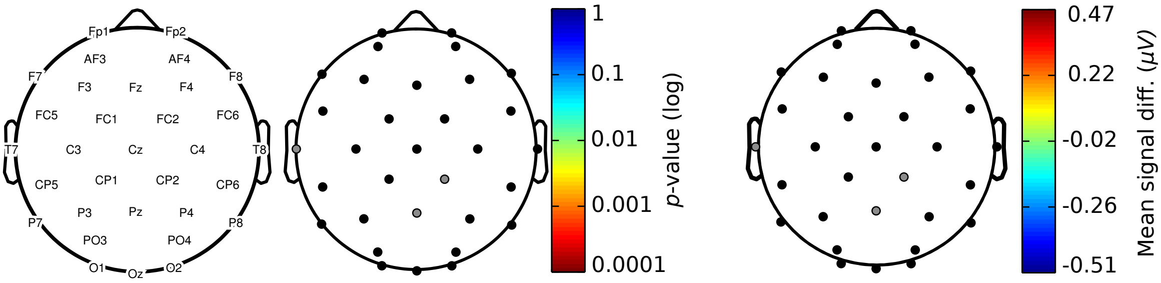


Figure 3.4: Left: Topoplot of Electrode locations, Middle: Topoplot of Significance of difference (p-value), Right: Topoplot of the Grand-average differences between 300 and 500 ms averaged over all 17 participants. Electrodes with significant differences are highlighted in grey.

from different electrodes can be expected to be strongly correlated. 13 electrodes have p-values lower than the uncorrected $\alpha = 0.05$.

Fig. 3.4 shows the location of observed differences in signal values. We can see that the differences are spatially mainly localised in two regions. The main region is located around the occipital and parietal lobe (covering electrodes CP1, Pz, PO3, CP2, C4 and Cz), where a more negative voltage deflection occurs when displaying invalid tags than when displaying valid tags. The occipital lobe is concerned primarily with vision tasks and the parietal lobe is, among other things, concerned with the location of visual attention[72]. The other region showing a difference in signal values is located in the left fronto-temporal regions around electrodes AF3, FC5, T7 and F7. One of the functions of the left fronto-temporal regions is the recognition of words, possibly explaining the activation there. In this case, the observed voltage is less negative for the case of invalid tags than for the case of valid tags.

Fig. 3.5 depicts the grand average waveforms for the 9 electrodes exhibiting the most significant differences between the two cases. The first four plotted electrodes show less negativity for invalid tags than for valid tags. The remaining electrodes show the opposite behaviour and display a higher negativity for the case of invalid tags than for valid tags. Clear examples of the N400 ERP can be observed. The differences are most clear in the 300-500 ms period after tag display.

From these results it is clear that the N400 occurs when participants are shown a combination of stimuli from the modalities of video and text (in the form of a tag). Furthermore, significant differences are present in a considerable number of electrodes between the cases of invalid and valid tags. However, the effect size ($\leq 1\mu V$) is smaller than that found in other studies (e.g. [106, 21]). This can be due to the semantic categories, the stimulus material, or other parameters of the experiment used here.

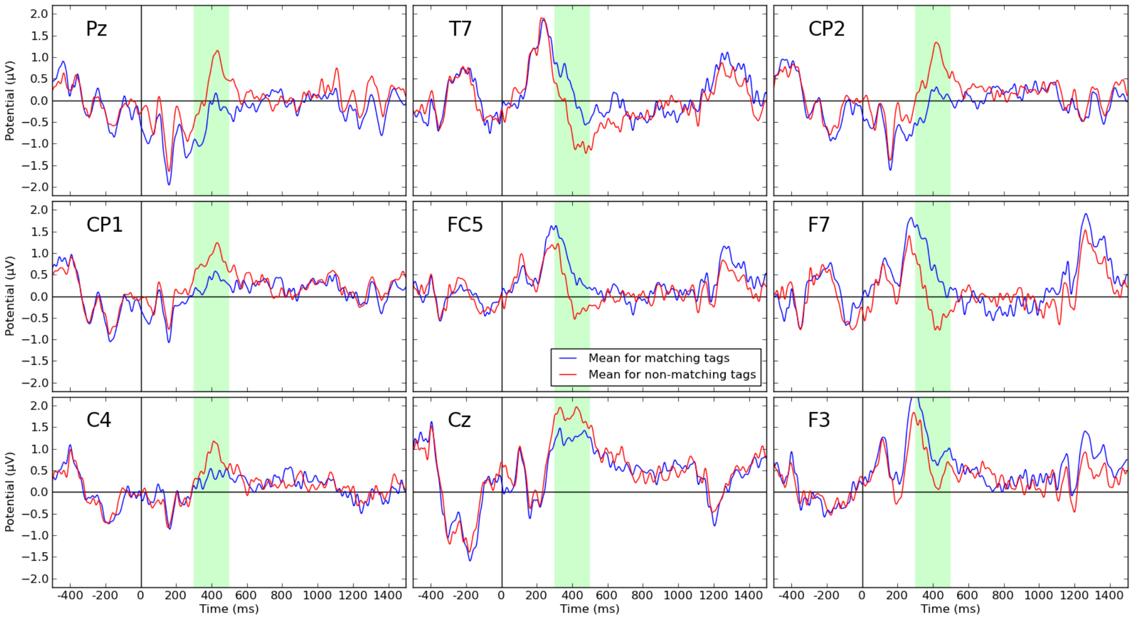


Figure 3.5: Grand average waveforms for the period 500 ms before to 1500 ms after tag presentation for the 9 electrodes with the strongest differences. The signal is averaged over all trials and participants. The lighter red line shows the average signal during presentation of valid tags and the darker blue line shows the average signal for invalid tags. Differences in signal values between the two categories can be observed in each plot around the 400 ms mark. The shaded period corresponds to the window used for t-test analysis. Note that for the y-axis, negative is up.

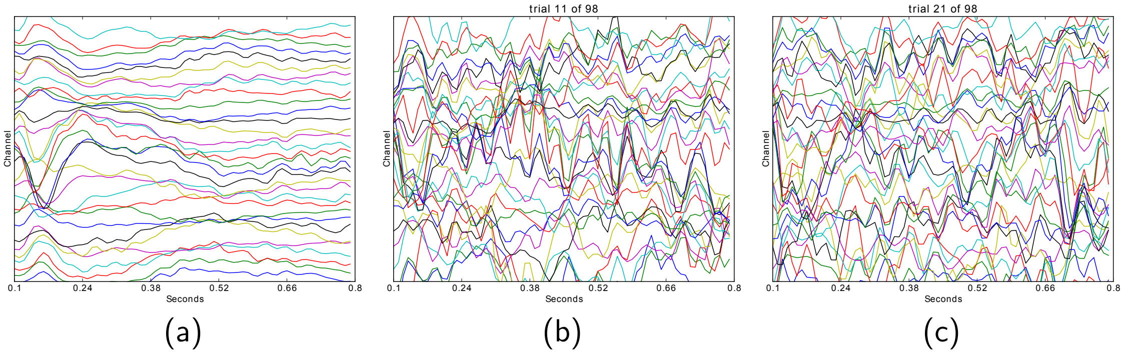


Figure 3.6: Comparison of an averaged EEG signal(left) with two single trial plots (middle,right)

3.4 single-trial analysis

Classifying single trials is a much more difficult challenge than finding significant challenges through statistical analysis. Whereas the t-test operates on averages over many trials, where signal noise is cancelled out to a degree, in single-trial classification we need to deal with this noise and with the large within-participant variability inherent in EEG signals.

To illustrate the difference between averaged signals and single trials, Figure 3.6 shows three EEG signal plot. Figure 3.6(a) shows the average EEG signal for participant 10 in all trials showing invalid tags. One can clearly see the ERP occurring, with a negative peak at approx. 170 ms, followed by a positive peak that slowly tapers off. Figures 3.6(b) and 3.6(c) show single trials of the same participant. Where in Figure 3.6(a) the ERP was clearly visible, in the single trials much processing is needed to detect its presence.

Before feature extraction, the data was down-sampled first to 100Hz using the Fourier transform resampling method (as implemented in the resample function of the scipy.signal software package), using a hamming filter to prevent aliasing of the data. The timeslot of 100ms to 800ms after tag display was used. We used three different methods for the extraction of feature vectors from raw EEG signals, which will now be shortly described in sections 3.4.1, 3.4.2 and 3.4.3.

3.4.1 Windowed Means

Firstly a simple baseline scheme was investigated, with the aim of distinguishing the N400 ERP as shown in Fig. 3.5. The raw data of 100 to 800ms following tag display was divided into sliding windows of 25 samples (0.25 seconds) each with a 50% overlap.

Then, for each window, the mean signal of each electrode was calculated.

We also consider the difference in means between pairs of electrodes for these sliding windows. This is motivated by the occurrence of dipolar fields in the ERP, and possibly these differences seen on the grand-average topoplots will be distinguishable to some extent in the single-trial case. This use of laterization properties has been used before in single-trial ERP analysis(e.g. [18]) The considerations of which electrodes to pair up should ideally be made inside the cross-validation loop, and not decided from a statistical analysis on the whole dataset. We try three different approaches here, using on the means as discussed in the previous paragraph, using the differences between symmetrical pairs on the left and right hemispheres, and using the differences between each pair of electrodes.

3.4.2 Common spatial patterns

Common spatial patterns (CSP) was originally proposed by Koles[81]. It is a technique to decompose the EEG signal into a number of components based on the variance of the signal that takes into account the class labels. In brief, it attempts to extract components for which the variance is maximal for one class and minimal for the other. Then, for a new, unclassified signal, one uses the variance of the components as features to classify the signal as belonging to one of the classes.

Let x denote the $N \cdot T$ matrix that is the EEG signal of one trial, where N is the number of electrodes and T is the number of samples. Further, let X_T represent the EEG signal of all trials in a training set and X_A and X_B the EEG signals for each class, where N_A and N_B denote the number of trials for each condition.

First, we calculate the mean normalized spatial covariance of the signal for each condition:

$$\bar{C}_A = \frac{1}{A} \sum_{a=1}^{N_A} \frac{X_a X_a^T}{trace(X_a X_a^T)} \tag{3.1}$$

$$\bar{C}_B = \frac{1}{B} \sum_{b=1}^{N_B} \frac{X_b X_b^T}{trace(X_b X_b^T)} \tag{3.2}$$

$$\bar{C}_c = \bar{C}_A + \bar{C}_B \tag{3.3}$$

Where $\bar{C}_c$ denotes the composite mean covariance of both conditions. Using eigenvalue decomposition, we find the eigenvectors and eigenvalues of $\bar{C}_c$. The eigenvalues and eigenvectors are sorted in descending order. These are then used to determine a whitening transformation W:

$$C_c = U_c \lambda_c U_c^T \tag{3.4}$$
$$W = \sqrt{\lambda_c^{-1}} U_c^T \tag{3.5}$$

Next, we use W to whiten the condition covariance matrices:

$$S_A = W \bar{C}_A W^T \tag{3.6}$$
$$S_B = W \bar{C}_B W^T \tag{3.7}$$

S_A and S_B now have the interesting property that they share common eigenvectors and their respective eigenvalue pairs sum up to 1. In other words, when we take the eigenvalue decomposition of S_A and S_B, that is,

$$S_A = U_A \lambda_A U_A^T \tag{3.8}$$
$$S_B = U_B \lambda_B U_B^T \tag{3.9}$$

,then:

$$S_{AB} \triangleq S_A = S_B \tag{3.10}$$
$$\lambda_A + \lambda_B = 1 \tag{3.11}$$
$$\tag{3.12}$$

This means that the largest eigenvalue of S_A corresponds to the smallest eigenvalue of S_B and the other way around. We can now define a projection matrix P that will transform a new EEG trial x^0 into the components:

$$P = (U_{AB}^T W)^T \tag{3.13}$$
$$x' = P x^0 \tag{3.14}$$

The variance of the rows of x' can now be used as features to classify x^0 into one of the classes. The first and last components correspond to the highest and lowest eigenvalues and it is thus in these components that we can expect the largest differences between the two conditions. It is quite common to only use a few components for classification.

Since variance can be seen as a measure for the power of a signal, CSP is most suitable for situations where the main differences in the signal lie in frequency power. It is therefore not so well suited for slow, non-oscillatory signal differences as seen in event-related potentials.

3.4.3 Discriminative spatial patterns

Since the CSP method is not well suited to slow, non-oscillatory signals, Liao et al.[92] proposed an alternative spatial filtering algorithm known as discriminative spatial patterns (DSP). Like CSP, it aims to find a transform from the original high-dimensional feature space to one where the separation between experimental conditions (or classes) is maximal. Rather than using covariance matrices and trying to maximize the variance for one class while minimizing it for the other class, it instead tries to maximize the between-class distance and minimize the within-class distance based on raw signal amplitudes.

Let x again denote the $N \cdot T$ matrix that is the EEG signal of one trial, where N is the number of electrodes and T is the number of samples. Further, let X_T represent the EEG signal of all trials in a training set and X_A and X_B the EEG signals for each class, where N_T,N_A and N_B denote the total number of trials and the number of trials for each class, respectively. The average EEG signal $\bar{X}_c$ for each class $c \in (A, B)$ is

$$\bar{X}_c = \frac{1}{N_c} \sum_{i=1}^{N_c} X_c(i) \tag{3.15}$$

Then, the within-class scatter matrix S_W can be found by

$$S_c = \sum_{i=1}^{N_c} (X_c(i) - \bar{X}_c)(X_c(i) - \bar{X}_c)^T \tag{3.16}$$

$$S_W = \sum_{c \in (A,B)} S_c \tag{3.17}$$

The between-class scatter matrix S_B is found by

$$\bar{X}_T = \frac{1}{N_T}\sum_{i=1}^{N_T} X_T(i) \tag{3.18}$$

$$S_B = \sum_{c\in(A,B)} N_c(\bar{X}_c - \bar{X}_T)(\bar{X}_c - \bar{X}_T)^T \tag{3.19}$$

We aim to find a projection matrix P that projects S_A and S_B into a new space:

$$\tilde{S}_B = P^T S_B P \tag{3.20}$$

$$\tilde{S}_W = P^T S_W P \tag{3.21}$$

The goal is to determine P such that $\tilde{S}_B$ is maximized and $\tilde{S}_W$ is minimized for optimal separation of the classes. Fisher's linear discriminant criterion is given by:

$$J(P) = \frac{|\tilde{S}_B|}{|\tilde{S}_W|} = \frac{|P^T S_B P|}{|P^T S_W P|} \tag{3.22}$$

Let λ_d and p_d be the eigenvalues and eigenvectors of P. Then, $J(P)$ can be maximized by solving the generalized eigenvalue problem

$$S_B p_d = \lambda_d S_W p_d \tag{3.23}$$

Having derived P, the final mapping of a new EEG trial x^0 is done by

$$x' = P^T x^0 - P^T \bar{X}_T \tag{3.24}$$

3.4.4 Experimental Results

Leave-one-trial-out cross-validation (i.e. 98-fold cross-validation) was used in testing in order to minimize overfitting. We investigated the use of each of the feature extraction approaches discussed earlier and compare their results. In each case, classification of the signals was performed using a linear support vector machine (SVM) classifier. We will describe the results for each method shortly.

As mentioned above, for the windowed means method we tried using the means only, the means and difference in means between left-right symmetrical pairs of elec-

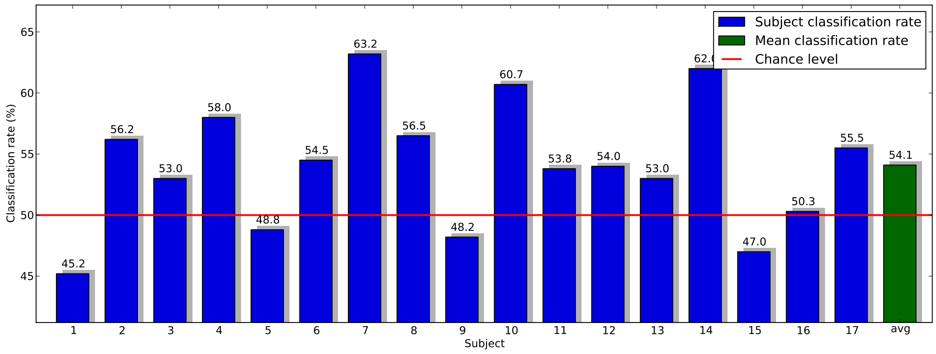


Figure 3.7: single-trial classification results per participant for the windowed means algorithm

Means only	**Means + diff. sym. pairs**	**Means + diff. all pairs**
53.6%	54.1%	54.1%

Table 3.3: Average results for the three tests run with the windowed means algorithm

trodes and differences between all electrodes. The average results over subjects for the three tests are given in Table 3.4.4. The use of differences between electrodes improves the classifier accuracy slightly, although no difference is observed between the use of all electrodes and only the symmetrical pairs.

Per subject results for the windowed means method with differences between all pairs of electrodes are given in Fig. 3.7. An independent one-sample t-test confirms that the average result of 54.1% is significantly higher than the 50% chance level ($p <= 0.005$). At 1% significance level ($p <= 0.01$), 12 out of the 17 participants show results significantly higher than chance level.

As discussed in Section 3.4.2, the CSP method is mainly useful in oscillatory signals where we expect differences between the power of certain frequencies. However, in ERP analysis the main difference is non-oscillatory. The results for this algorithm are shown in Fig. 3.8. Here, we used the 3 highest and the 3 lowest components.

As expected, the results are not very good, with an average classification rate of only 49.5%. Only for 6 out of the 17 participants the results are significantly above chance level ($p <= 0.01$).

We also used the Discriminative Spatial Patterns (DSP) method (See Section 3.4.3) to project the features onto maximally separable components. Only the first (most discriminative) DSP component was retained (the classification rate did not improve

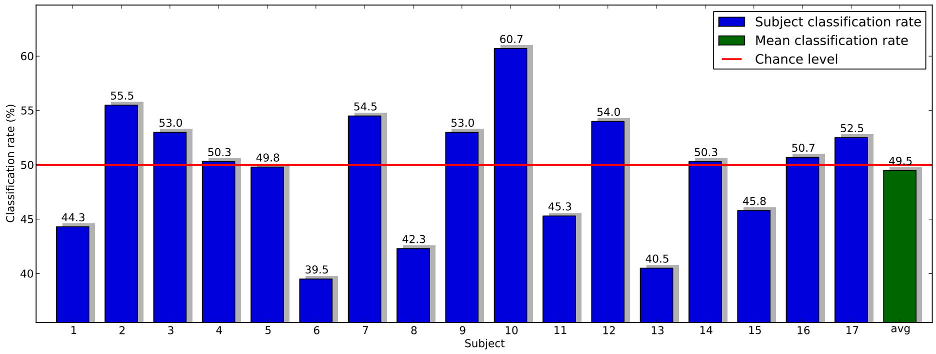


Figure 3.8: single-trial classification results per participant for the CSP algorithm

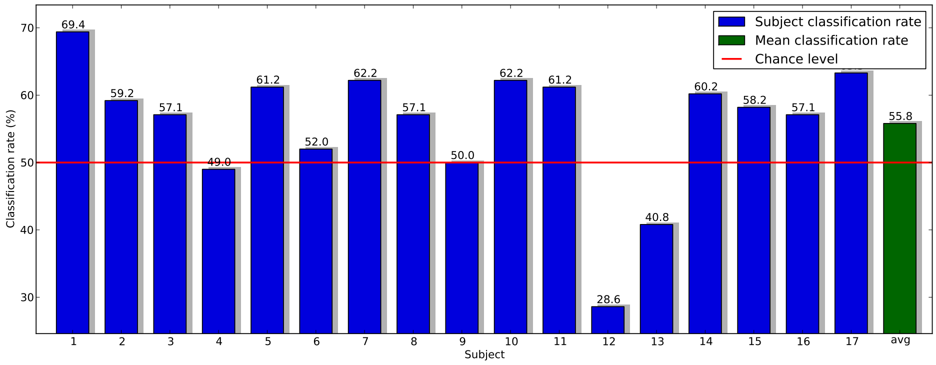


Figure 3.9: single-trial classification results per participant for the DSP algorithm

by using more components). The results show a mean classification rate of 55.8%, significantly higher than chance level ($p <= 0.03$). For 12 out of the 17 individual participants, the results are significantly higher than chance level. Full results per participant are given in Fig. 3.9.

Table 3.4.4 shows the results for all three methods side by side. DSP gives the best result of the three, followed by the windowed means method. It's clear that CSP is not meant for this type of non-oscillatory problems as the result is not significantly different from chance level. Obtaining significant results for DSP and windowed means indicates that the use of EEG signal may form a useful tool for implicit tag validation. While the single-trial classification rate is still low, it may be possible to obtain reliable tag validation results when aggregating the results of many users.

WM	CSP	DSP
54.1%‡	49.5%	55.8%†

Table 3.4: Average classification rates for the three investigated methods (WM = windowed means). Daggers indicate whether the classification rate distribution over participants is significantly higher than 50% according to an independent one-sample t-test ($^{\ddagger}= p < .01$, $^{\dagger}= p < .05$)

3.5 Conclusions

In this Chapter, we proposed a method for the implicit validation of tags displayed in conjunction with video clips. We described an experiment where data was collected for 17 participants and each participant was shown 98 videos, 49 followed by with valid tags and 49 followed by invalid tags. The collected dataset was made publicly available.

Independent Component Analysis was used to remove noise (including eye blink artefacts) from the data. A paired t-test showed significant differences in the EEG signal between the two cases of valid and invalid tags. single-trial analysis was performed using three different algorithms. For the Discriminative Spatial Patterns method a classification rate of 55.8% was attained, yielding significant results for 12 out of 17 participants.

The results indicate that this technique may be used as a mechanism to validate spurious tags. In similar P300 experiments usually the EEG signal of several trials is averaged to increase the signal-to-noise ratio and increase the accuracy of ERP detection. Similarly, here validity of tags could be determined by averaging out the validation result of many users.

However, in order to achieve a practical tag validation system several parameters will have to be studied and optimized. Questions that need to be answered include: how long after a stimulus does an invalid tags still elicit the ERP? What types of categories elicit the most robust mismatches? Does a subliminal presentation, not consciously perceived by the viewer, also elicit N400 responses? Does the manner of tag presentation (alongside, after, in between the video) influence the recognition rate? These questions are left as future work.

Chapter 4

EEG for implicit affective tagging

Contents

4.1 Introduction

In this chapter, we describe our research in the use of EEG signals for automated affective, implicit tagging of multimedia data. We focus here on affective tags defined in terms of valence and arousal judgements. Affective tags form a valuable, high-level description of the data that is very hard to obtain fully automatically due to the semantic gap. Examples of the use of affective tags in retrieval include recommendation of music to match a user's mood, or searching for films by the emotions they are likely to elicit. Previous work has given indications that it may be possible to observe information about affective states from EEG signals[134, 129].

In this work, several methods are investigated and a series of experiments described on classifying participants' emotional states (in terms of valence and arousal) as they view music videos. To the best of our knowledge, this is the first work to perform affect estimation from EEG signals with music videos as the stimuli data.

This work is carried out as part of the EU project PetaMedia[1]. The goal of our part of the project was the realization of a real-time affective music recommendation system(as described in Appendix A). The main focus of PetaMedia was on the investigation of tagging, social networks, and multimedia content analysis. Within our part of the project, we focussed on affective tags, recommendation (using social network data), and multimedia content analysis of the experiment stimuli. Given the requirements of multimedia stimuli combined with recommendation systems, the use of music videos seemed the most natural fit. Advantages of using music videos are:

- One of the main motivations in music composition is the elicitation of a wide range of emotions. For a lesser extent, this can also be said of the accompanying videos. Ergo, one can expect music(videos) to be powerful affect elicitation stimuli.
- Possibly, the combination of video and audio modalities may enhance the affect elicitation, compared to the use of music only.
- Music videos are relatively short, in the range of several minutes. This makes them well suited for use in time-limited experiments.
- Music videos are good examples of 'real-world' stimuli (e.g. highly variable and found in everyday life). A system capable of classifying affect based on such stimuli is more likely to generalize well compared to 'laboratory' stimuli (e.g. short segments specifically selected to elicit a single emotion).

[1] http://www.petamedia.eu

Disadvantages of music videos as stimuli include:

- A single music video may elicit a range of different emotions sequentially. This makes the ground truth labelling ambiguous and thus classification can be significantly harder compared to single-emotion stimuli.
- It proved relatively hard to find videos that elicit a negative valence and positive arousal. This may be easier to do when using for instance (scary) movie clips.
- The combination of video and audio modalities make it impossible to distinguish which modality is the cause of the elicited affect.
- Affect elicitation of music is highly subjective, probably more so than compared to using movie clips. This makes it hard to select a set of stimuli materials that will elicit a wide range of emotions for all participants.

Here, we describe first a pilot study in section 4.2, where the EEG signals of 6 participants were recorded and classified in terms of valence and arousal. The goal of this study was mainly to test the feasibility of such an experiment and to see whether any significant classification results can be achieved. Encouraged by the results of the pilot study, we performed an expanded version of this experiment with 32 participants (described in section 4.3), where a novel stimuli selection method is proposed and both single-trial classification as well as regression results are presented. Table 4.1 gives a comparison of the two experiments in terms of stimuli selection and experimental conditions.

The dataset collected during this second experiment is made publicly available and described in detail in Appendix B. Based upon the methods described in this chapter, a first of its kind real-time music recommendation system, which utilises the affective state estimates, was built as a proof of concept. This system is described in Appendix A.

A significant part of the work in this Chapter was carried out within a group of researchers in the PetaMedia project. Analysis on peripheral physiological signals, MCA signals and emotional highlight detection was carried out by Mohammad Soleymani from the University of Geneva. Some analysis on participant ratings and EEG correlates was performed by Christian Mühl from The University of Twente. Jong-Seok Lee from the École polytechnique fédérale de Lausanne (EPFL) has contributed a section on multimodal fusion. All Sections containing work from these colleagues have been clearly labelled in the text.

Parts of the work presented in this chapter have been published in the proceedings of the Brain Informatics conference 2010[5] and the EmoSPACE Workshop

Stimuli selection	**Pilot study**	**DEAP study**
stimuli material	music videos	music videos
initial selection	manual	automatic/manual
number of videos	70	120
video fragment	full video	60s emotional highlight
num. participants	55	not recorded
minimum num. ratings per video	7	14
rating dimensions	arousal valence liking	arousal valence dominance
Experiment	**Pilot study**	**DEAP study**
number of videos	20	40
video fragment	120s manual selection	60s emotional highlight
num. participants	6	32
rating dimensions	arousal valence liking	arousal valence dominance familiarity
modalities classified	EEG, physiological	EEG, physiological, MCA, fusion
modalities recorded	EEG, EOG, physiological	EEG, EOG, physiological, face video
dataset published	no	yes

Table 4.1: Comparison of stimuli selection and experimental conditions for the two experiments discussed in this chapter.

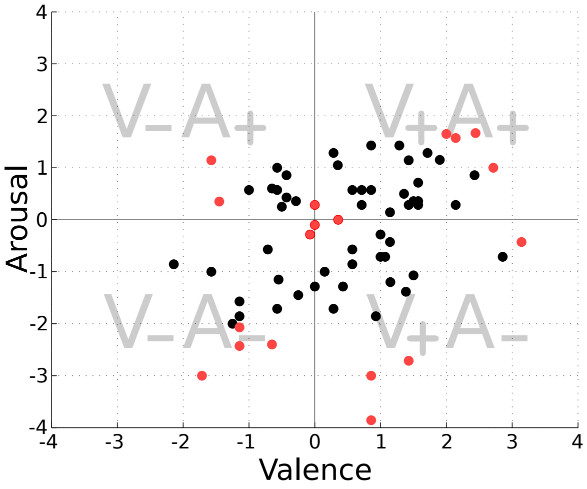


Figure 4.1: Ratings for 70 videos from the online questionnaire. Selected videos are shown in red.

2011[6]. In addition, an article was published in the IEEE Transactions on Affective Computing[1].

4.2 Pilot study

In this study, EEG signals of 6 participants were recorded as they each watched 20 music videos. In between each video, participants were asked to give ratings of general liking, arousal and valence. Then, the EEG signals were classified into low/high arousal, valence and liking.

4.2.1 Stimuli selection

In order to select 20 videos that would elicit the most differing emotions, an online questionnaire was set up. First, a total of 70 videos were manually selected with the aim of fully covering the valence-arousal space. These were partitioned into 5 sets of 14 videos. Participants in the online questionnaire were asked to rate one set of 14 videos in terms of valence and arousal using SAM manikins. Participants were instructed to rate their felt (not perceived) emotion as a result of the music video (as a whole). 55 persons participated in the online study in total and each video was rated by at least 7 individuals.

The results are shown in Figure 4.1. 20 videos (the four most extreme from each quadrant in the arousal-valence space and the four most neutral videos) were then selected to be used in the experiment. Each video was cut down to 2 minutes in length. The original length of the music videos varied between 3:03 and 8:23.

Artist	**Title**	**Quad**
Fool's Garden	Lemon tree	V_+A_+
Aqua	Barbie girl	
Joe Satriani	Summer song (live 2006)	
Katrina and the Waves	Walking on sunshine	
Afro Man	Because I got high	V_+A_-
Jeff Buckley	Hallelujah	
J. Birkin & S. Gainsbourg	Je t'aime moi non plus	
Wilhelm Kempff	Beethoven's moonlight sonata	
Simonopetra monastery	Agni parthene	V_-A_-
Fragma	Toca me	
Baz Luhrmann	Everybody's free to wear sunscreen	
Johnny Cash	Hurt	
Arch Enemy	My apocalypse	V_-A_+
Scooter	Hyper hyper	
Madonna	Like a prayer	neutral
Nada Surf	Popular	
The Cranberries	Zombie	
Within temptation	Ice queen	

Table 4.2: The music videos used in the experiment and their placement in the arousal-valence space according to the online questionnaire. Note: the two songs in the V_-A_+-quadrant were used twice.

The videos were shortened mainly to keep the experiment duration below 1 hour to minimize participant fatigue and ensure the quality of the recorded EEG signals (as the effectivity of the electrode gel decreases over time). For one quadrant negative valence, positive arousal) only 2 videos had somewhat extreme ratings, these were both used twice (using different parts of the video each time). The selected videos are given in Table 4.2.

4.2.2 Data acquisition

The EEG of 6 participants was recorded. Each watched the videos in a random order. Before each video, a fixation cross was displayed for five seconds. At the end of each trial, participants performed self-assessment of their levels of arousal, valence, and liking (how high did they rate the video?). SAM manikins were used to visualize the scales. For the liking scale, thumbs down/thumbs up symbols were used. The manikins were displayed in the middle of the screen with the numbers 1-9 printed

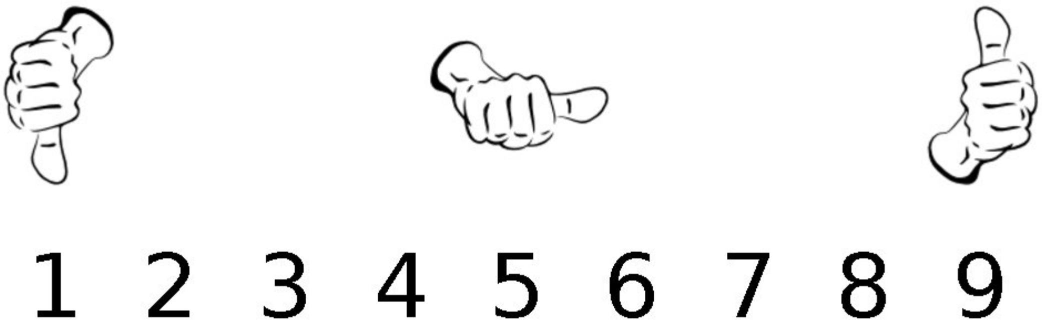


Figure 4.2: self-assessment screen for liking.

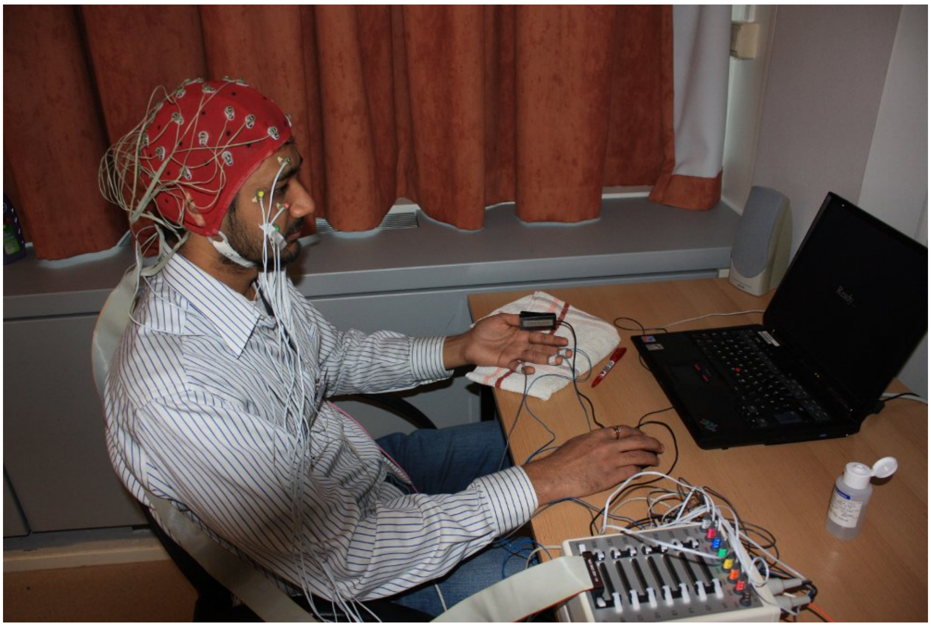


Figure 4.3: A participant shortly before the experiment

below (see Fig. 4.2). The participants could then position the mouse cursor at or in between the numbers and click to give their rating. Participants were instructed to rate their felt (not perceived) emotion as a result of the music video (as a whole).

EEG and peripheral physiological signals were recorded using a Biosemi ActiveTwo system[2] on a dedicated recording laptop (Pentium M, 1.8 GHz) using the BioSemi Actiview recording software. Stimuli were presented on a dedicated stimulus laptop (P4, 3.2GHz) that sent synchronization markers directly to the recording PC. For presentation of the stimuli the Presentation software[3] was used. In order to minimise eye movements, the video stimuli were all shown with a width of 640 pixels, filling

[2] BioSemi instrumentation (http://www.biosemi.com)
[3] Presentation by Neurobehavioral systems (http://www.neurobs.com)

Table 4.3: The correlations between the rating scales of valence, arousal, like/dislike and the order of the presentation of stimuli. Significant correlations are indicated by stars.

	Valence	**Arousal**	**Like/Dislike**	**Order**
Valence	1	0.46*	0.66*	-0.24
Arousal	-	1	0.56*	-0.17
Like/Dislike	-	-	1	-0.18
Order	-	-	-	1

approximately a quarter of the screen. Each participant signed an informed consent form. 32 active AgCl electrodes were used (placed according to the international 10-20 system) with a sampling rate of 512 Hz. At the same time, 13 peripheral physiological signals (which will be introduced in section 4.2.6) were also recorded. Fig. 4.3 shows a participant shortly before the start of the experiment.

4.2.3 Analysis of Subjective Ratings

The analysis in this Section was provided by Christian Mühl from the University of Twente.

To validate the affect induction approach and identify possible threats to reliability (e.g. due to extreme habituation or fatigue), we computed the (Spearman) correlations between the rating scales and the stimulus order.

The correlation analysis revealed a medium correlation between the ratings on the valence, arousal, and like/dislike scales (Table 4.3). That could be due to the fact that people liked positive emotions evoking and arousing clips more. Despite the correlations between valence, arousal, and like/dislike, the results suggest that subjects did differentiate between these concepts.

Furthermore, no significant correlation between stimulus order and the ratings was observed. This indicates that any effects of habituation and fatigue were kept to an acceptable minimum.

4.2.4 Correlations Between EEG Frequencies and Ratings

The analysis in this Section was provided by Christian Mühl from the University of Twente.

For the investigation of the correlates of the subjective ratings with the EEG signals, the EEG data was referenced to the common average, re-sampled to 256 Hz, and high-pass filtered with a 0.5 Hz cutoff-frequency using EEGlab[4]. Eye movement and

[4] http://sccn.ucsd.edu/eeglab/

blinking artefacts were removed with a blind source separation technique from the AAR toolbox[5] for EEGlab. Then the signals from the last 30 seconds of each trial (video) were extracted for further analysis. To correct for stimulus-unrelated variations in power over time the EEG signal from the five seconds before each video was used as a baseline.

The frequency power of trials and baselines between 2 and 40Hz was extracted with Welch's method with windows of 256 samples. The baseline power was then subtracted from the trial power, yielding the change of power relative to the prestimulus period. These changes of power were then Spearman correlated with the valence ratings. This was done for each subject separately and the six p-values per frequency and electrode were then combined to one p-value via Fisher's method [95].

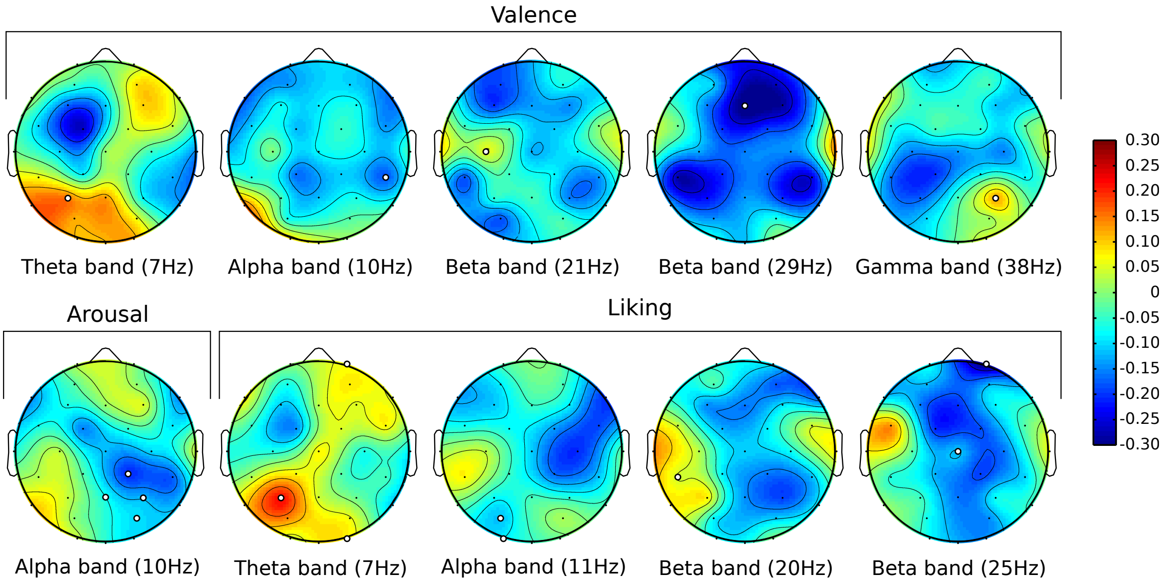


Figure 4.4: The plots show the mean correlation coefficients over all 6 subjects for specific narrow frequency bands. Electrodes showing highly significant ($p < 0.01$) differences are highlighted.

The results of the correlation analysis between participant ratings and EEG frequency power suggest that brain activity from different regions of the scalp can be related to the subjective emotional states of the participants along the axes of arousal and valence, and to their preference for the clips (Fig. 4.4). The large number of tests computed may lead to an increase in false positives. To attenuate this risk, only highly significant ($p < 0.01$) correlations are discussed.

For valence a strong positive correlation with left parietal-occipital power in the theta band, and a negative correlation with right posterior alpha power is observed.

[5] http://www.cs.tut.fi/g̃omezher/projects/eeg/aar.htm

This pattern of increasing low frequency band and decreasing alpha band power can be understood in the context of emotion regulation and increased sensory processing [78]. Furthermore, a left central increase and a right frontal decrease in high beta band power is visible with higher valence. Especially, the frontal response might indicate a relative deactivation of cortical regions related to negative mental states [34]. Additionally, a positive correlation with right posterior gamma is observed, possibly hinting again to a role of right posterior cortices in emotion-related sensory processes.

For states of higher arousal, a robust decrease of right posterior alpha power can be observed. This is consistent with the role of (posterior) alpha in sensory processes, and the role of the right hemisphere in affective processing [34].

Like/dislike shows a similar positive correlation in the theta range and negative correlation in the alpha range as observed for valence. This is presumably due to the correlations seen between the valence and like/dislike ratings. Interestingly, a decrease of beta power with higher liking is observed over the central cortical region, known to be involved in imaginary and real (foot) movement [140].

4.2.5 EEG signal classification

4.2.5.1 Data processing

The data was referenced to the common average (CAR). Also, the data was bandpass-filtered between 0.5 and 35Hz to remove DC drifts and suppress the 50Hz power line interference. We down-sampled the data to 100Hz in order to speed up the classification processes. To improve statistical accuracy of the single-trial classification, each of the 6 participant's trials was segmented into 200 12 second trials. Next, Two different methods were used to extract features from these segments: Common spatial patterns (CSP) and Power spectral density (PSD). CSP was discussed earlier in section 3.4.2. PSD will now be described shortly.

4.2.5.2 Power spectral density

Power spectral density analyses concern the spectral domain and look at the rhythmic activity of brainwaves. A standard analysis of band power includes taking the Fourier transform of the signal and calculating the power in each frequency band. One can also include the coherence between electrode sites by taking the difference in band power between every pair of electrodes. It is also possible to use the wavelet transform rather than the Fourier transform to also include the temporal differences (i.e. changes in band power during the experiment).

The power in each of the frequency bands was calculated using the Fourier transform of the signal. Often, the delta theta, alpha, beta and gamma wave bands are used, but here we tried different fixed bandwidths (from 1 to 10Hz) with 50% band overlap. We also included the difference in band power between every pair of electrodes as features.

4.2.5.3 Results

Three different targets were considered for classification: the liking rating, the arousal rating and the valence rating. The problem is posed as a two-class classification problem, where the given ratings were thresholded (at rating 5) into two classes per target. For the arousal and liking targets, we had to exclude participant 1, as this participant gave 17/20 videos a high arousal rating and 19/20 videos a high liking rating. As a result, we did not have enough samples to train the classifier for low arousal and low liking for this participant. All other participants rated the videos in a more balanced manner.

We investigated two different feature extraction methods (CSP and PSD). For each method and each participant, a linear Support Vector Machine was used to classify the data. Testing was performed using leave-one-trial-out cross-validation. The results of each algorithm and each classification target are described below.

As mentioned before, in the EEG single-trial classification, we compared two different feature extraction methods for feature extraction from the EEG signals, PSD and CSP. With the PSD method, we tested several options for the width of the frequency bands (1,2,3,4,5 and 10Hz). Only the results of the best scoring bandwidth are reported. The results of each algorithm and each classification target are given in Table 4.4 and discussed below.

Arousal For CSP two components were used and for PSD 10Hz frequency bands. Overall, CSP outperforms the PSD method (55.7% vs. 51.9%). For participant 2, the result is below chance level for both methods. When excluding this participant, CSP has a mean classification rate of 58.5% vs. 60.0% for PSD.

Valence The performance for valence prediction is better than for arousal prediction. For CSP using two components gave the best result. The best result for the PSD method was obtained using 3Hz frequency bands (with 50% overlap). Overall, both algorithms lead to the same classification accuracy (58.8%). Participant 2 scores below random level for both methods. When excluding this participant, CSP scores 62.8% vs. 61.6% for the PSD method.

Target	Method	P1	P2	P3	P4	P5	P6	Avg.
Arousal	PSD (10Hz)	—	19.5	67.0	63.5	56.5	53.5	**51.9**
	CSP (2)	—	44.5	55.0	65.0	59.5	54.5	**55.7**
Valence	PSD (3Hz)	59.0	45.0	63.0	51.5	58.5	76.0	**58.8**‡
	CSP (2)	60.0	38.5	54.0	60.0	65.0	75.0	**58.8**†
Like/Dislike	PSD (4Hz)	—	54.0	50.5	57.5	32.5	52.5	**49.4**
	CSP (4)	—	53.0	54.0	63.5	17.0	56.5	**48.8**

Table 4.4: Single-trial two-class classification rates for the valence, arousal and like/dislike targets. The number following the methods denote bandwidth for PSD and number of components for CSP. Daggers indicate whether the average classification rate over participants is significantly higher than 50% according to an independent one-sample t-test ($^\ddagger = p < .01$, $^\dagger = p < .05$).

Like/dislike For CSP four components were used and for PSD 4Hz frequency bands. The PSD method obtains the highest classification accuracy, though the difference is minimal (49.4% vs. 48.8%). Participant 5 had a very low accuracy for both methods. When excluding this participant, CSP outperforms the PSD method (56.8% vs. 53.6%).

In general, it appears from the results that the CSP method slightly outperforms the PSD method. Furthermore, it seems that the classification rates are strongly subject-dependent. For valence, results for both methods are significantly better than chance level. For arousal this is also the case when excluding participant 2.

4.2.6 Peripheral physiological signals classification

The analysis in this Section was provided by Mohammad Soleymani from the University of Geneva.

4.2.6.1 Features

The following peripheral nervous system signals were recorded: galvanic skin response (GSR), respiration amplitude, skin temperature, electrocardiogram, blood volume by plethysmograph, electromyograms of Zygomaticus and Trapezius muscles, and electrooculogram (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. This resistance decreases due to an increase of perspiration, which usually occurs when one is experimenting emotions such as stress or surprise. Moreover, Lang et al. discovered that the mean value of the GSR is related to the level of arousal [86].

A plethysmograph measures blood volume in the participant's thumb. This measurement can also be used to compute heart rate (HR) by identification of local maxima (i.e. heart beats), inter-beat periods, and heart rate variability (HRV). Blood pressure and heart rate variability correlate with emotions, since stress can increase blood pressure. Pleasantness of stimuli can increase peak heart rate response [86]. In addition to the HR and HRV features, spectral features derived from HRV were shown to be a useful feature in emotion assessment [101].

Skin temperature was also recorded since it changes in different emotional states. The respiration amplitude was measured by tying a respiration belt around the abdomen of the participant. Slow respiration is linked to relaxation while irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear.

Regarding the EMG signals, the Trapezius muscles (neck) activity was recorded to investigate the possible head movements during music listening. 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. Eye-blinking affects the EOG signal and results in easily detectable peaks in that signal.

In total 53 features were extracted from peripheral physiological responses based on the proposed features in the literature [25, 160]. A summary of the features is given below. For all features, the baseline of 5 seconds before each trial was first subtracted from the data.

GSR Mean and standard deviation of skin resistance, mean of derivative, mean of absolute of derivative, mean of derivative for negative values only (mean decrease rate during decay time), proportion of negative samples in the derivative vs. all samples, spectral power in the bands (0-0.1Hz, 0.1-0.2Hz, 0.2-0.3Hz, 0.3-0.4Hz)

Blood volume pressure Mean and standard deviation of HR and its derivative, HRV, mean and standard deviation of inter beat intervals, energy ratio between the frequency bands 0.04-0.15Hz and 0.15-0.5Hz, spectral power in the bands (0.1-0.2Hz, 0.2-0.3Hz, 0.3-0.4Hz), low (0.01-0.08Hz), medium (0.08-0.15Hz) and high (0.15-0.5Hz) frequency components of HRV power spectrum.

Respiration Mean respiration signal, mean of derivative (variation of the respiration signal), standard deviation, range or greatest breath, breathing rate, spectral

Table 4.5: Classification rates using an SVM classifier and FCBF feature selection

Target	P1	P2	P3	P4	P5	P6	Avg.
Valence	40.5	37.0	78.5	40.5	65.5	63.0	54.2
Arousal	—	44.5	85.5	55.0	49.0	60.5	58.9
Like/Dislike	—	73.0	69.0	55.5	32.0	60.0	57.9

power in the bands (0-0.1Hz, 0.1-0.2Hz, 0.2-0.3Hz, 0.3-0.4Hz)

Skin Temperature Range, mean, standard deviation, mean of its derivative, spectral power in the bands (0-0.1Hz, 0.1-0.2Hz, 0.2-0.3Hz, 0.3-0.4Hz)

EMG and EOG Eye blinking rate, energy, mean and variance of the signal.

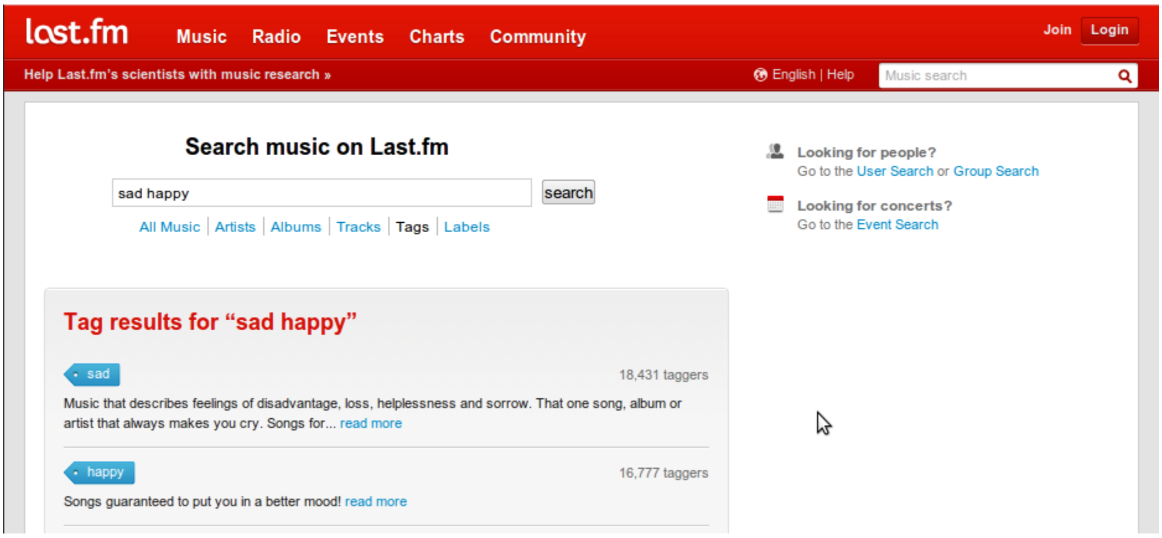
Normalization was applied on each feature separately by subtracting the minimum and dividing by the difference between the maximum and the minimum value of the features. The normalization parameters, maximum and minimum values, were obtained from the training set.

4.2.6.2 Results

As mentioned above, 53 features were extracted from each physiological signal sample. The classification scheme remains the same as for EEG-based classification. The fast correlation based filter (FCBF) feature selection method was used to select the most discriminating features at each iteration of cross-validation[170].

The classification rates for valence, arousal and like/dislike are given in Table 4.5. On average, valence results using peripheral physiological signals are worse than like/dislike and arousal. Arousal relates most to peripheral nervous system activities; therefore, the best classification results were obtained for arousal classification. All the results of participant 3 are shown to be amongst the best obtained results. This may be due to a better self-assessment for this participant.

However, the low number of participants in the study as well as the low number of trials make it hard to draw any definite conclusions. Nevertheless, we considered the results of this small scale test encouraging enough to warrant further investigation, and a more extensive experiment is described in the next section.


last.fm
Music
Radio
Events
Charts
Community
Join
Login
Help Last.fm's scientists with music research »
English | Help
Music search
All Music | Artists | Albums | Tracks | Tags | Labels
search

Search music on Last.fm

sad happy

Looking for people?
Go to the User Search or Group Search

Looking for concerts?
Go to the Event Search

Tag results for "sad happy"

sad 18,431 taggers

Music that describes feelings of disadvantage, loss, helplessness and sorrow. That one song, album or artist that always makes you cry. Songs for... read more

happy 16,777 taggers

Songs guaranteed to put you in a better mood! read more

Figure 4.5: Screenshot of Last.fm, showing the number of people tagging songs as 'sad' and 'happy'.

4.3 DEAP Experiments

The results from the pilot study described above were sufficiently encouraging to warrant further investigation. In this section we describe an extension of that study, designed to address several shortcomings of the pilot study. First, the relatively small number of participants and number of trials per participant in the pilot study limited the statistical significance of the results. Therefore, in the next experiment we recorded 32 participants with 40 trials per participant. Secondly, the stimulus material in the pilot study was all manually selected by the experimenters. In this experiment we introduce a semi-automated method of stimuli selection based on emotional keywords and the Last.fm database. We report results on both single-trial classification and single-trial regression.

4.3.1 Stimuli selection

The stimuli used in the experiment were selected in several steps. First, we selected 120 initial stimuli, half of which were chosen semi-automatically and half manually. Then, a one-minute highlight part was determined for each stimulus. Finally, through a web-based subjective assessment experiment, 40 final stimuli were selected. Each of these steps is explained below.

4.3.1.1 Initial stimuli selection

Eliciting emotional reactions from test participants is a difficult task and selecting the most effective stimulus materials is crucial. We propose here a semi-automated method for stimulus selection, with the goal of minimizing the bias arising from manual stimuli selection.

60 of the 120 initially selected stimuli were selected using the Last.fm[6] music enthusiast website. Last.fm allows users to track their music listening habits and receive recommendations for new music and events. Additionally, it allows the users to assign tags to individual songs, thus creating a folksonomy of tags. Many of the tags carry emotional meanings, such as 'depressing' or 'aggressive'. Last.fm offers an API, allowing one to retrieve tags and tagged songs (see also Figure 4.5).

A list of emotional keywords was taken from[114] and expanded to include inflections and synonyms, yielding a total of 304 keywords. Next, for each keyword, corresponding tags were found in the Last.fm database. For each found affective tag, the ten songs most often labelled with this tag were selected. This resulted in a total of 1084 songs.

The valence-arousal space can be subdivided into 4 quadrants, namely low arousal and low valence (LALV), low arousal and high valence (LAHV), high arousal and low valence (HALV) and high arousal and high valence (HAHV). In order to ensure diversity of induced emotions, from the 1084 songs, 15 were selected manually for each quadrant according to the following criteria:

Does the tag accurately reflect the emotional content?

Examples of songs subjectively rejected according to this criterion include songs that are tagged merely because the song title or artist name corresponds to the tag. Also, in some cases the lyrics may correspond to the tag, but the actual emotional content of the song is entirely different (e.g. happy songs about sad topics).

Is a music video available for the song?

Music videos for the songs were automatically retrieved from YouTube, corrected manually where necessary. However, many songs do not have a music video and thus had to be rejected for the purposes of this experiment.

[6] http://www.last.fm

Is the song appropriate for use in the experiment?

Since our test participants were mostly European students, we selected those songs most likely to elicit emotions for this target demographic. Therefore, mainly European or North American artists were selected.

In addition to the songs selected using the method described above, 60 stimulus videos were selected manually, with 15 videos selected for each of the quadrants in the arousal/valence space. The goal here was to select those videos expected to induce the most clear emotional reactions for each of the quadrants. The combination of manual selection and selection using affective tags produced a list of 120 candidate stimulus videos.

4.3.1.2 Detection of one-minute highlights

The analysis in this Section was provided by Mohammad Soleymani from the University of Geneva. For each of the 120 initially selected music videos, a one-minute segment for use in the experiment was extracted. In order to extract a segment with maximum emotional content, an affective highlighting algorithm is used.

Soleymani et al.[138] used a linear regression method to calculate arousal for each shot in movies. In their method, the arousal and valence of shots was estimated using a linear regression on the content-based features. 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. There are other content features such as colour variance and key lighting that have been shown to be correlated with valence[159]. The detailed description of the content features used in this work is given in Section 4.3.8.

In order to find the best weights for arousal and valence estimation using regression, the regressors were trained on all shots in 21 annotated movies in the dataset presented in [138]. The linear weights were estimated by means of a relevance vector machine (RVM) from the RVM toolbox provided by Tipping[145]. The RVM is able to reject uninformative features during its training hence no further feature selection was used for arousal and valence determination.

The music videos were segmented into one-minute segments with 55 seconds overlap between segments. Content features were extracted and provided the input for the regressors. The emotional highlight score of the i-th segment e_i was estimated using the following equation:

$$e_i = \sqrt{a_i^2 + v_i^2} \tag{4.1}$$

The arousal, a_i, and valence, v_i, were centered. 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. In these cases, the one-minute highlight was selected so that these segments were included.

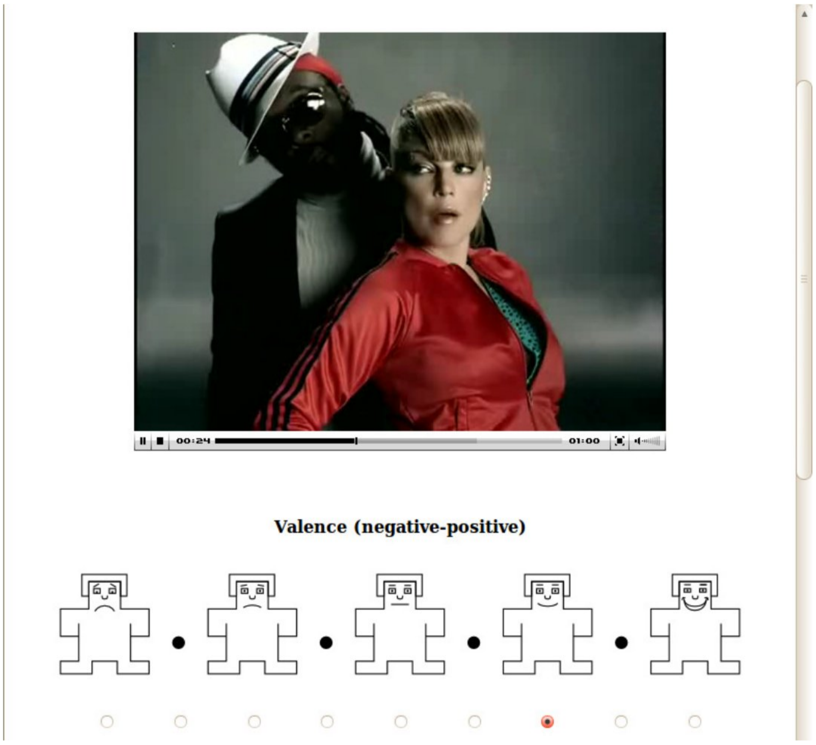
Given the 120 one-minute music video segments, the final selection of 40 videos used in the experiment was made on the basis of subjective ratings by volunteers, as described in the next section.

4.3.1.3 Online subjective annotation

From the initial collection of 120 stimulus videos, the final 40 test video clips were chosen by using a web-based subjective emotion assessment interface. Participants watched music videos and rated them on a discrete 9-point scale for valence, arousal and dominance. Participants were instructed to rate their felt (not perceived) emotion as a result of the music video (as a whole). A screenshot of the interface is shown in Fig. 4.6. Each participant watched as many videos as he/she wanted and was able to end the rating at any time. The order of the clips was randomized, but preference was given to the clips rated by the least number of participants. This ensured a similar number of ratings for each video (14-16 assessments per video were collected). It was ensured that participants never saw the same video twice.

The dominance dimension was included here as several researchers have argued that a two-dimensional model of emotion is not sufficient[48]. The dominance scale ranges from submissive (feeling weak, without control) to dominant (feeling strong and in control). It allows distinction of certain emotions that lie close to each other in the velence-arousal model, but intuitively differ substantially, such as fear and anger. While difficult to grasp intuitively in relation to the setting described here (the viewing of music videos in a laboratory setting), the dominance measure has been used before in similar settings[135] and was shown to vary significantly with different musical compositions. Section 4.3.4 provides a further analysis of the ratings given by participants.

After all of the 120 videos were rated by at least 14 volunteers each, the final 40 videos for use in the experiment were selected. 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. To this end, for each video x we calculated a normalized



Valence (negative-positive)

○ ○ ○ ○ ○ ○ ● ○ ○

Figure 4.6: Screenshot of the web interface for subjective emotion assessment.

arousal and valence score μ_x/σ_x by taking the mean rating μ_x divided by the standard deviation σ_x.

Then, for each quadrant in the normalized valence-arousal space, we selected the 10 videos that lie closest to the extreme corner of the quadrant. Fig. 4.7 shows the score for the ratings of each video and the selected videos highlighted in green. The video whose rating was closest to the extreme corner of each quadrant is mentioned explicitly. Of the 40 selected videos, 17 were selected via Last.fm affective tags, indicating that useful stimuli can be selected via this method.

4.3.2 Materials and Setup

The experiments were performed in two laboratory environments with controlled illumination. EEG and peripheral physiological signals were recorded using a Biosemi ActiveTwo system[7] on a dedicated recording PC (Pentium 4, 3.2 GHz). Stimuli were presented using a dedicated stimulus PC (Pentium 4, 3.2 GHz) that sent synchronization markers directly to the recording PC. For presentation of the stimuli and recording the users' ratings, the Presentation software[8] was used. The music videos

[7] BioSemi instrumentation (http://www.biosemi.com)

[8] Presentation by Neurobehavioral systems (http://www.neurobs.com)

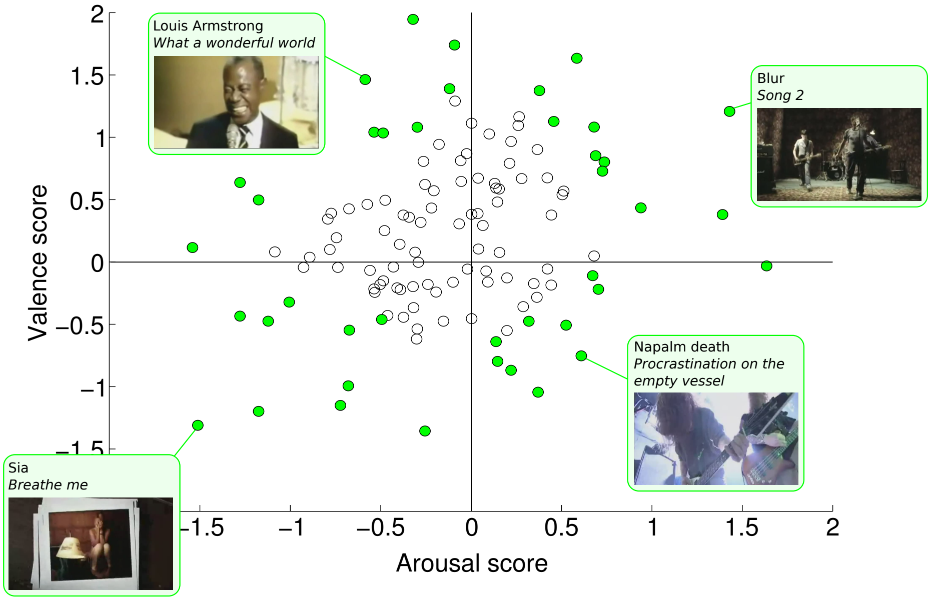


Figure 4.7: μ_x/σ_x value for the ratings of each video in the online assessment. Videos selected for use in the experiment are highlighted in green. For each quadrant, the most extreme video is detailed with the song title and a screenshot from the video.

were presented on a 17-inch screen (1280 × 1024, 60 Hz) and in order to minimize eye movements, all video stimuli were displayed at 800 × 600 resolution, filling approximately 2/3 of the screen. Participants were seated approximately 1 meter from the screen. Stereo Philips speakers were used and the music volume was set at a relatively loud level, however each participant was asked before the experiment whether the volume was comfortable and it was adjusted when necessary.

EEG was recorded with a sampling rate of 512 Hz using 32 active AgCl electrodes (placed according to the international 10-20 system). Thirteen peripheral physiological signals (which will be further discussed in section 4.3.7) were also recorded. Additionally, for the first 22 of the 32 participants, frontal face video was recorded in DV quality using a Sony DCR-HC27E consumer-grade camcorder. Fig. 4.8 illustrates the electrode placement for acquisition of peripheral physiological signals. For more details regarding the contents of the dataset, see also Appendix B.

The collected dataset was made publicly available to the research community[9]. To the best of our knowledge, the dataset is the largest (in terms of number of participants) publicly available dataset concerning emotion assessment from EEG and peripheral signals. As of the writing of this thesis, researchers from 81 different institutions have requested access.

[9] DEAP Dataset: http://www.eecs.qmul.ac.uk/mmv/datasets/deap/

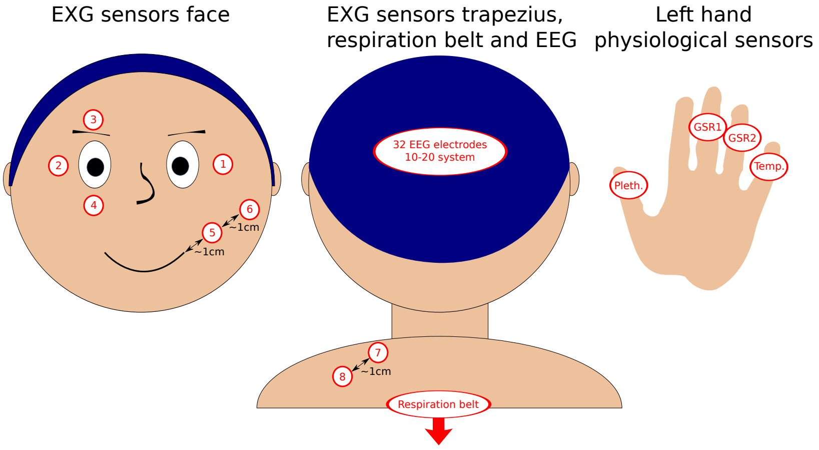

EXG sensors face
EXG sensors trapezius, respiration belt and EEG
Left hand physiological sensors
3
2
1
4
6
5
~1cm
~1cm
32 EEG electrodes 10-20 system
7
8
~1cm
Respiration belt
GSR1
GSR2
Temp.
Pleth.

Figure 4.8: Placement of peripheral physiological sensors. For Electrodes were used to record EOG and 4 for EMG (zygomaticus major and trapezius muscles). In addition, GSR, blood volume pressure (BVP), temperature and respiration were measured.

4.3.3 Experiment protocol

32 Healthy participants (50% female), aged between 19 and 37 (mean age 26.9), participated in the experiment. Prior to the experiment, each participant signed a consent form and filled out a questionnaire. Next, they were given a set of instructions to read informing them of the experiment protocol and the meaning of the different scales used for self-assessment. An experimenter was present during this time to answer any questions. When the instructions were clear to the participant, he/she was led into the experiment room. After the sensors were placed and their signals checked, the participants performed a practice trial to familiarize themselves with the system. In this unrecorded trial, a short video was shown, followed by a self-assessment by the participant. Next, the experimenter started the physiological signals recording and left the room, after which the participant started the experiment by pressing a key on the keyboard.

The experiment started with a 2 minute baseline recording, during which a fixation cross was displayed to the participant (who was asked to relax during this period). Then the 40 videos were presented in 40 trials, each consisting of the following steps:

1. A 2 second screen displaying the current trial number to inform the participants of their progress.

2. A 5 second baseline recording (fixation cross).

3. The 60 second display of the music video.

4. Self-assessment for arousal, valence, liking and dominance. Participants were

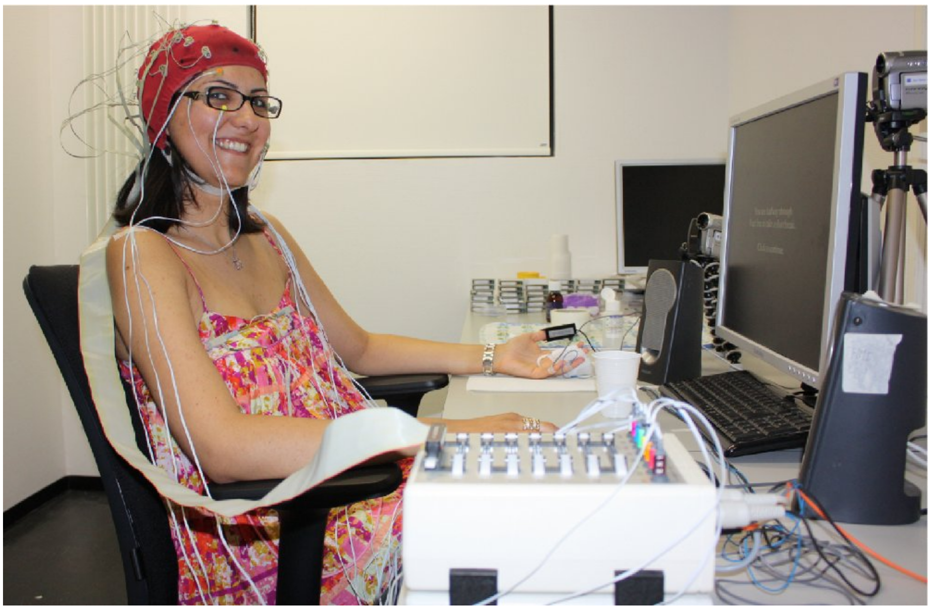


Figure 4.9: A participant shortly before the experiment.

instructed to rate their felt (not perceived) emotion as a result of the music video (as a whole).

After 20 trials, the participants took a short break. During the break, they were offered some cookies and non-caffeinated, non-alcoholic beverages. The experimenter then checked the quality of the signals and the electrodes placement and the participants were asked to continue the second half of the test. Fig. 4.9 shows a participant shortly before the start of the experiment.

At the end of each trial, participants performed a self-assessment of their levels of arousal, valence, liking and dominance on a level from 1 to 9 in the same fashion as during the pilot study. Finally, after the experiment, 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”).

4.3.4 Analysis of subjective ratings

The analysis in this Section was partially provided by Christian Mühl from the University of Twente and Mohammad Soleymani from the University of Geneva.

In this section we describe the effect the affective stimulation had on the subjective ratings obtained from the participants. Firstly, we will provide descriptive statistics for the recorded ratings of liking, valence, arousal, dominance, and familiarity. Secondly, we will discuss the covariation of the different ratings with each other.

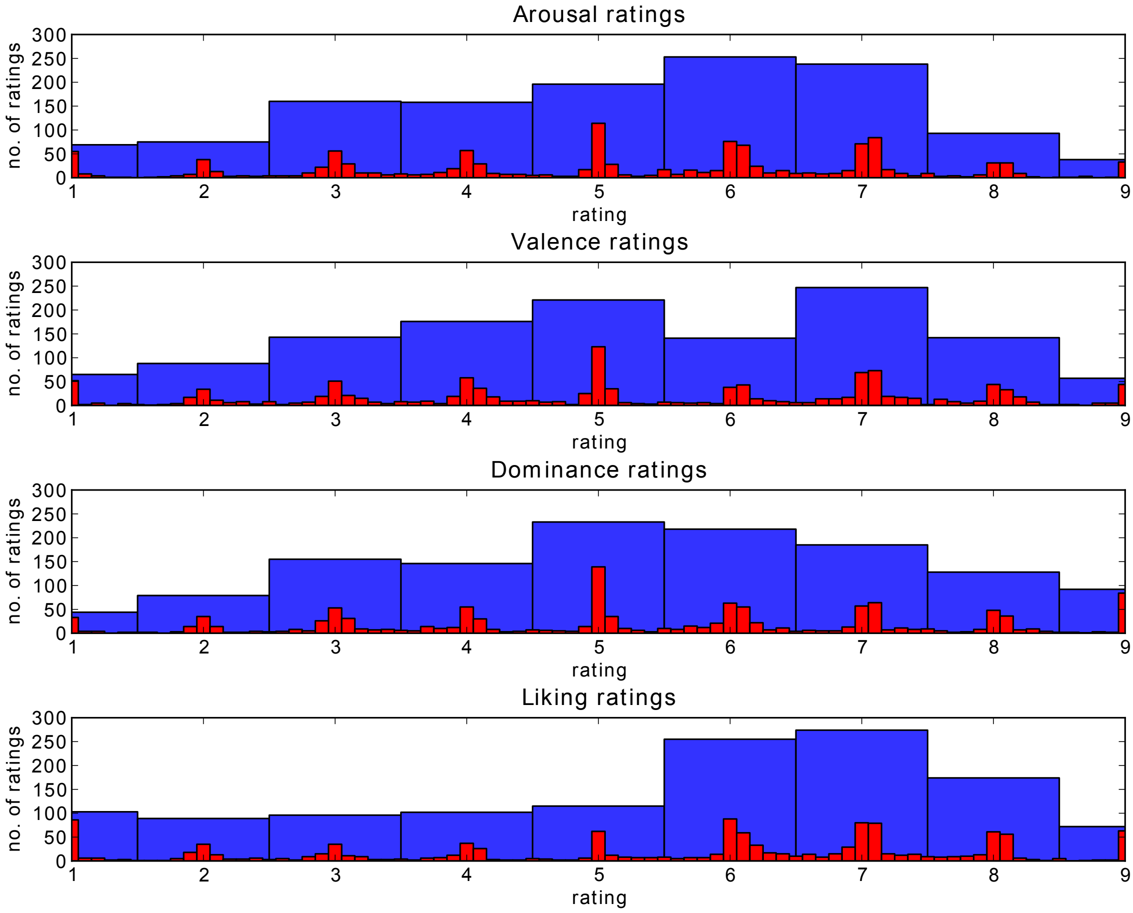


Figure 4.10: Histogram of arousal, valence, dominance and liking ratings given to all videos by the 32 participants. The blue bars are the histogram of the quantized ratings in nine levels.The red bars are showing the ratings quantized in 80 levels (The quantization step is equal to 0.1).

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. 4.10). The blue bars on Fig. 4.10 show the ratings' histograms quantized in nine levels. From the blue bars, we can see that the distribution of the ratings are skewed towards higher scores.

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 affective ratings with $mean(\kappa) = 0.02 \pm 0.06$ for arousal, $mean(\kappa) = 0.08 \pm 0.08$ for valence, and $mean(\kappa) = 0.05 \pm 0.08$ 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 ($p = 4.5 \times 10^{-7}$).

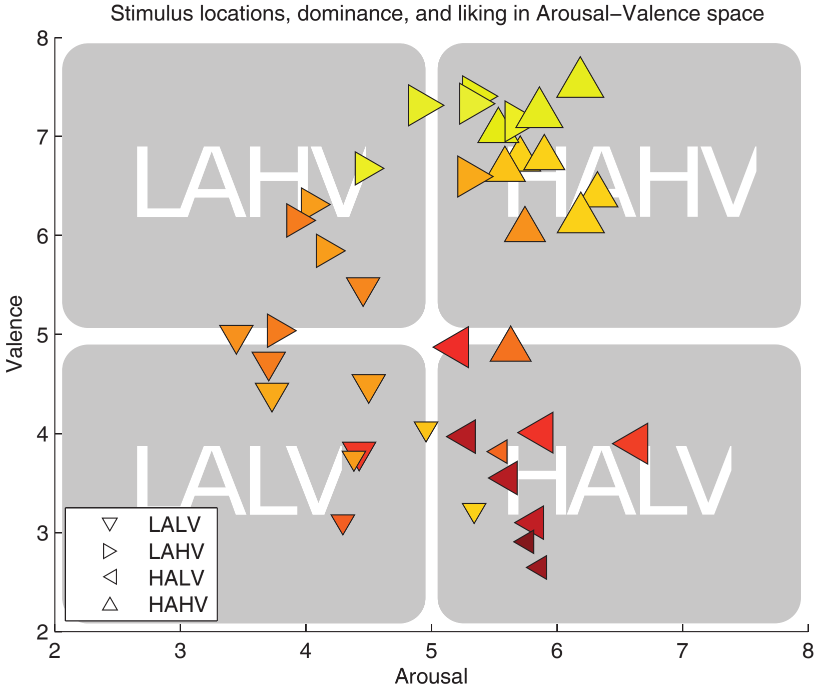


Figure 4.11: The mean locations of the stimuli on the arousal-valence plane for the 4 conditions (LALV, HALV, LAHV, HAHV). Liking is encoded by color: dark red is low liking and bright yellow is high liking. Dominance is encoded by symbol size: small symbols stand for low dominance and big for high dominance.

Stimuli were selected to induce emotions in the four quadrants of the valence-arousal space (LALV, HALV, LAHV, HAHV). The stimuli from these four affect elicitation conditions generally resulted in the elicitation of the target emotion aimed for when the stimuli were selected, ensuring that large parts of the arousal-valence plane (AV plane) are covered (see Fig. 4.11). Wilcoxon signed-rank tests showed that low and high arousal stimuli induced different valence ratings ($p < .0001$ and $p < .00001$). Similarly, low and high valenced stimuli induced different arousal ratings ($p < .001$ and $p < .0001$).

The emotion elicitation worked specifically well for the high arousing conditions, yielding relative extreme valence ratings for the respective stimuli. The stimuli in the low arousing conditions were less successful in the elicitation of strong valence responses. Furthermore, some stimuli of the LAHV condition induced higher arousal than expected on the basis of the online study. Interestingly, this results in a C-shape of the stimuli on the valence-arousal plane also observed in the well-validated ratings for the international affective picture system (IAPS) [85] and the international affective digital sounds system (IADS) [22], indicating the general difficulty to induce

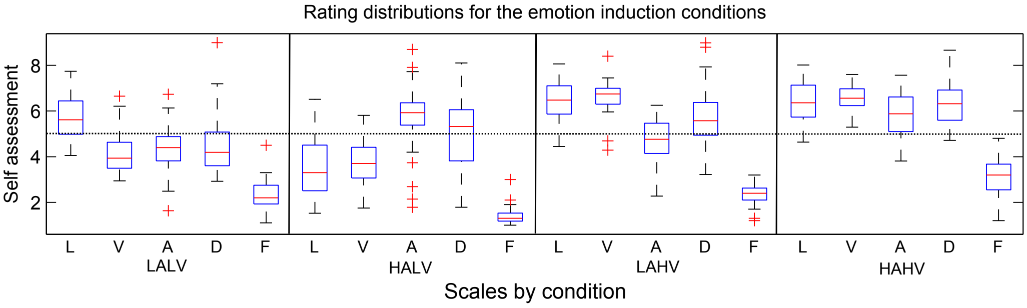


Figure 4.12: The distribution of the participants' subjective ratings per scale (L - general rating, V - valence, A - arousal, D - dominance, F - familiarity) for the 4 affect elicitation conditions (LALV, HALV, LAHV, HAHV).

emotions with strong valence but low arousal. The distribution of the individual ratings per conditions (see Fig. 4.12) shows a large variance within conditions, resulting from between-stimulus and -participant variations, possibly associated with stimulus characteristics or inter-individual differences in music taste, general mood, or scale interpretation. However, the significant differences between the conditions in terms of the ratings of valence and arousal reflect the successful elicitation of the targeted affective states (see Table 4.6).

Table 4.6: The mean values (and standard deviations) of the different ratings of liking (1-9), valence (1-9), arousal (1-9), dominance (1-9), familiarity (1-5) for each affect elicitation condition.

Cond.	**Liking**	**Valence**	**Arousal**	**Dom.**	**Fam.**
LALV	5.7 (1.0)	4.2 (0.9)	4.3 (1.1)	4.5 (1.4)	2.4 (0.4)
HALV	3.6 (1.3)	3.7 (1.0)	5.7 (1.5)	5.0 (1.6)	1.4 (0.6)
LAHV	6.4 (0.9)	6.6 (0.8)	4.7 (1.0)	5.7 (1.3)	2.4 (0.4)
HAHV	6.4 (0.9)	6.6 (0.6)	5.9 (0.9)	6.3 (1.0)	3.1 (0.4)

The distribution of ratings for the different scales and conditions suggests a complex relationship between ratings. We explored the mean inter-correlation of the different scales over participants (see Table 4.7), as they might be indicative of possible confounds or unwanted effects of habituation or fatigue. We observed high positive correlations between liking and valence, and between dominance and valence. Seemingly, without implying any causality, people liked music which gave them a positive feeling and/or a feeling of empowerment. Medium positive correlations were observed between arousal and dominance, and between arousal and liking. Familiarity correlated moderately positive with liking and valence. As already observed above, the scales of valence and arousal are not independent, but their positive correlation is rather low, suggesting that participants were able to differentiate between these two important concepts. Stimulus order had only a small effect on liking and dominance ratings, and no significant relationship with the other ratings, suggesting that effects of habituation and fatigue were kept to an acceptable minimum.

Table 4.7: The means of the subject-wise inter-correlations between the scales of valence, arousal, liking, dominance, familiarity and the order of the presentation (i.e. time) for all 40 stimuli. Significant correlations ($p < .05$) according to Fisher's method are indicated by stars.

Scale	Liking	Valence	Arousal	Dom.	Fam.	Order
Liking	1	0.62*	0.29*	0.31*	0.30*	0.03*
Valence		1	0.18*	0.51*	0.25*	0.02
Arousal			1	0.28*	0.06*	0.00
Dominance				1	0.09*	0.04*
Familiarity					1	-
Order						1

In summary, the affect elicitation was in general successful, though the low valence conditions were partially biased by moderate valence responses and higher arousal. High scale inter-correlations observed are limited to the scale of valence with those of liking and dominance, and might be expected in the context of musical emotions. The rest of the scale inter-correlations are small or medium in strength, indicating that the scale concepts were well distinguished by the participants.

4.3.5 Correlates of EEG and ratings

The analysis in this Section was provided by Christian Mühl from the University of Twente.

For the investigation of the correlates of the subjective ratings with the EEG signals, the EEG data was common average referenced, down-sampled to 256 Hz, and high-pass filtered with a 2 Hz cutoff-frequency using the EEGlab[10] toolbox. We removed eye artefacts with a blind source separation technique[11]. Then, the signals from the last 30 seconds of each trial (video) were extracted for further analysis. To correct for stimulus-unrelated variations in power over time, the EEG signal from the five seconds before each video was extracted as baseline.

[10] http://sccn.ucsd.edu/eeglab/

[11] http://www.cs.tut.fi/~gomezher/projects/eeg/aar.htm

	Theta				**Alpha**			
	Elec.	$\bar{R}$	R^{-}	R^{+}	Elec.	$\bar{R}$	R^{-}	R^{+}
Arousal	CP6*	-0.06	-0.47	0.25	Cz*	-0.07	-0.45	0.23
Valence	Oz**	0.08	-0.23	0.39	PO4*	0.05	-0.26	0.49
	PO4*	0.05	-0.26	0.49				
Liking	C3*	0.08	-0.35	0.31	AF3*	0.06	-0.27	0.42
					F3*	0.06	-0.42	0.45

	Beta				**Gamma**			
	Elec.	$\bar{R}$	R^{-}	R^{+}	Elec.	$\bar{R}$	R^{-}	R^{+}
Arousal	FC2*	-0.06	-0.40	0.28				
Valence	CP1**	-0.07	-0.49	0.24	T7**	0.07	-0.33	0.51
	Oz*	0.05	-0.24	0.48	CP6*	0.06	-0.26	0.43
	FC6*	0.06	-0.52	0.49	CP2*	0.08	-0.21	0.49
	Cz*	-0.04	-0.64	0.30	C4**	0.08	-0.31	0.51
					T8**	0.08	-0.26	0.50
					FC6**	0.10	-0.29	0.52
					F8*	0.06	-0.35	0.52
Liking	FC6*	0.07	-0.40	0.48	T8*	0.04	-0.33	0.49

Table 4.8: The electrodes for which the correlations with the scale were significant (*=p < .01, **=p < .001). Also shown is the mean of the subject-wise correlations ($\bar{R}$), the most negative (R^{-}), and the most positive correlation (R^{+}).

The frequency power of trials and baselines between 3 and 47 Hz was extracted with Welch's method with windows of 256 samples. The baseline power was then subtracted from the trial power, yielding the change of power relative to the pre-stimulus period. These changes of power were averaged over the frequency bands of theta (3 - 7 Hz), alpha (8 - 13 Hz), beta (14 - 29 Hz), and gamma (30 - 47 Hz). For the correlation statistic, we computed the Spearman correlated coefficients between the power changes and the subjective ratings, and computed the p-values for the left- (positive) and right-tailed (negative) correlation tests. This was done for each participant separately and, assuming independence [88], the 32 resulting p-values per correlation direction (positive/negative), frequency band and electrode were then combined to one p-value via Fisher's method [95].

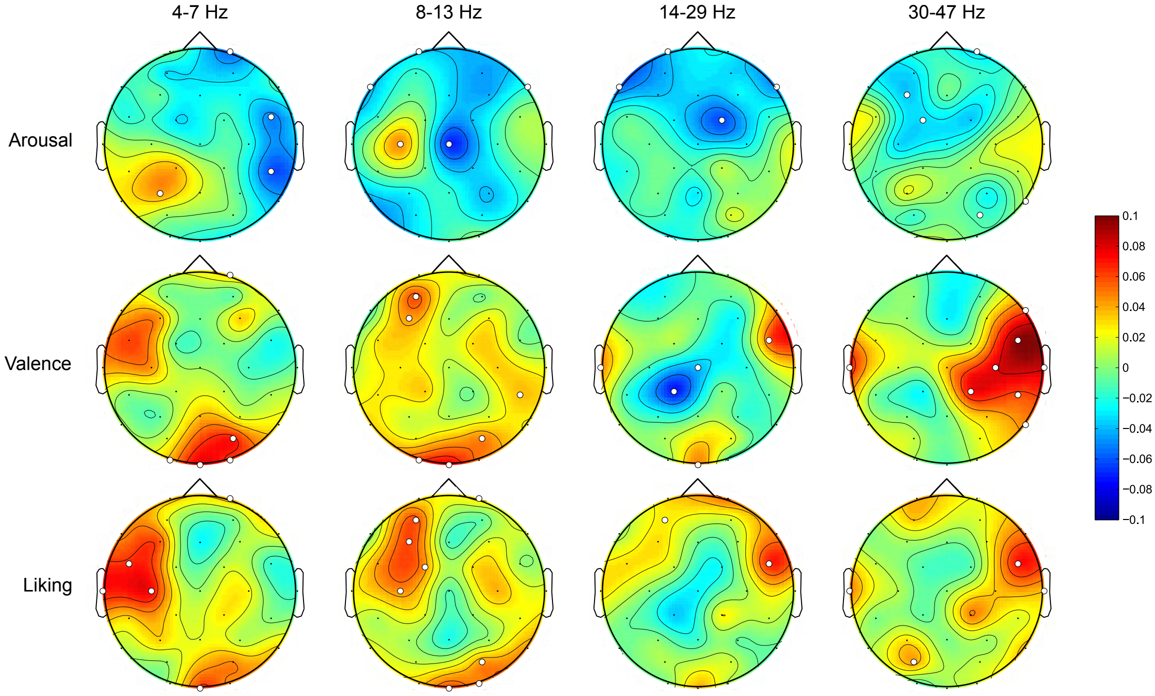


Figure 4.13: The mean correlations (over all participants) of the valence, arousal, and general ratings with the power in the broad frequency bands of theta (4-7 Hz), alpha (8-13 Hz), beta (14-29 Hz) and gamma (30-47 Hz). The highlighted sensors correlate significantly ($p < .05$) with the ratings.

Fig. 4.13 shows the (average) correlations with significantly ($p < .05$) correlating electrodes highlighted. Below we will report and discuss only those effects that were significant with $p < .01$. A comprehensive list of the effects can be found in Table 4.3.5.

For arousal we found negative correlations in the theta, alpha, and gamma band. The central alpha power decrease for higher arousal matches the findings from our earlier pilot study [5] and an inverse relationship between alpha power and the general level of arousal has been reported before [12, 11].

Valence showed the strongest correlations with EEG signals and correlates were found in all analysed frequency bands. In the low frequencies, theta and alpha, an increase of valence led to an increase of power. This is consistent with the findings in the pilot study. The location of these effects over occipital regions, thus over visual cortices, might indicate a relative deactivation, or top-down inhibition, of these due to participants focusing on the pleasurable sound [77]. For the beta frequency band we found a central decrease, also observed in the pilot, and an occipital and right temporal increase of power. Increased beta power over right temporal sites was associated with positive emotional self-induction and external stimulation by [30]. Similarly, [105] has reported a positive correlation of valence and high-frequency power, including beta and gamma bands, emanating from anterior temporal cerebral sources. Correspondingly, we observed a highly significant increase of left and especially right temporal gamma power. However, it should be mentioned that EMG (muscle) activity is also prominent in the high frequencies, especially over anterior and temporal electrodes [53].

The liking correlates were found in all analysed frequency bands. For theta and alpha power we observed increases over left fronto-central cortices. Liking might be associated with an approach motivation. However, the observation of an increase of left alpha power for a higher liking conflicts with findings of a left frontal activation, leading to lower alpha over this region, often reported for emotions associated with approach motivations [59]. This contradiction might be reconciled when taking into account that it is well possible that some disliked pieces induced an angry feeling (due to having to listen to them, or simply due to the content of the lyrics), which is also related to an approach motivation, and might hence result in a left-ward decrease of alpha. The right temporal increases found in the beta and gamma bands are similar to those observed for valence, and the same caution should be applied. In general the distribution of valence and liking correlations shown in Fig. 4.13 seem very similar, which might be a result of the high inter-correlations of the scales discussed above.

Summarising, we can state that the correlations observed partially concur with observations made in the pilot study and in other studies exploring the neuro-physiological correlates of affective states. They might therefore be taken as valid indicators of emotional states in the context of multi-modal musical stimulation. However, the mean correlations are seldom bigger than ±0.1, which might be due to high inter-participant variability in terms of brain activations, as individual correlations between ±0.5 were observed for a given scale correlation at the same electrode/frequency combination. The presence of this high inter-participant variability justifies a participant-specific classification approach, as we employ it, rather than a single classifier for all participants.

4.3.6 Single-trial classification

In this section we present the methodology and results of single-trial classification of the videos. Three different modalities were used for classification, namely EEG signals, peripheral physiological signals and MCA. Conditions for all modalities were kept equal and only the feature extraction step varies.

Three different binary classification problems were posed: the classification of low/high arousal, low/high valence and low/high liking. To this end, the participants' ratings during the experiment are used as the ground truth. The ratings for each of these scales are thresholded into two classes (low and high). On the 9-point rating scales, the threshold was simply placed in the middle. Note that for some participants and scales, this leads to unbalanced classes. To give an indication of how unbalanced the classes are, the mean and standard deviation (over participants) of the percentage of videos belonging to the high class per rating scale are: arousal 59%(15%), valence 57%(9%) and liking 67%(12%).

In light of this issue, in order to reliably report results, we report the F1-score, which is commonly employed in information retrieval and takes the class balance into account, contrary to the mere classification rate. In addition, we use a naïve Bayes classifier, a simple and generalizable classifier which is able to deal with unbalanced classes in small training sets.

First, the features are extracted for each trial (video). Then, for each participant, the F1 measure was used to evaluate the performance of emotion classification in a leave-one-trial-out cross validation scheme. We use Fisher's linear discriminant J for feature selection:

$$J(f) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2} \tag{4.2}$$

where μ and σ are the mean and standard deviation for feature f. We calculate this criterion for each feature and then apply a threshold to select the maximally discriminating ones. This threshold was empirically determined at 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:

$$G(f_1, .., f_n) = \arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c) \tag{4.3}$$

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 modelling these from the training set.

The following section explains the feature extraction steps for the EEG and peripheral physiological signals. Section 4.3.8 presents the features used in MCA classification. In section 4.3.9 we explain the method used for decision fusion of the results. Finally, section 4.3.10 presents the classification results.

4.3.7 EEG and peripheral physiological features

The analysis in this Section was partially provided by Mohammad Soleymani from the University of Geneva.

Most of the current theories of emotion [31, 130] agree that physiological activity is an important component of an emotion. For instance several studies have demonstrated the existence of specific physiological patterns associated with basic emotions [41].

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 electrooculogram (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. This resistance decreases due to an increase of perspiration, which usually occurs when one is experiencing emotions such as stress or surprise. Moreover, Lang et al. discovered that the mean value of the GSR is related to the level of arousal [86].

A plethysmograph measures blood volume in the participant's thumb. This measurement can also be used to compute the heart rate (HR) by identification of local maxima (i.e. heart beats), inter-beat periods, and heart rate variability (HRV). Blood pressure and HRV correlate with emotions, since stress can increase blood pressure. Pleasantness of stimuli can increase peak heart rate response [86]. In addition to the HR and HRV features, spectral features derived from HRV were shown to be a useful feature in emotion assessment [101].

Skin temperature and respiration were recorded since they varies with different emotional states. Slow respiration is linked to relaxation while irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear.

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. For further reading on psychophysiology of emotion, we refer the reader to [83].

All the physiological responses were recorded at a 512Hz sampling rate and later down-sampled to 256Hz to reduce prcoessing 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 [25, 75, 123, 160, 137] (see also Table 4.9).

EEG Signals were acquired from 32 electrodes placed on the participants' scalps according to the international 10-20 system. Signals were recorded with a sampling rate of 512Hz and then down-sampled to 128Hz to simplify further processing. EOG was recorded from 4 electrodes placed to the sides and above/below the eyes, and eye artefacts were suppressed using the algorithm proposed in [133]. Next, a band pass filter of 4-45Hz was used to further reduce signal artefacts and remove the 50Hz power line interference. Finally, the data was referenced to the common average.

As in the pilot study, power spectral density (PSD) in different frequency bands was estimated using Welch's method for each 60 second trial. The frequency bands were: theta (4 - 8Hz), slow alpha (8 - 10Hz), alpha (8 - 12Hz), beta (12 - 30Hz) and gamma (30 - 45Hz). We also computed the lateralization for 14 left-right pairs of electrodes by computing the difference in PSD for each frequency band. The EEG feature vector is composed of 230 features (5 × (32 channels + 14 asymmetry pairs)). Table 4.9 summarizes the list of features extracted from the physiological signals.

4.3.8 MCA Features

The analysis in this Section was provided by Mohammad Soleymani from the University of Geneva.

Music videos were encoded into the MPEG-1 format to extract motion vectors and I-frames for further feature extraction. The video stream has been segmented at the

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
Blood volume pressure	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 Spectral power for each electrode. The spectral power asymmetry between 14 pairs of electrodes.

shot level using the method proposed in [73].

From a movie director's point of view, lighting key [124, 159] and color variance [159] are important tools to evoke emotions. We therefore extracted lighting key from frames in the HSV space by multiplying the average value V (in HSV) by the standard deviation of the values V (in HSV). Color variance was obtained in the CIE LUV color space by computing the determinant of the covariance matrix of L, U, and V.

Hanjalic and Xu [58] showed the relationship between video rhythm and affect. The average shot change rate, and shot length variance were extracted to characterize video rhythm. Fast moving scenes or objects' movements in consecutive frames are also an effective factor for evoking 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 are important parameters to elicit emotions [149]. A 20 bin color histogram of hue and lightness values in the HSV space was computed for each I-frame and subsequently averaged over all frames. The resulting bin averages were used as video content-based features. The median of the L value in HSL space was computed to obtain the median lightness of a frame.

Finally, visual cues representing shadow proportion, visual excitement, grayness and details were also determined according to the definition given in [159].

Sound also has an important impact on affect. For example, loudness of speech (energy) is related to evoked arousal, while rhythm and average pitch in speech signals are related to valence [117]. The audio channels of the videos were extracted and encoded into mono MPEG-3 format at a sampling rate of 44.1 kHz. All 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 4.10, are commonly used in audio and speech processing and audio classification [91, 96]. MFCC, formants and the pitch of audio signals were extracted using the PRAAT software package [20].

4.3.9 Fusion of single-modality results

The analysis in this Section was provided by Jong-Seok Lee from the École polytechnique fédérale de Lausanne (EPFL).

Fusion of the multiple modalities explained above aims at improving classification results by exploiting the complementary nature of the different modalities. In general, approaches for modality fusion can be classified into two broad categories, namely, feature fusion (or early integration) and decision fusion (or late integration) [89]. In feature fusion, the features extracted from signals of different modalities are con-

Table 4.10: Low-level features extracted from audio signals.

Feature category	Extracted features
MFCC	MFCC coefficients (13 features) [91], Derivative of MFCC (13 features), Autocorrelation of MFCC (13 features)
Energy	Average energy of audio signal [91]
Formants	Formants up to 5500Hz (female voice) (five features)
Time frequency	MSpectrum flux, Spectral centroid, Delta spectrum magnitude, Band energy ratio[91, 96]
Pitch	First pitch frequency
Zero crossing rate	Average, Standard deviation [91]
Silence ratio	Proportion of silence in a time window [27]

catenated to form a composite feature vector and then inputted to a recognizer. In decision fusion, on the other hand, each modality is processed independently by the corresponding classifier and the outputs of the classifiers are combined to yield the final result. Each approach has its own advantages. For example, implementing a feature fusion-based system is straightforward, while a decision fusion-based system can be constructed by using existing unimodal classification systems. Moreover, feature fusion can consider synchronous characteristics of the involved modalities, whereas decision fusion allows us to model asynchronous characteristics of the modalities flexibly.

An important advantage of decision fusion over feature fusion is that, since each of the signals are processed and classified independently in decision fusion, it is relatively easy to employ an optimal weighting scheme to adjust the relative amount of the contribution of each modality to the final decision according to the reliability of the modality. The weighting scheme used in our work can be formalized as follows: For a given test datum X, the classification result of the fusion system is

$$c^* = \arg\max_i \left\{ \prod_{m=1}^{M} P_i(X|\lambda_m)^{\alpha_m} \right\} \tag{4.4}$$

where M is the number of modalities considered for fusion, λ_m is the classifier for the

m-th modality, and $P_i(X|\lambda_m)$ is its output for the i-th class. The weighting factors α_m, which satisfy $0 \leq \alpha_m \leq 1$ and $\sum_{m=1}^{M} \alpha_m = 1$, determine how much each modality contributes to the final decision and represent the modality's reliability.

We adopt a simple method where the weighting factors are fixed once their optimal values are determined from the training data. The optimal weight values are estimated by exhaustively searching the regular grid space, where each weight is incremented from 0 to 1 by 0.01 and the weighting values producing the best classification results for the training data are selected.

4.3.10 Results

Parts of the analysis in this Section were provided by Mohammad Soleymani from the University of Geneva.

	Arousal		Valence		Liking	
Modality	**CR**	**F1**	**CR**	**F1**	**CR**	**F1**
EEG	0.620	$0.583^{\ddagger}$	0.576	$0.563^{\ddagger}$	0.554	0.502
Peripheral	0.570	$0.533^{\dagger}$	0.627	$0.608^{\ddagger}$	0.591	$0.538^{\ddagger}$
MCA	0.651	$0.618^{\ddagger}$	0.618	$0.605^{\ddagger}$	0.677	$0.634^{\ddagger}$
Random	0.500	0.483	0.500	0.494	0.500	0.476
Majority class	0.644	0.389	0.586	0.368	0.670	0.398
Class ratio	0.562	0.500	0.525	0.500	0.586	0.500

Table 4.11: Average classification rates (CR) and F1-scores (F1, average of score for each class) over participants. Daggers indicate whether the F1-score distribution over participants is significantly higher than 0.5 according to an independent one-sample t-test ($^{\ddagger}= p < .01$, $^{\dagger}= p < .05$). For comparison, expected results are given for classification based on random classification, classification according to the majority class and classification with the ratio of the classes.

Table 4.11 shows the average classification rates and F1-scores (average F1-score for both classes) over participants for each rating scale. We compare the results to the expected values (analytically determined) of classifying randomly, classifying according to the majority class in the training data, and classifying by choosing a class with the probability of its occurrence in the training data. For determining the expected values of majority voting and class ratio classification, we used the class ratio of each participant's ratings during the experiment. These results are slightly too high, as in reality the class ratio would have to be estimated from the training set in each fold of the leave-one-trial-out cross-validation.

	Arousal			**Valence**			**Liking**		
	Mod.	W_{opt}	W_{eq}	Mod.	W_{opt}	W_{eq}	Mod.	W_{opt}	W_{eq}
Best mod.	MCA	0.618	—	PER	0.608	—	MCA	0.634	—
Best 2 mod.	EEG MCA	0.631	0.629	MCA PER	0.638	0.652	MCA PER	0.622	0.642
All 3 mod.	All	0.616	0.618	All	0.647	0.640	All	0.618	0.607

Table 4.12: F1-scores for fusion of the best two modalities and all three modalities using the equal weights(W_{eq}) and optimal weights(W_{opt}) scheme. For comparison, the F1-score for the best single modality is also given.

Classifying according to the class ratio gives an expected F1-score of 0.5 for each participant. To test for significance, an independent one-sample t-test was performed, comparing the F1-distribution over participants to the 0.5 baseline. As can be seen from Table 4.11, 2 of the 3 obtained F1-scores are significantly better than the class ratio baseline. The exception is the classification of liking ($p = 0.068$). When classifying according to the majority class, relatively high accuracies are achieved, due to the imbalanced classes. However, this classification scheme also has the lowest F1-scores.

Overall, classification using the MCA features fares significantly better than EEG and peripheral ($p < 0.0001$ for both), while EEG and peripheral scores 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 complementary, where EEG scores best for arousal, peripheral for valence and MCA for liking. Of the different rating scales, valence classification performed best, followed by liking and lastly arousal.

Table 4.12 gives the results of multi-modal fusion. Two fusion methods were employed; the method described in section 4.3.9 and the basic method where each modality is weighed equally. The best results were obtained when only the two best-performing modalities were considered. Though fusion generally outperforms the single modalities, it is only significant for the case of MCA,PER weighted equally in the valence scale ($p = 0.025$).

The results in this experiment are much better than for the pilot study. This may be due to the increased number of training videos (39 vs. 19). Concerning the differences between the rating scales, it appears that arousal consistently scores best. The prediction of liking on the other hand is consistently poor. This may be due to liking in general being a more high-level decision.

While the presented results are significantly higher than chance level, there remains much room for improvement. Signal noise, individual physiological differences and limited reliability of self-assessments make single-trial classification challenging.

4.3.11 Single-trial regression

Parts of the analysis in this Section were provided by Mohammad Soleymani from the University of Geneva.

Using the same feature set as for classification, we also performed single-trial regression on the dataset. The goal here is to train a regression function to map the features to the rating given by the participants in the experiments. We used a basic linear ridge regression algorithm with $\alpha = 10$ as implemented in the mlpy package[12]. Other, more complicated methods such as Guassian Process Regression[164] and Relevance Vector Machine regression[145] were also investigated. Those regression methods were not found to improve the results, possibly due to the noisy nature of the signals, leading to overfitting.

For each participant, a regressor is trained and tested using a leave-one-trial-out cross validation strategy, for which the results are reported below. For comparison, we also present results obtained by a random regressor and a Π-regressor. The Π-regressor uses an estimate of the probability density function of the ground truth ratings to generate its output. This Π-regressor is included for comparison as the ground truth is not uniformly distributed (see Fig. 4.10). The random and Π-regressors were each run 10000 times and their performance was averaged.

Table 4.13 shows the regression MAE (Mean Absolute Error) for the different modalities and rating scales, with the ratings on a continuous scale between 1 and 9. We performed a two-sided repeated samples t-test on the MAE-scores per participant between each modality and the Π-regressor. The MAE-scores are always significantly higher than Π-regression for the EEG and MCA modalities. For the peripheral physiological signals, the results are only significant for the arousal modality. This indicates information regarding a participant's emotional state exists in the EEG measurements. It also seems evidence to predict the user's emotional reaction to a video is available in the content features extracted from the videos. Although for valence and liking, better results are attained, they are not significantly better. The fact that for regression all results are significantly better than random suggests that the method was able to benefit from the extra information present in the ratings that is discarded when using thresholding in the binary classification case.

[12] https://mlpy.fbk.eu/

	Arousal	Valence	Dominance	Liking
EEG	1.53(0.40)‡	**1.59(0.39)**‡	**1.53(0.49)**‡	1.78(0.51)‡
Peripheral	1.70(0.51)†	1.81(0.41)	1.64(0.49)	1.96(0.64)
MCA	**1.50(0.45)**‡	1.65(0.35)‡	**1.47(0.46)**‡	**1.68(0.45)**‡
EEG/Per/MCA	1.49(0.42)‡	1.56(0.36)‡	1.51(0.49)‡	1.66(0.46)‡
EEG/MCA	**1.47(0.42)**‡	**1.55(0.39)**‡	**1.46(0.48)**‡	**1.62(0.45)**‡
EEG/Per	1.58(0.43)‡	1.63(0.39)‡	1.57(0.50)‡	1.83(0.53)*
Per/MCA	1.53(0.52)‡	1.66(0.38)‡	1.50(0.46)‡	1.72(0.52)‡
Random regr.	2.51(0.05)	2.58(0.05)	2.57(0.05)	2.69(0.05)
Π-regressor	2.05(0.04)	2.30(0.05)	1.99(0.04)	2.33(0.05)

Table 4.13: MAE (Mean Absolute Error) and its standard deviation over participants. MAE is the mean difference between the true and predicted rating (ratings on a scale of 1-9). For comparison, results from the random and Π regressors are also presented. Daggers indicate significance of results compared to the Π-regressor according to a two-sided repeated samples t-test ($^\ddagger = p < .01$), ($^\dagger = p < .05$).

4.4 Conclusions

In this chapter a pilot study was described to automatically recognize emotions induced by watching music video clips. Six participants were asked to watch 20 music videos each and rate them according to perceived levels of valence, arousal and general like/dislike. As they watched the videos, their EEG and peripheral physiological signals were recorded.

We posed the affect recognition problem as a two-class classification problem, classifying the videos as having low or high arousal, valence and liking. For EEG classification, the average (maximum) classification rates are 55.7% (67%) for arousal, 58.8% (76%) for valence and 49.4% (63.5%) for liking ratings. However, the low number of samples makes it difficult to draw any definite conclusions. Nevertheless, the results of the single-trial classification indicated that there is information in the EEG signals regarding users' emotional states. This encouraged expanding the experiment with more participants and trials per participant to improve statistical significance of the results, and a more extensive video clip dataset in order to elicit more diverse emotions.

For the extended experiment, a database for the analysis of spontaneous emotions was collected. The database contains physiological signals of 32 participants (and frontal face video of 22 participants), where each participant watched and rated their emotional response to 40 music videos along the scales of arousal, valence, and dominance, as well as their liking of and familiarity with the videos. We presented a novel semi-automatic stimuli selection method using affective tags. Single-trial classification was performed for the scales of arousal, valence and liking. The results were shown to be significantly better than random classification for the arousal and valence targets. The database collected is made publicly available and is further detailed in Appendix B.

A regression method for continuous emotional characterization of music videos was applied on the feature set. The performance of this single-trial regression was evaluated in a leave-one-trial-out cross-validation strategy for each participant. The continuous emotion detection performance was shown to be significantly superior to random estimation for all rating scales.

This method can be used in the heart of music video recommendation systems to improve viewers' experience. On the basis of the methods described in this chapter, such a real-time affective music video recommendation system was implemented, which is described in detail in Appendix A.

Chapter 5

Facial Expression Recognition

Contents

5.1 Introduction

In this chapter, we discuss our research into automated facial Action Unit (AU) detection. Facial expressions are one of the best modalities for obtaining information about a person's internal state without any active involvement of said person. The AUs we detect here correspond to the smallest discernible facial muscle movements. As AUs are independent of interpretation, they can be used for any high-level decision making process including recognition of basic emotions according to Emotional FACS (EMFACS) rules[1], recognition of various affective states according to the FACS Affect Interpretation Database (FACSAID)[1] introduced by Ekman et al.[40], as well as for recognition of other complex psychological states such as depression[42] or pain[163]. AUs are very suitable to be used in studies on human facial behaviour as thousands of anatomically possible facial expressions (independently of their high-level interpretation) can be described as combinations of 27 basic AUs (see Fig. 2.7) and a number

[1] http://face-and-emotion.com/dataface/general/homepage.jsp

Dataset	MMI	CK	SAL
color	color	grayscale	color
num. sequences used	264	143	77
mean duration(frames)	85.0	19.5	103.1
num. participants	15	65	10
num. Action Units	27	18	10
posed/spontaneous	posed	posed	spontaneous
out-of-plane rotations	minimal	minimal	moderate
temporal segments	all	onset only	all
temporal annotation	yes	no	for 6 subjects only
average MHI F1-score	40.6	Not tested	Not tested
average FFD F1-score	65.1	72.14	75.52
average temporal offset	2.46 frames	-	-

Table 5.1: Comparison of datasets used in this chapter.

of AU descriptors. It is for these reasons that the automatic detection of AUs forms a valuable tool in performing implicit tagging.

In this chapter, we propose a method for AU classification in face videos based on dynamic textures (DT). We combine the use of Free-form Deformations, a non-uniform decomposition of the facial area based on quadtrees for feature extraction, a frame-based GentleBoost classifier and a dynamic, generative HMM model. This is the second DT-based method for AU recognition proposed. We compare our method to the earlier method[153], and show a clear improvement in performance. Unlike most previous approaches, we detect AUs on a frame-by-frame level rather than sequence-based. In addition, unlike most earlier works, we present results on both posed and spontaneous data and investigate the method's generalization performance by training and testing on different datasets.

Table 5.1 summarizes the attributes of the datasets used in this chapter. The MMI dataset was used mainly for it's unparalelled number of posed facial image sequences, allowing us to train and test our methods for all Action Units. The Cohn Kanade dataset was selected mainly due to its predominance in the field, allowing easy comparisons to other methods. Finally, we used the SAL dataset to test our method on more challenging spontaneous expressions.

The outline of this chapter is as follows: Section 5.2 presents the two utilized approaches to modelling dynamics and appearance in the face region of an input video (MHI and FFD) and explains the methodology used to detect AUs and their temporal segments. Section 5.3 describes the utilized datasets, the evaluation study and discusses the results. Section 5.4 concludes the chapter. The work presented

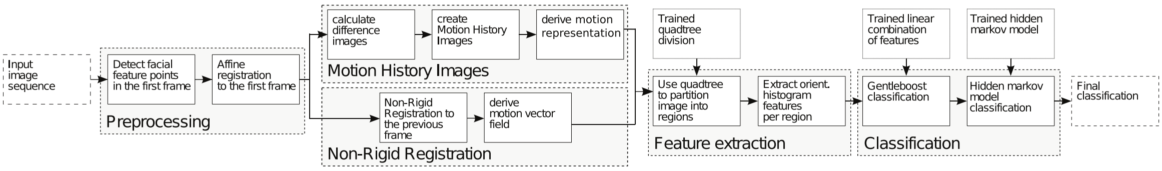

Input image sequence
Detect facial feature points in the first frame
Affine registration to the first frame
Preprocessing
calculate difference images
create Motion History Images
derive motion representation
Motion History Images
Non-Rigid Registration to the previous frame
derive motion vector field
Non-Rigid Registration
Trained quadtree division
Use quadtree to partition image into regions
Extract orient. histogram features per region
Feature extraction
Trained linear combination of features
Trained hidden markov model
Gentleboost classification
Hidden markov model classification
Classification
Final classification

Figure 5.1: Outline of the proposed method.

here has been published in the IEEE Transactions on Pattern Analysis and Machine Intelligence[3].

5.2 Appearance-based Action Unit detection

Fig. 5.1 gives an overview of our system. In the preprocessing phase, the face is located in the first frame of an input video and head motion is suppressed by an affine rigid face registration. Next, non-rigid motion is estimated between consecutive frames by the use of either Non-rigid Registration using Free-form Deformations (FFDs) or Motion History Images (MHIs). For each AU, a quadtree decomposition is defined to identify face regions related to that AU. In these regions, orientation histogram feature descriptors are extracted. Finally, a combined GentleBoost classifier and a Hidden Markov Model (HMM) are used to classify the sequence in terms of AUs and their temporal segments. In the remainder of this section the details of each processing phase are described.

5.2.1 Rigid face registration

In order to locate the face in the first frame of the sequence, we assume the face is expressionless and in a near-frontal position in that frame and use the fully automatic face and facial point detection algorithm proposed in [158]. This algorithm uses an adapted version of the Viola-Jones face detector to locate the face. 20 facial characteristic points and a facial bounding box are detected by using Gabor-feature-based boosted classifiers.

To suppress inter-sequence variations (i.e. facial shape differences) and intra-sequence variations (i.e. rigid head motion), registration techniques are applied to find a displacement field T that registers each frame to a neutral reference frame, while maintaining the facial expression:

$$T = T_{inter} \circ T_{intra}. \tag{5.1}$$

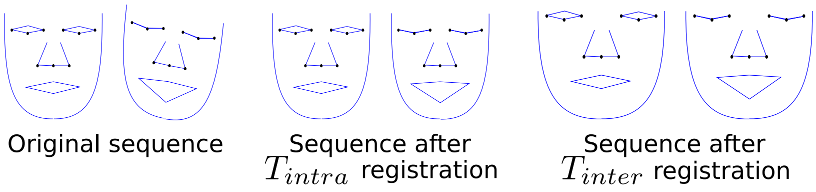

Original sequence
Sequence after T_{intra} registration
Sequence after T_{inter} registration

Figure 5.2: An illustration of the rigid registration process. Also shown are the 10 facial feature points used for registration.

The intra-sequence displacement field T_{intra} is modelled as a simple affine registration. The facial part of each frame in the sequence is registered to the facial part of the first frame to suppress minor head motions. This is done using a gradient descent optimization, with the squared sum of differences (SSD) of the grey level values as a distance metric.

The inter-subject displacement field T_{inter} is again modelled as an affine registration. A subset of 9 of the 20 facial points detected in the first frame that are stable (i.e., their location is mostly unaffected by facial expressions) is registered to a predefined reference set of facial points. This predefined set of reference points is taken from an expressionless image of a subject that was not used in the rest of the experiments. The displacement field T_{inter} is applied to the entire image sequence to eliminate inter-subject differences in facial shape.

The T_{intra} and T_{inter} registrations are performed separately since T_{inter} is a geometric registration of two sets of fiducial facial points, whereas T_{intra} is an appearance-based registration based on the minimization of the sum of squares of the motion-compensated image intensities. Therefore, we can not combine the two registrations. Let us also note here, that intra-sequence transforms (i.e., from a frame to the previous one) are in general smaller and therefore more easily estimated than the combined transform to a global reference frame. However, once estimated, T_{inter} and T_{intra} are combined and applied as a single transformation. An illustration of the two steps and the used facial points is given in Fig. 5.2.

This registration method is quite effective in suppressing rigid face motion in the frontal-face image sequences used in this Chapter. However, it will not solve the general general problem of head pose estimation. Significant out-of-plane head rotations, large and sudden head motions and occlusion of the face area will likely cause this algorithm to fail, and these issues are considered beyond the scope of this work.

5.2.2 Motion representation

Most existing approaches base their classification on either single frames or entire videos. Here, we use overlapping sliding windows of different sizes and classify each window in terms of depicted AUs and their temporal segments. In any given frame, each AU can be in one of four different temporal segments: neutral (inactive), onset, apex, or offset. Different AUs have different onset and offset durations. Therefore it is useful to have a flexible θ (size of temporal window) and consider several sizes. The onset of AU 45(blink), for instance, has an average duration of 2.4 frames (in the utilized datasets). On the other hand, the offset of AU 12 (smile) lasts 15.4 frames on average. A temporal window of 2 frames is well-suited to find the onset of AU 45, but it is hard to detect the onset of AU 12 using such a window. Therefore, several window sizes are tested, ranging from 2 frames to 20 frames. 96.4% of all onsets/offsets in our dataset last 20 frames or less, so this size suffices to easily capture most activations.

To represent the motion in the face due to facial expressions, two different methods of Motion History Images and Non-rigid registration using Free-form Deformations have been investigated, which will now be discussed in detail.

5.2.2.1 Motion history images

Motion history images (MHIs) were first proposed by Davis and Bobick[33]. MHIs compress the motion over a number of frames into a single image. This is done by layering the thresholded differences between consecutive frames one over the other. In doing so, an image is obtained that gives an indication of the motion occurring in the observed time-frame.

Let t be the current frame and let θ be the temporal window size. Then, MHI_t^θ consists of the weighted layered binary difference images for each consecutive two frames $(t-\frac{\theta}{2}, t-\frac{\theta}{2}+1), \ldots, (t+\frac{\theta}{2}-2, t+\frac{\theta}{2}-1)$. A binary difference image for the pair $(t, t+1)$ is denoted with d_t and is defined as

$$d_t(x,y) = b\left(\begin{cases} 1 & |g(x,y,t) - g(x,y,t+1)| > \gamma \\ 0 & otherwise \end{cases}\right), \tag{5.2}$$

where $g(\cdot,\cdot,t)$ is frame t filtered by a Gaussian filter of size 2, γ is a noise threshold set to 4 (this means that two pixels must differ 4 grey levels to be classified as different), b is a binary opening filter applied to the difference image to remove remaining isolated small noise spots with an area smaller than 5 pixels. g was varied between 0 and 10, γ was varied between 1 and 20, b was varied between 0 and 20. The parameters were varied on a small set of videos and the values as used above gave the best results for

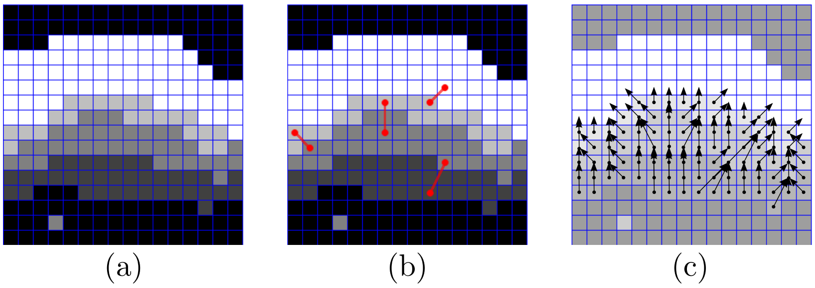


Figure 5.3: Illustration of the estimation of a motion vector field from an MHI. (a): Original MHI. (b): for each pixel, the closest neighbouring brighter pixel is found (without crossing background pixels). (c): This process is repeated for each pixel, resulting in the motion vector field shown here.

recognition.

Using weighted versions of these binary difference images, the MHI is then defined as:

$$M_t^\theta = \frac{1}{\theta}\max_s(\{(s+1)d_{t-\frac{\theta}{2}+s}|0 \leq s \leq \theta - 1\}). \tag{5.3}$$

That is, the value at each pixel of the MHI is the weight of the last difference image in the window that depicts motion, or 0 if the difference images do not show any motion.

In the original implementation by Davis[33], motion vectors are retrieved from the MHI by simply taking the Sobel gradient of the image. This will however only give motion vectors at the borders of each grey level intensity in the image. This works well in the case that the MHIs show smooth and large motion, but in our case the motion is usually shorter and over a smaller distance, leading to less smooth gradients in the image. Applying the Sobel gradient in such a case leads to a very sparse motion representation. The approach taken here is as follows. For each pixel that is not a background pixel (i.e. pixels where M_t^θ is 0 since no motion was detected), we search in its vicinity for the nearest pixel of higher intensity (without crossing through background pixels). The direction in which a brighter pixel lies (if there is one) is the direction of motion in that pixel. In the case that multiple brighter pixels are found at the same distance, the pixel closest to the centre of gravity of those pixels is chosen. This gives us a dense and informative representation of the occurrence and the direction of motion. This is illustrated in Fig. 5.3.

5.2.2.2 Non-rigid registration using free-form deformations

This method is an adapted version of the method proposed by Rueckert et al.[126], which uses a free-form deformation (FFD) model based on b-splines. The method was originally used to register breast MR images, where the breast undergoes local shape changes as a result of breathing and patient motion.

Let Ω_t denote the grey-level image of the face region at frame t, where $\Omega_t(x, y)$ is the grey-level intensity at pixel (x, y). Given a pixel (x, y) in frame t, let $(\hat{x}, \hat{y})$ be the unknown location of its corresponding pixel in frame $t - 1$. Then, the non-rigid registration method is used to estimate a motion vector field $\hat{F}_t$ between frames t and $t - 1$, such that:

$$(\hat{x}, \hat{y}) = (x, y) + \hat{F}_t(x, y) \tag{5.4}$$

To estimate $\hat{F}_t$, we select a $U \times V$ lattice Φ_t of control points with coordinates $\phi_t(u, v)$ in Ω_t, evenly spaced with spacing d. Then, non-rigid registration is used to align Φ_t with Ω_{t-1}, resulting in a displaced lattice $\hat{\Phi}_{t-1} = \Phi_t + \Phi_\delta$. Then, $\hat{F}_t$ can be derived by b-spline interpolation from Φ_δ. To estimate $\hat{\Phi}_{t-1}$, a cost function C is minimized. Rueckert et al.[126] use normalized mutual information as the image alignment criterion. However, in the 2D low-resolution case considered here, not enough sample data is available to make a good estimate of the image probability density function from the joint histograms. Therefore, we use the sum of squared differences (SSD) as the image alignment criterion, i.e. :

$$C(\hat{\Phi}_{t-1}) = \sum_{x,y} (\Omega_t(x, y) - \Omega_{t-1}(\hat{x}, \hat{y}))^2 \tag{5.5}$$

The full algorithm for estimating $\hat{\Phi}_{t-1}$ (and therefore Φ_δ) is given in Fig. 5.4. We can calculate $\hat{F}_t$ using b-spline interpolation on Φ_δ.

For a pixel at location (x, y), let $\phi_t(u, v)$ be the control point with coordinate (x_0, y_0) that is the nearest control point lower and to the left of (x, y), i.e. it satisfies:

$$x_0 \leq x < x_0 + d, \qquad y_0 \leq y < y_0 + d \tag{5.6}$$

In addition, let $\phi_\delta(u, v)$ denote the vector that displaces $\phi_t(u, v)$ to $\hat{\phi}_{t-1}(u, v)$. Then, to derive the displacement for any pixel (x, y), we use a b-spline interpolation between its 16 closest neighbouring control points (see Fig. 5.5). This gives us the estimate of the displacement field $\hat{F}_t$

$$\hat{F}_t(x, y) = \sum_{k=0}^{3} \sum_{l=0}^{3} B_k(a) B_l(b) \phi_\delta(u + k - 1, v + l - 1), \tag{5.7}$$

where $a = x - x_0, b = y - y_0$ and B_n is the n^{th} basis function of the uniform cubic

```
Find the 20 facial points in the first frame of the sequence
Find $T_{inter}$ (affine transformation to reference facial points)
Apply $T_{inter}$ to the entire sequence
foreach frame $t$ do
    Find $T_{intra}$ (affine transformation to frame 1) and apply it
    Initialize the control point lattice $\hat{\Phi}^0_{t-1}$ as $\Phi^0_t$
    foreach control point density $d$ do
        Calculate the gradient vector of the cost function $C$
        in terms of $\hat{\Phi}^d_{t-1}$: $\nabla C = \frac{\delta C(\hat{\Phi}^d_{t-1})}{\delta \hat{\Phi}^d_{t-1}}$
        while $||\nabla C|| > \varepsilon$ do
            Recalculate the control point positions:
            $\hat{\Phi}^d_{t-1} = \hat{\Phi}^d_{t-1} + \mu \frac{\nabla C}{||\nabla C||}$
            Recalculate $\nabla C$
        end
        Increase the density $d$ of the control point lattice
        Add points to $\hat{\Phi}^{d+1}_{t-1}$ from $\hat{\Phi}^d_{t-1}$ by b-spline interpolation
    end
    Derive $\Phi_\delta$: $\Phi_\delta = \hat{\Phi}_{t-1} - \Phi_t$
    Use b-spline interpolation to derive $\hat{F}_t$ from $\Phi_\delta$
end
```

Figure 5.4: The non-rigid registration algorithm. ε is a stopping criterion and μ is the step size in the recalculation of control point positions. The values for both are taken from [126].

b-spline, i.e.:

$$\begin{aligned}
B_0(a) &= (-a^3 + 3a^2 - 3a + 1)/6, \\
B_1(a) &= (3a^3 + 6a^2 + 4)/6, \\
B_2(a) &= (-3a^3 + 3a^2 + 3a + 1)/6, \\
B_3(a) &= a^3/6.
\end{aligned}$$

To speed up the process, and avoid local minima, we use a hierarchical approach in which the lattice density is being doubled at every level in the hierarchy. The coarsest lattice Φ^0_t is placed around the point $c = (c_x, c_y)$ at the intersection of the horizontal line that connects the inner eye corners, and the vertical line passing through the tip of the nose and the centre of the upper the and bottom lip. Then,

$$\Phi^0_t = \left\{ (u, v) \;\middle|\; \begin{array}{l} u \in [c_x - 2id, \ldots, c_x + 2id], \\ v \in [c_y - 2id, \ldots, c_y + 4id] \end{array} \right\} \tag{5.8}$$

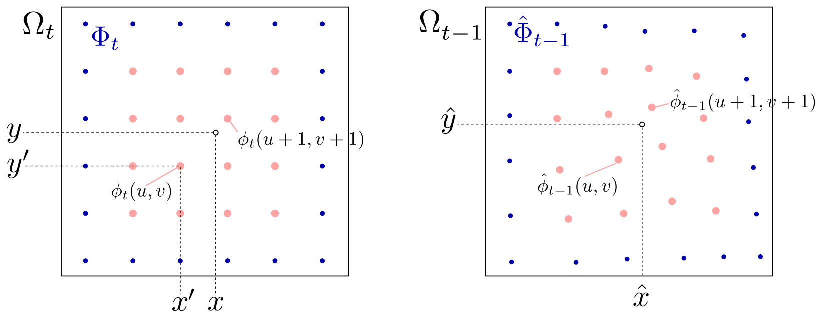
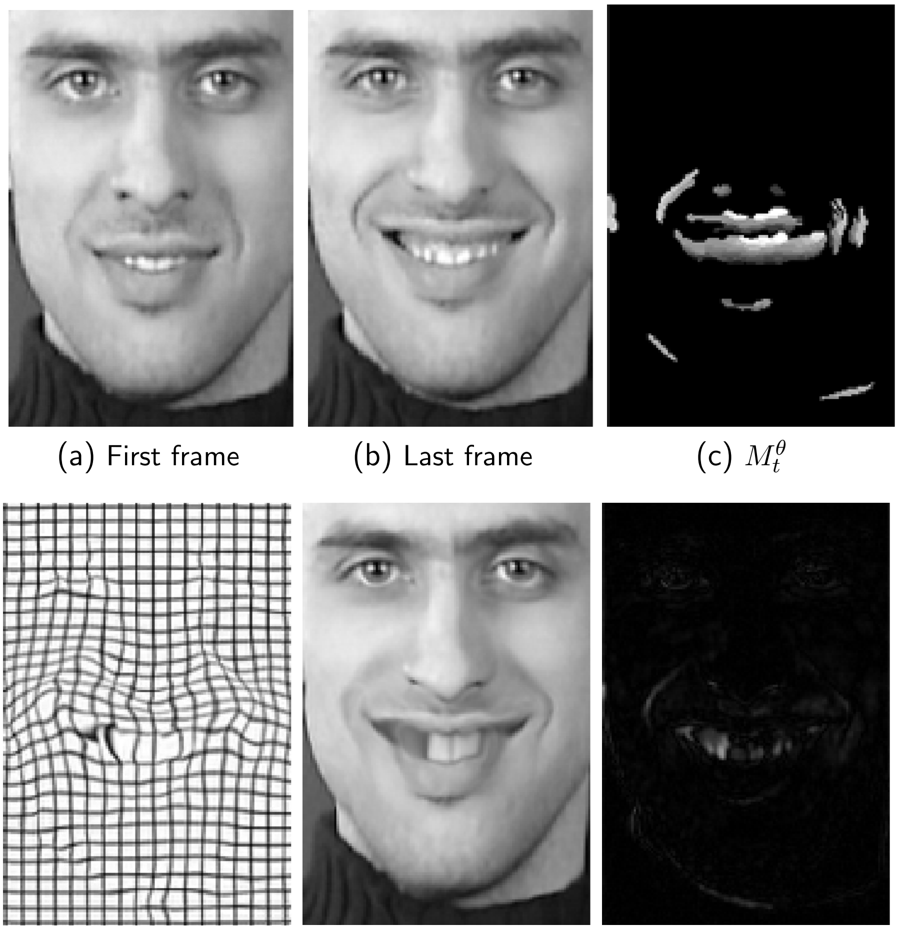


Figure 5.5: Illustration of the B-spline interpolation showing an image Ω_t and the control point lattice Φ_t, as well as the estimated $\hat{\Phi}_{t-1}$ aligned with Ω_{t-1}. To estimate the new position $(\hat{x},\hat{y})$ of the point at (x,y), only the 16 control points shown in a lighter, red colour are used.

where id is the distance between the eye pupils (i.e. Φ_t^0 consists of 35 control points). New control points are iteratively added in between, until the spacing becomes $0.25id$ (approximately the size of a pupil), giving 1617 control points. This has proven sufficient to capture most movements and gives a good balance between accuracy and calculation speed.

Having estimated $\hat{F}_t$, we now have a motion vector field depicting the facial motion between frame $t-1$ and t, from which orientation histogram features can be extracted. For feature extraction, we actually consider the motion vector field sequence $\hat{F}_t^\theta$ over a sliding window of size θ around frame t.

Fig. 5.6 shows an example of the MHI and FFD methods. Fig. 5.6(a) and 5.6(b) show the first and last frame of the sequence. Fig. 5.6(c) shows the resulting MHI M_t^θ, where θ is set such as to include the entire sequence. It is quite easy for humans to recognize the face motion from the MHI. Fig. 5.6(d) shows the motion field sequence $\hat{F}_t^\theta$ from the FFD method applied to a rectangular grid. The face motion (Fig. 5.6(f)) is less clear to the human eye from this visualization of the transform. However, when we transform the first frame by applying $\hat{F}_t^\theta$ to get an estimate of the last frame, the similarity is clear as shown in Fig. 5.6(e). In addition, one can see that between Fig. 5.6(a) and 5.6(b), the subject shows a slight squinting of the eyes (AU6). While this is invisible in the resulting MHI (Fig. 5.6(c)), it is visible in the motion field derived from FFD (Fig. 5.6(d)), indicating that the FFD method is more sensitive to subtle motions than the MHI method.



(a) First frame (b) Last frame (c) M_t^θ

(d) $\hat{F}_t^\theta$ applied to a grid (e) $\hat{F}_t^\theta$ applied to first frame (f) Difference between (b),(e)

Figure 5.6: Example of MHI and FFD techniques.

5.2.3 Feature extraction

5.2.3.1 Quadtree decomposition

In order to define the face sub-regions at which features will be extracted, we use a quadtree decomposition. Instead of dividing the face region into a uniform grid (e.g. as in [173]) or manually partitioning the face, a quadtree decomposition is used to divide the regions in a such a manner that areas showing much motion during the activation of a specific AU are divided in a large number of smaller sub-regions, while those showing little motion are divided into a small number of large sub-regions. This results in an efficient allocation of the features. We note that different features (i.e. different quadtree decompositions) are used for the analysis of different AUs.

Some AUs are very similar in appearance but differ greatly in the temporal domain. For instance, AU 43 (closed eyes) looks exactly like AU 45 (blink) but lasts significantly longer. Therefore, we also use a number of temporal regions to extract features. Let

$\Theta_{a,s}$ be the collection of all sliding windows of size θ around the frames depicting a particular AU a in a particular temporal segment s in the training set. We then use a quadtree decomposition specific to each AU and the segments onset and offset on a set of projections of $\Theta_{a,s}$ to decide where to extract features to recognize the target AU and its target temporal segment.

Three projections of each window are made showing the motion magnitude, the motion over time in the horizontal direction, and the motion over time in the vertical direction:

$$P^{\theta}_{mag}(x,y) = \sum_t u(x,y,t)^2 + v(x,y,t)^2, \tag{5.9}$$

$$P^{\theta}_{tx}(t,x) = \sum_y u(x,y,t)^2, \tag{5.10}$$

$$P^{\theta}_{ty}(t,y) = \sum_x v(x,y,t)^2 \tag{5.11}$$

where $u(x,y,t)$ and $v(x,y,t)$ are the horizontal and vertical components of the motion vector field sequence $\hat{F}^{\theta}_t$. These projections are then summed over all windows in $\Theta_{a,s}$ to get the final projections used for the quadtree decomposition:

$$P^{\Theta_{a,s}}_{mag}(x,y) = \sum_{\theta\in\Theta_{a,s}} P^{\theta}_{mag}(x,y), \tag{5.12}$$

$$P^{\Theta_{a,s}}_{tx}(t,x) = \sum_{\theta\in\Theta_{a,s}} P^{\theta}_{tx}(t,x), \tag{5.13}$$

$$P^{\Theta_{a,s}}_{ty}(t,y) = \sum_{\theta\in\Theta_{a,s}} P^{\theta}_{ty}(t,y) \tag{5.14}$$

These three images then undergo a quadtree decomposition to determine a set of 2D regions ((x,y)-, (t,x)-, and (t,y)- regions) where features will be extracted. The defined projections show us exactly where much motion occurs for a particular AU and a particular temporal segment and where there is less motion. The quadtree decomposition algorithm is outlined in Fig. 5.7. The splitting threshold τ was set to 0.1, meaning a region in the quadtree will be split if the region accounts for 10% of the total motion in the frame. This gives a reasonable balance between having too large regions, so the detail is lost, and too many small regions, where the features become less effective as facial features do no longer always fall in the same region. The minimum region size σ is defined to be $0.25id$, where id is the interocular distance. In other words, the minimum region size is about the size of a pupil. Extracting features in smaller regions will not be very informative due to small variations in facial feature locations in different subjects. Some examples of motion magnitude images and the

```
Initialize $R$ with a single region (the entire face region)
Define $p_{total}$ as the summed value of all pixels in $P$
Define $\tau$ as the splitting threshold
Define $\sigma$ as the minimum size of a region
while True do
    foreach region $r$ in $R$ do
        Calculate $p_r$, the summed value of all pixels in $r$
        if $p_r < \tau \cdot p_{total}$ and $size(r) > \sigma$ then
            Remove $r$ from $R$
            Split $r$ in 4 equally sized rectangles
            Add these to $R$
        end
    end
    if no region was split then stop
end
```

Figure 5.7: The quadtree decomposition algorithm. τ is the threshold for splitting, σ is the minimum region size.

resulting quadtree decompositions are shown in Fig. 5.8. We can see in Fig. 5.8(e) that for AU46R (right eye wink) most of the features will be extracted in the eye area, where all the motion occurs.

In $\Theta_{a,s}$, some frames also show the activation of other AUs than a. Usually, the activation of other AUs does not occur frequently enough to significantly alter the decomposition. However, in some cases, AUs co-occur very frequently and the decomposition shows some of the motion of the co-occurring AU. It may then happen that some features corresponding to the co-occurring AU are then selected to classify a.

5.2.3.2 Motion features

After generating the quadtree decompositions, we extract the features for the sliding window around each frame in the dataset. We consider the $u(x, y, t)$ and $v(x, y, t)$ components from $\hat{F}_t^\theta$ in the sub-regions determined by the quadtree decomposition of $P_{mag}^{\Theta_{a,s}}(x, y)$. In each sub-region 11 features are extracted from the components: an orientation histogram of 8 directions, the divergence, the curl, and the motion magnitude.

For the temporal regions determined by the decompositions of $P_{tx}^{\Theta_{a,s}}(t, x)$ and $P_{ty}^{\Theta_{a,s}}(t, y)$, we first determine the projections $P_{tx}^\theta(t, x)$ and $P_{ty}^\theta(t, y)$ for the test frame in question. For each sub-region in the projections, we extract 3 features: the average absolute motion, the average amount of positive (i.e. left, upward) motion and the

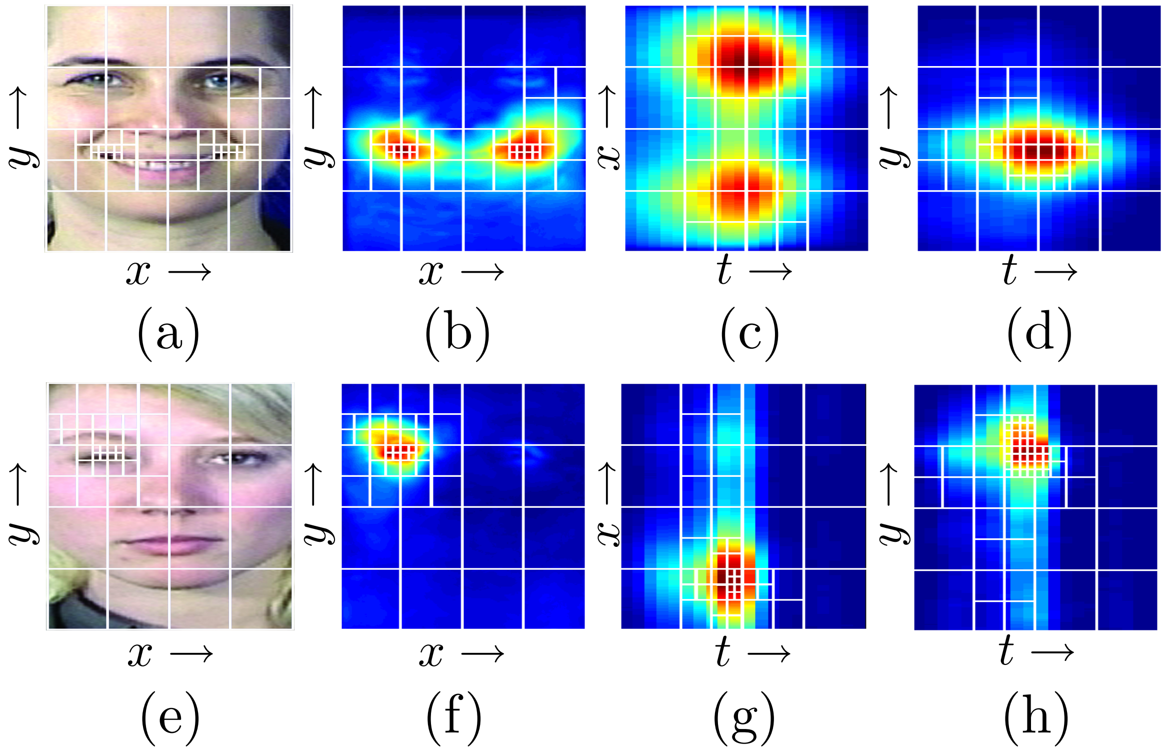


Figure 5.8: Quadtree decompositions: (a,b,c,d) Onset of AU 12(smile); (e,f,g,h) Onset of AU 46R(right eye wink). Shown for each AU are example frames (a,e) and the three projections $P_{mag}^{\Theta_{a,s}}$ (b,f), $P_{tx}^{\Theta_{a,s}}$ (c,g), $P_{ty}^{\Theta_{a,s}}$ (d,h). Overlaid on each projection is the resulting quadtree decomposition.

average amount of negative (i.e. right, downward) motion.

5.2.4 Action Unit classification

5.2.4.1 Gentleboost

We use the GentleBoost algorithm[49] for feature selection and classification. Advantages of GentleBoost over AdaBoost are that it converges faster and is more reliable when stability is an issue[49]. For each AU and each temporal segment characterised by motion (i.e. onset, offset), we train a dedicated one-vs-all GentleBoost classifier. Since our dataset is rather unbalanced (over 95% of the frames in the database depict expressionless faces), we initialize the weights such that both the positive and the negative classes carry equal weight. This prevents that all frames are classified as neutral. The GentleBoost algorithm is used to select a linear combination of features one at a time until the classification no longer improves by adding more features. This gives a reasonable balance between speed and complexity. The number of features selected for each classifier range between 19 and 93, with an average of 74 features

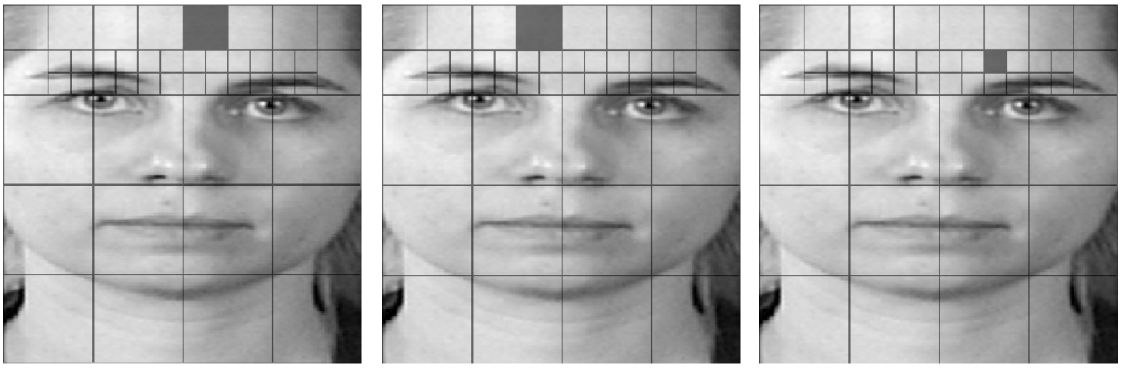
AU	**1**	**5**	**9**	**12**	**16**	**24**	**27**
onset original	2013	2013	1551	1551	1551	1650	1386
onset selected	67	67	34	47	19	87	12
offset original	1452	1815	1551	1683	1551	1749	1683
offset selected	90	85	76	86	34	86	73

Table 5.2: Original number of features and number of features selected by GentleBoost per AU when trained on the entire MMI dataset with a window-size of 20 frames.

selected. Table 5.2 gives an overview of the number of selected features for several AUs.

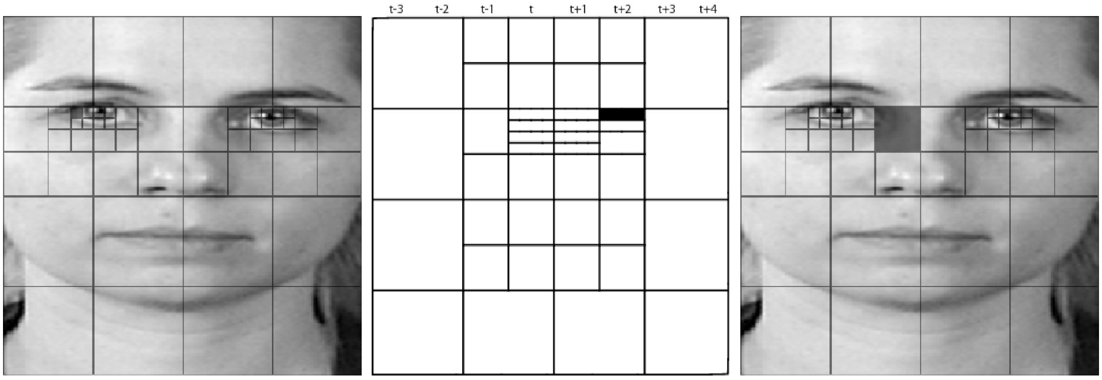
The first three selected features for some of the classifiers are shown in Figures 5.9-5.10. In the images, for each feature selected from the $P_{mag}^{\Theta_{a,s}}$-projection, a neutral face image is overlaid to indicate the location of the region. The selected features correspond reasonably well to the intuitively interesting features/regions for each AU. The $P_{mag}^{\Theta_{a,s}}$-projection is the most important (and most often selected) projection since most information is available in the spatial domain. This is also the reason why the problem of facial expression recognition can be solved (to a certain extent) using static images (e.g. [111]). However, for some AUs, the information in the spatial magnitude projection is insufficient to distinguish them from other AUs. One example is AU 43 (closed eyes), which only differs from AU 45 (blink) in the temporal domain. Since AU 45 is much more common, an AU 43 detector that does not take the temporal domain into account would detect many false positives. Fig. 5.10 shows that a temporal feature is the second most important one in the detection of the onset of AU 43. The feature in question measures the amount of upward motion in the eyelid area for the next 2 frames. If the depicted AU were AU 45, then the next 2 frames after any of the onset frames should show upward motion as the eye would be be opening again. In AU 43 however, the next 2 frames after any of the onset frames will show no motion as the eyes will still be closed. Thus, the absence of upward motion in this area in a period of 2 frames after an onset frame is a very good way to tell apart AU 43 from AU 45 onset segments.

Each onset/offset GentleBoost classifier returns a single number per frame indicating the confidence that that frame depicts the target AU and the target temporal segment.



(a) $P_{mag}^{\Theta_{1,8}}$: divergence (b) $P_{mag}^{\Theta_{1,8}}$: divergence (c) $P_{mag}^{\Theta_{1,8}}$: divergence

Figure 5.9: First three selected features for onset of AU 1 (inner brow raiser), window size 8, superimposed on a neutral frame.



(a) $P_{mag}^{\Theta_{43,8}}$: divergence (b) $P_{ty}^{\Theta_{43,8}}$: no upward motion (c) $P_{mag}^{\Theta_{43,8}}$: divergence

Figure 5.10: First three selected features for onset of AU 43 (closed eyes), window size 8, superimposed on a neutral frame. (b) depicts the absence of upwards motion in shown y-area of frame $t + 2$.

5.2.4.2 Hidden Markov Models

In order to combine the onset/offset GentleBoost classifiers into one AU recognizer, a continuous HMM is used. The motivation for using an HMM is to use the knowledge that we can derive from our training set about the prior probabilities of each temporal segment of an AU and its duration (represented in the HMM's transition matrix). Hence, an HMM is trained for the classification of each AU.

HMMs are defined by $\lambda = \{\Lambda, B, \Pi\}$, where Λ is the transition matrix, B is the emission matrix and Π is the initial state probability distribution. These are all estimated from the training set, where the outputs of the onset- and offset-GentleBoost classifiers are used to calculate the emission matrix B for the HMM by fitting a Gaussian to the values of both outputs in any temporal state. Then, the probability for each state can be calculated given the output of the GentleBoost classifiers in a particular frame.

The HMM has four states, one corresponding to each of the temporal segments. The initial probabilities Π show that the sequences in our dataset usually start in the neutral segment (i.e. no AU is depicted), but on rare occasions the AU is already in one of the other states. Based on the initial probabilities Π, the transition probabilities Λ and emission probability matrix B, the HMM decides the mostly likely path through the temporal segment states for the input image sequence, using the standard Viterbi algorithm. This results in the classification of the temporal segment for each frame in the tested image sequence.

The HMM facilitates a degree of temporal filtering. For instance, given that the input data temporal resolution is 25 fps and given the facial anatomy rules, it is practically impossible to have an apex followed by a neutral phase and this is reflected in the transition probabilities Λ. Also, the HMM tends to smooth out the results of the GentleBoost classifiers (for instance, short incorrect detections are usually filtered out). However, it only captures the temporal dynamics to a limited degree, since it operates under the Markov assumption that a signal value at time t is only dependent on the signal value at time $t-1$. For example, the HMM does not explicitly prevent onsets that last only one frame (even though in most AUs, the minimum onset duration is much longer). Yet it does model these dynamics implicitly through its use of transition probabilities between the states.

An example of the learned transition probabilities Λ for one HMM, trained to recognize AU 1, is given in Fig. 5.11. The transition probabilities say something about the state duration. For instance, the transition probability for *neutral* → *neutral* is very high, since the duration of a neutral state is usually very long (it is as long as the video itself when the video does not contain the target AU). The normal sequence of states is *neutral* → *onset* → *apex* → *offset* → *neutral*. However, the transition probabilities show that, although highly unlikely, transitions *apex* → *onset* or *offset* → *apex* do occur. This is typical for spontaneously displayed facial expressions which are characterized by multiple apexes[42, 109]. As both utilized datasets, the MMI and the Cohn-Kanade dataset, contain recordings of acted (rather than spontaneously displayed) facial expressions, occurrence of multiple apexes is rare and unlikely. In the SAL spontaneous expression dataset on the other hand, multiple apexes occur quite frequently. However, especially in the MMI dataset and especially by brow actions (AU1, AU2), smiles (AU12), and parting of the lips (AU25), some recordings seem to be capturing spontaneous (unconsciously displayed) rather than purely acted expressions.

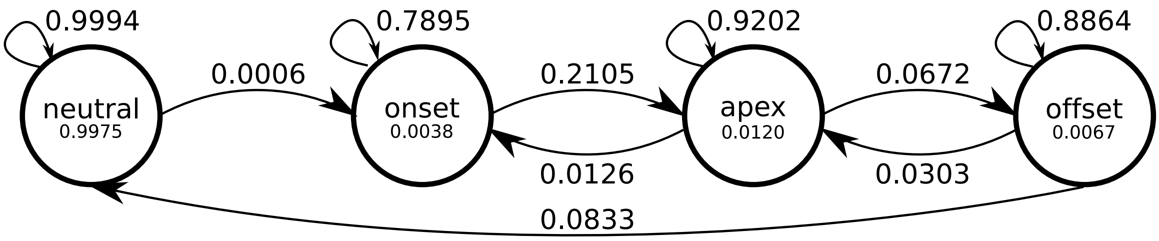


Figure 5.11: The states and transition probabilities for an HMM trained on AU 1. Initial probabilities are denoted below the state names. Transitions with probability 0 are not shown.

5.3 Experiments

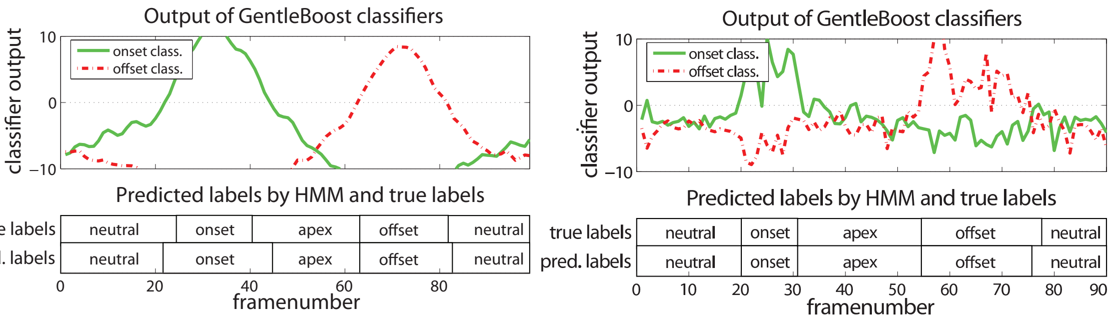
5.3.1 Datasets

5.3.1.1 Posed facial expressions

The first dataset consists of 264 image sequences taken from the MMI facial expression database[112][2]. To the best of our knowledge, this data is the largest freely available dataset of facial behaviour recordings. Each image sequence used in this study depicts a (near-)frontal view of a face showing one or more AUs. The image sequences are chosen such that all AUs under consideration are present in at least ten of the sequences and distributed over 15 subjects. The image sequences last on average 3.4 seconds and were all manually coded for the presence of AUs. Ten-fold cross-validation was used, with the folds divided such that each fold contains at least one example of each AU. Temporal window sizes ranging from 4 to 20 frames were all tested independently and the window size that yielded the best result was chosen.

To test the generalization performance of the system, we have also evaluated the proposed FFD-based method on the Cohn-Kanade (CK) dataset[69], arguably the most widely used dataset in the field. We only tested the system on those AUs for which more than ten examples existed in the CK dataset. This resulted in examples of 18 AUs shown in 143 sequences in total. The original CK dataset only has event coding for the AUs (stating only whether an AU occurs in the sequence, not a frame-by-frame temporal segment coding). Here, we have used frame-by-frame annotations provided by Valstar& Pantic[151] based on the given event coding.

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



(a) AU 27, $\theta = 20$

(b) AU 27, $\theta = 2$

Figure 5.12: Example classification results. Top: The output of the GentleBoost-classifiers. Bottom: The true and estimated frame labels (as predicted by the HMM). θ is the used temporal window size.

5.3.1.2 Spontaneous facial expressions

We also tested the method on the SAL (Sensitive Artificial Listener) dataset containing displays of spontaneous expressions[38]. The expressions were elicited in human-computer conversations through a 'Sensitive Artificial Listener' interface. Subjects converse with one of four avatars, each having its own personality. The idea is for subjects to unintentionally and spontaneously mirror the emotional states of the avatars. 10 subjects were recorded for around 20 minutes each. The speech sections were removed from the data, leaving 77 sequences that depict spontaneous facial expressions. For 4 subjects, the data has been FACS-coded on a frame-by-frame basis, for the other 6 subjects only event coding exists. Since our method requires frame-by-frame annotations to train the classifiers, we used data of 4 subjects for training and we tested on the remaining 6 subjects. We only tested our method on the 10 AUs for which there were at least 5 training examples.

5.3.2 Results

Fig. 5.12 shows two typical results for AU 27 (mouth stretch). As can be seen in Fig. 5.12(a), the GentleBoost classifiers yield good results and the resulting labelling is almost perfect for $\theta = 20$. For $\theta = 2$, the GentleBoost classifiers yield less smooth results (Fig. 5.12(b)). Even so, the HMM filters out the jitter very effectively.

5.3.2.1 Event coding

Table 5.3 gives the results for all AUs tested with the MHI and the FFD technique on the MMI dataset (per AU, the window width θ that gave the highest F1-score is mentioned). The F1-score is a weighted mean of the precision and recall measures.

	Results FFD method					**Results MHI method**				
AU	θ	**CR**	**RC**	**PR**	**F1**	θ	**CR**	**RC**	**PR**	**F1**
1	20	97.7	61.5	88.9	**72.7**	20	93.9	53.9	41.2	**46.7**
2	20	97.7	66.7	80.0	**72.7**	20	96.2	50.0	60.0	**54.6**
4	20	91.3	74.3	65.0	**69.3**	20	76.1	91.2	34.1	**49.6**
5	20	93.6	66.7	38.1	**48.5**	12	93.6	27.3	25.0	**26.1**
6	20	96.2	82.4	66.7	**73.7**	20	93.6	76.9	41.7	**54.1**
7	8	92.1	54.6	27.3	**36.4**	8	86.0	45.5	13.9	**21.3**
9	20	97.0	81.8	60.0	**69.2**	20	93.6	70.0	33.3	**45.2**
10	20	97.4	78.6	73.3	**75.9**	20	95.8	42.9	66.7	**52.2**
11	12	94.7	77.8	58.3	**66.7**	16	89.0	33.3	26.1	**29.3**
12	20	93.6	82.4	50.0	**62.2**	20	80.3	100	24.6	**39.5**
13	12	95.5	90.0	45.0	**60.0**	12	87.9	20.0	7.7	**11.1**
14	16	91.3	75.0	38.7	**51.1**	16	91.7	68.8	39.3	**50.0**
15	8	94.7	75.0	45.0	**56.3**	20	95.1	25.0	42.9	**31.6**
16	16	97.0	85.7	66.7	**75.0**	16	95.8	57.1	61.5	**59.3**
17	16	83.7	75.3	77.8	**76.5**	20	74.2	86.7	58.2	**69.6**
18	16	91.7	63.6	50.0	**56.0**	12	84.9	47.8	28.2	**35.5**
20	20	95.1	45.5	41.7	**43.5**	20	91.3	36.4	20.0	**25.8**
22	12	93.2	72.7	34.8	**47.1**	20	94.3	36.4	33.3	**34.8**
23	16	92.4	58.3	31.8	**41.2**	16	91.3	8.3	7.7	**8.0**
24	16	89.4	61.1	34.4	**44.0**	16	89.0	20.0	15.0	**17.1**
25	8	90.5	92.0	78.4	**84.7**	20	71.6	86.7	50.0	**63.4**
26	20	95.5	81.8	81.8	**81.8**	20	82.2	61.3	35.2	**44.7**
27	20	99.6	100	92.9	**96.3**	20	95.8	100	54.2	**70.3**
28	16	93.6	92.9	44.8	**60.5**	20	88.6	42.9	21.4	**28.6**
28B	16	95.5	72.7	47.1	**57.1**	12	92.8	36.4	25.0	**29.6**
28T	12	92.4	80.0	30.8	**44.4**	16	84.5	50.0	12.2	**19.6**
43	20	95.1	60.0	56.3	**58.1**	20	86.4	20.0	11.1	**14.3**
45	8	93.6	90.8	93.4	**92.1**	4	85.6	96.3	75.4	**84.6**
46L	8	99.2	90.9	90.9	**90.9**	8	97.0	54.6	66.7	**60.0**
46R	8	99.2	81.8	100	**90.0**	12	97.0	27.3	100	**42.9**
avg	-	94.3	75.7	59.7	**65.1**	-	89.2	52.4	37.7	**40.6**

θ = Window Size, CR = Classification Rate
RC = Recall Rate, PR = Precision Rate, F1 = F1-score

Table 5.3: Results for 27 AUs (30 classes) on 264 sequences from the MMI dataset for the MHI and the FFD method.

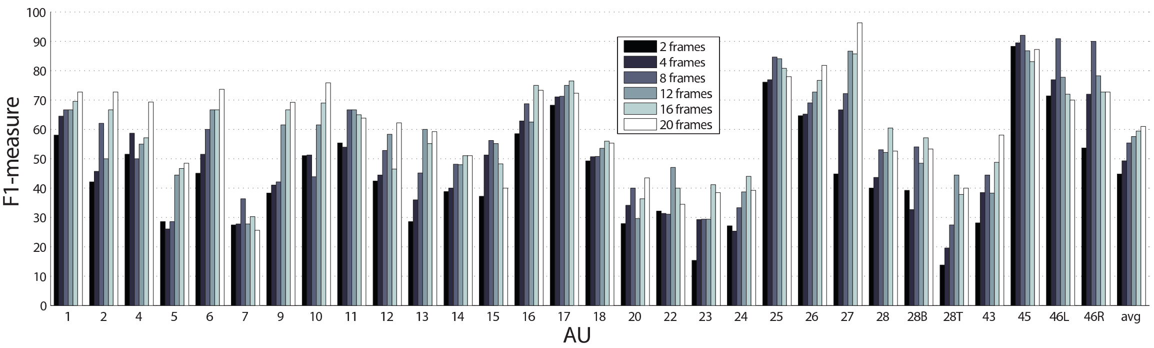


Figure 5.13: F1-score per AU for different window sizes for the FFD method.

In the manual labelling of the dataset, AU 46 (wink) has been split up into 46L and 46R, since the appearance differs greatly depending on which eye is used to wink. Similarly, AU 28 (lip suck) is scored when both lips are sucked into the mouth, and AU 28B and AU 28T are scored when only the lower or the upper lip is sucked in. This gives us a total of 30 classes, based on the 27 AUs defined in FACS. As can be seen from Table 5.3, both techniques have difficulties with subtle AUs (i.e. 5 (upper lid raiser), 7 (eye squint), 23 (lip tightener)). These problems possibly stem from the method of extracting motion statistics over larger regions. If the regions are too large, these subtleties are easily lost (however, having the regions too small generates errors relating to the rigid registration and inter-subject differences). Possibly, geometric approaches are better equipped to handle these AUs (e.g. AU5, AU7), since their activation is clearly observable from displacements of facial fiducial points and no averaging of the motion over regions is needed.

It is clear that, overall, the FFD technique produces superior results to those obtained for the MHI-based approach. Therefore, in the remainder of this work, only the FFD-based approach is investigated further. One reason for the inferior performance of the MHI-based approach is that only intensity differences above the noise threshold are registered in the MHI. For instance, if the mouth corner moves (e.g. in AU12), only the movement of the corner of the mouth is registered in the related MHI. More subtle and smoother motion of the skin (e.g., on the cheeks) is not registered in the related MHI (see Fig. 5.6). In the FFD method however, we will see the entire cheek deform as a result. Also, in MHIs earlier movements can obscure later movements (e.g. in AU 28) and fast movements can show up as disconnected regions that do not produce motion vectors (e.g. in AU 27).

In general, the F1-score is reasonably high for most AUs when the FFD technique is applied, but there is still room for improvement. In particular, there are many false positives. Most of these occur in AUs that have a similar appearance. The AUs performing below 50% are AUs 5 (upper lid raiser), 7 (eye squint), 20 (lip stretcher), 22 (lip funneller), 23 (lip tightener) and 28T (upper lip inward suck). For most of these AUs, the reasons for the inaccurate performance lie in the confusion of the target AU with other AUs. For instance, the onset of AU7 (eye squint) is often confused with the onset of AU45 (blink), the offset of AU5 is very similar to the onset of AU45 (and vice versa), and AUs 20, 23, 24 and 28T are often confused with each other since they all involve downward movement of the upper lip.

Another cause of some false positives is a failure of the affine registration meant to stabilize the face throughout the sequence. Out-of-image-plane head motions, for instance, if not handled well, result in some classifiers classifying rigid face motions as non-rigid AU activations. We partially address this issue for spontaneous expressions in Section 5.3.2.3 by incorporating the results of a facial point tracker in the rigid registration process. However, we should note that for very large out-of-plane rotations, affine registration is not sufficient. The use of 3D models seems a promising direction. However, they require the construction of a 3D model that might be difficult to obtain from monocular image sequences.

Though most AUs perform best with the largest window size tested, it is clear from the results that AUs with shorter durations such as AU 45 benefit from a smaller window size.

Fig. 5.13 shows the results for all AU classifiers for all tested window widths for the FFD technique. Overall, we see that the F1-score improves as the temporal window increases. Exceptions include AUs with particularly short durations, such as 7 (eye squint), 45 (blink), 46L (left eye wink), and 46R (right eye wink).

5.3.2.2 Temporal analysis

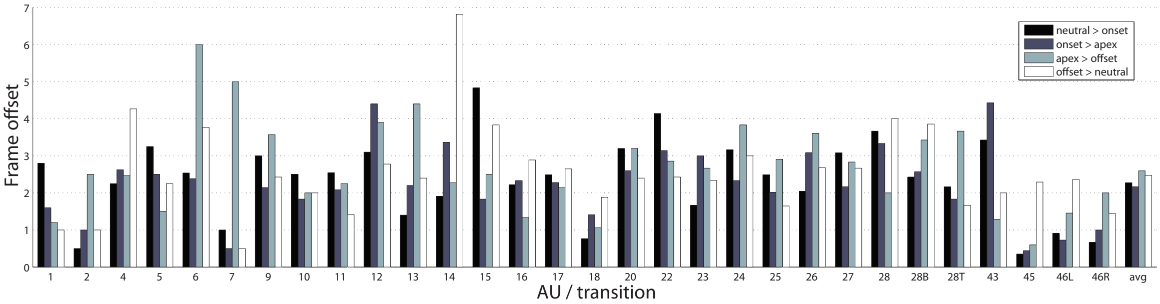


Figure 5.14: Average detection offsets per AU and temporal segment transition.

We were also interested in the timing of the temporal segment detections with respect to the timing delimited by the ground truth. This test was run using the optimal window widths as summarized in Table 5.3. Only sequences that were correctly classified in terms of AUs were considered in this test. Four different temporal segment transitions can be detected, *neutral* → *onset*, *onset* → *apex*, *apex* → *offset*, and *offset* → *neutral*. Fig. 5.14 shows the average absolute frame deviations per AU and temporal segment transition. The overall average deviation is 2.46 frames. 44.12% of the detections are early and 38.18% are late. The most likely cause of late detection is that most AUs start and end in a very subtle manner, visible to the human eye but not sufficiently pronounced to be detected by the system. Early detections usually occur when a larger temporal window width is used, where the AU's segment in question is already visible in the later frames of the window, but it is not actually occurring at the frame under consideration (this can also be seen in Fig. 5.12a). In general, AUs of shorter duration also show smaller deviations. Also, the transitions that score badly are usually subtle ones. The high deviations for *apex* → *offset* in AUs 6 (cheek raiser and lid compressor) and 7 (eye squint) can be explained by considering that these transitions are first only slightly visible in the higher cheek region before becoming apparent in the motion of the eyelids. Since the eyelid motion is much clearer, our method targets that motion and misses the cheek raising in the start of the transition. Similarly, the *offset* → *neutral* transition in AU 14 (mouth corner dimpler) has almost all of the motion in the first few frames and then continues very slowly and subtly. Our method picks up only the first few frames of this transition.

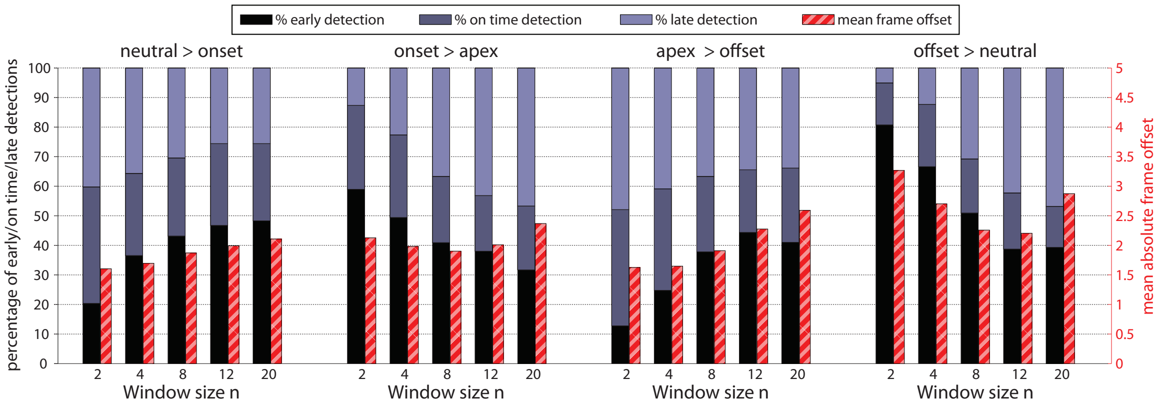


Figure 5.15: Percentages of early/on time/late detection per transition and window size. Also shows average frame offset.

Another way to look at the temporal analysis results is to analyse them per window size and transition type. Fig. 5.15 illustrates that. It shows the proportion of early, timely, and late detections for all correctly detected transitions per window size. It also shows the mean absolute frame offset per transition and per window size (this is depicted by the narrow bar, placed on the right side of each of the main bars in the graph). Interestingly, for the *neutral* → *onset* and *apex* → *offset* transitions the most accurate results are obtained for the lowest window size and the results deteriorate as the window size increases. For the other two transitions, the lower window sizes are actually less accurate and the best results are obtained at window sizes 8 and 12. This behaviour might be explained by a few factors. Firstly, most motion occurs in the beginning of the onset and offset segments, with the endings of those segments containing slower, more subtle motions. Hence, the transitions indicating the end of motion (*onset* → *apex* and *offset* → *neutral*) are detected early since the subtle motion at the end of the onset and offset segments remains undetected by the system. The transitions indicating the start of motion (*neutral* → *onset* and *apex* → *offset*) are quite unlikely to be early, simply because there is no prior motion which could be classified as the transition in question. The results change as the window size increases. This is due to the smoothing effect discussed earlier, due to which the start of motion is detected earlier and the end of motion is detected later.

5.3.2.3 Spontaneous expressions

We performed tests on the SAL dataset, containing 77 sequences of spontaneous expressions, mostly smiles and related expressions. We tested for the 10 AUs that occurred 5 or more times. We trained on the sequences of 4 of the 10 subjects, that were annotated frame-by-frame for AUs, and tested on the data of the other 6 subjects, that were annotated per sequence.

The dataset contains relatively large head motions and moderate out-of-plane rotations. We note that in the datasets used in this paper all facial fiducial points were visible at all times. If that is not the case, one could train a different set of classifiers for each facial viewpoint.

The results for the SAL dataset are given in Table 5.4. The obtained classification rate is 80.2%, which is lower than the results on the posed data sets (89.8% on CK and 94.3% on MMI). However, we achieve a satisfactory average F1-score of 75.5%, which is in fact higher than for the MMI (65.1%) and CK (72.1%) datasets. The worst performance is reported for AUs 2, 7, and 10. AUs 2 and 10 are much exaggerated in posed expressions and therefore harder to detect in subtle spontaneous depictions. AU 7 is here also often confused with AU 45, just as in the MMI dataset. The best performing AUs are 12, 25, and 6. In fact, these AUs perform much better than in the MMI dataset. This can be explained by the fact that many more training samples were available here, indicating that more training examples can greatly benefit the performance. In addition, these AUs also occur more frequently in the test set than in the MMI case, making the test set less unbalanced compared to the other datasets.

AU	θ	NT	CR	RC	PR	F1
1	12	8	92.86	60.00	75.00	66.67
2	20	10	88.10	57.14	66.67	61.54
6	4	28	85.71	85.71	96.77	90.91
7	4	7	57.14	42.86	60.00	50.00
10	8	13	66.67	80.00	52.17	63.16
12	2	35	95.24	94.87	100.00	97.37
23	12	6	83.33	91.67	64.71	75.86
25	2	33	92.86	92.86	100.00	96.30
26	4	18	76.19	76.32	96.67	85.29
45	16	17	64.29	53.33	94.12	68.09
avg	-	-	80.24	73.48	80.61	75.52

θ = Window Size, NT = No. of training examples
CR = Classification Rate, RC = Recall Rate
PR = Precision Rate, , F1 = F1-score

Table 5.4: Results for testing the system for 10 AUs on 77 sequences from the SAL dataset for the FFD method.

We note that here the selected window sizes are much shorter than for the MMI dataset. A possible explanation for this is that spontaneous expressions are generally less smooth and depict multiple apexes interleaved with onset and offset segments. As a result, each segment occurs for a shorter time-period.

5.3.2.4 Generalisation performance

To test the robustness and generalization ability of the proposed FFD method, we performed a smaller test on the Cohn-Kanade (CK) dataset[69]. We only tested on those AUs for which at least ten examples exist in the dataset (18 AUs in 143 sequences). The 10-fold cross-validation results are shown in Table 5.5. As a reference, the F1-scores for the MMI dataset are also repeated. The results achieved for the CK dataset are on average similar to those for the MMI dataset. AUs 2, 5, 12, 15, 20, 24 and 25 perform much better in the CK dataset. Possible explanations for the inferior performance of AU 10, 11, 14 and 45 lie in the differences in ground truth labelling and the absence of offset segments in the CK dataset. The two datasets were labelled in different ways. More specifically, in the CK database, trace activations (FACS intensity A) were also coded, whereas in the MMI dataset only AUs of FACS intensity B and higher were considered. Trace activations (especially in AU 10, 11 and 14) involve very subtle changes in the facial skin appearance, that remain undetected by our method.

AU	θ	CR	RC	PR	F1	F1^V	F1^{MMI}
1	2	88.81	86.89	86.89	86.89	**87.6**	72.73
2	4	94.41	92.31	87.80	90.00	**94.0**	72.73
4	20	74.83	85.96	63.64	73.13	**87.4**	69.33
5	2	92.31	75.86	84.62	**80.00**	78.3	48.48
6	16	94.41	84.21	76.19	80.00	**88.0**	73.68
7	16	71.33	72.00	34.62	46.75	**76.9**	36.36
9	8	93.01	89.47	68.00	**77.27**	76.4	69.23
10	16	89.51	46.67	50.00	48.28	**50.0**	75.86
11	4	88.81	50.00	37.50	**42.86**	—	66.67
12	8	95.10	90.00	78.26	83.72	**92.1**	62.22
14	8	93.01	33.33	42.86	**37.50**	—	51.06
15	8	92.31	68.42	72.22	**70.27**	30.0	56.25
17	4	83.92	72.55	80.43	**76.29**	—	76.50
20	20	90.91	73.53	86.21	**79.37**	60.0	43.48
24	4	90.21	70.59	57.14	**63.16**	14.3	44.00
25	2	95.10	92.68	98.70	**95.60**	95.3	84.66
27	8	95.80	95.45	80.77	87.50	**89.3**	96.30
45	2	92.31	81.48	78.57	**80.00**	—	92.09

Averages	CR	RC	PR	F1
avg. ours, 18 AUs	89.78	75.63	70.25	72.14
avg. ours, 14 AUs	89.86	80.29	73.21	75.85
avg. [151], 14 AUs	90.3	73.3	79.8	72.83

θ = Window Size, CR = Classification Rate
RC = Recall Rate, PR = Precision Rate
F1 = F1-score CK dataset, F1^V = F1-score from [151]
F1^{MMI} = F1-score on MMI dataset

Table 5.5: Results for testing the system for 18 AUs on 143 sequences of the CK dataset.

Test	CR	RC	PR	F1
Trained on MMI, tested on CK	82.52	55.17	65.95	56.13
Trained on MMI, tested on MMI	93.52	76.02	58.79	65.40
Trained on CK, tested on CK	89.78	75.63	70.25	72.14

CR = Classification Rate, RC = Recall Rate
PR = Precision Rate, F1 = F1-score

Table 5.6: Results for cross database testing, 18 AUs.

Another difference between the results is that for the CK dataset, lower window sizes are selected than for the MMI dataset. Since each sequence in the CK dataset ends at the apex of the expression with the offset segments cut off, no GentleBoost classifiers could be trained for the detection of offsets and the HMM classification relies solely on the onset detections. Since the duration of onsets is generally shorter than offsets, shorter window sizes tend to be selected. The absence of offset phases, especially for fast AUs like AU 45, in which onset phases can often not be captured in more than 1-2 frames and the detection relies heavily on the detection of offset phases, explains the inferior performance for such AUs. A possible explanation for better performance for AU 2, 5, 12 and 15 lies in the intensity of these expressions present in the CK dataset. More specifically, facial expression displays constituting the CK dataset are shorter and more exaggerated than it is the case with data from the MMI dataset. The better performance for AUs 24 and 25 can be explained by the greater number of examples present in the CK dataset.

We compare our results to those reported earlier by Valstar & Pantic[151], the only other authors that addressed the problem of AU temporal segments recognition. Valstar & Pantic use 153 sequences from the CK dataset, where we use 143. Their geometric-feature-based approach gives on average very similar results. Interestingly, on this dataset, the results of Valstar & Pantic are much better for AUs 4 and 7 (the related facial displays are characterized by large morphological changes which can easily be detected based on facial point displacements) and the results obtained by the FFD-based method are much better for AUs 15, 20 and 24 (which activations involve distinct changes in skin texture without large displacements of facial fiducial points). Also, the method of Valstar & Pantic is unable to deal at all with AUs 11 (nasolabial furrow deepener), 14 (mouth corner dimpler) and 17 (chin raiser), the activation of which is only apparent from changes in skin texture and cannot be uniquely detected from displacements of facial fiducial points only[111, 109].

A cross-database test was also performed with the MMI and CK dataset. Average

Authors	†	Features	Classification
Bartlett et al. 2005[15]	a,f	Gabor filters	AdaBoost+SVM
Bartlett et al. 2006[16]	a,f	Gabor filters	AdaBoost+SVM
Chang 2006[26]	a,f	manifold embed.	Bayesian
Whitehill & Omlin '06[162]	a,f	Haar wavelets	AdaBoost
Littlewort et al. 2006[94]	a,f	Gabor filters	AdaBoost+SVM
Lucey et al. 2007[98]	a,f	AAM	SVM
Valstar & Pantic 2004[153]	a,t	MHIs	kNN/rule-based
Pantic & Patras 2005[108]	g,t	tracked points	temporal rule-base
Valstar & Pantic 2006[151]	g,t	tracked points	AdaBoost+SVM
Valstar & Pantic 2007[152]	g,t	tracked points	AdaBoost+SVM
Tong et al. 2007[147]	a,t	Gabor filters	AdaBoost+DBN
This work	a,t	FFD	GentleBoost+HMM

†: geometric/appearance-based(g/a), temporal-/frame-based(t/f)

Table 5.7: Comparison of AU recognition methods.

results are shown in table 5.6. The tests were run on those AUs available in both datasets using a temporal window size of 20 frames. The average result is slightly lower than the result for training and testing on the MMI dataset, but this is to be expected given the different coding styles and other differences between the two datasets.

5.3.2.5 Comparison to earlier work

We compared our method to earlier works that reported results on either the CK or the MMI dataset. Table 5.7 gives an overview of these works. It is interesting to note that most works are image-based, which means they derive the classification per frame independently and do not take temporal information into consideration. Additionally, it means that the results reported for those works are found using manually selected ”peak” frames, that is, frames showing the AU in question at maximum intensity. In contrast, sequence-based approaches take the whole sequence into account without prior information as to the location of the peak intensity.

Table 5.8 shows results reported previously on the CK and MMI datasets. While the classification rate (the percentage of correctly classified frames / sequences) is the most commonly reported measure, it is also the one that is the least informative. Especially in cases where the dataset is highly unbalanced, it can be misleading. For example, in our subset of the CK dataset, the percentage of true positive sequences is below 10% for most AUs. This means that it is possible to report a 90% classification rate by simply classifying every sequence as negative. Therefore, we report the F1-

CK dataset, Image-based works							
Authors	**AU**	**NI**	**CR**	**F1**	**FA**	**Hit**	**FRR**
Bartlett '05[15]	17	313	94.8	-	3.9	60.2	-
Bartlett '06[16]	20	2568	90.9	-	8.2	80.1	-
Chang '06[26]	23	258	89.4	-	-	-	-
Whitehill '06[162]	11	580	92.4	-	-	-	-
Littlewort '06[94]	7	313	92.9	-	-	-	-
Lucey '07[98]	15	?	95.5	-	16.7	-	1.9
CK dataset, Sequence-based works							
Authors	**AU**	**NS**	**CR**	**F1**	**FA**	**Hit**	**FRR**
Valstar '04[153]	10	344	68	-	32.0	-	-
Pantic '05[108]	21	90	93.3	-	-	-	-
Valstar '06[151]	15	?	90.2	72.9	-	-	-
Tong '07[147]	14	?	93.3	-	5.5	86.3	-
This work	18	143	89.8	72.1	6.4	75.6	29.1
This work, 15 best AUs	15	143	92.5	72.5	4.8	73.8	26.1
MMI dataset, Image-based works							
Authors	**AU**	**NI**	**CR**	**F1**	**FA**	**Hit**	**FRR**
Chang '06[26]	29	584	91.9	-	-	-	-
MMI dataset, Sequence-based works							
Authors	**AU**	**NS**	**CR**	**F1**	**FA**	**Hit**	**FRR**
Valstar '04[153]	22	253	61	-	-	-	-
Pantic '05[108]	9	45	86.7	-	-	-	-
Valstar '07[152]	23	196	-	66.0	-	-	-
This work	27	264	94.3	65.1	-	-	-

AU = No. of AUs recognized, NI = number of images used
NS = number of sequences used, CR = Classification Rate
F1 = F1-score, FA = False Alarm/Accept Rate
Hit = Hit Rate, FRR=False Rejection Rate

Table 5.8: Comparison of results on CK and MMI dataset.

score, which gives a better understanding of the quality of the classifier. Our results in terms of the classification rate on the CK dataset are largely comparable to those reported in the other works, 89.8% vs. 90.2%, 93.3%. For the MMI dataset, we outperform the other works. The main reason for the worse comparative performance on the CK dataset is probably the absence of offset segments. In contrast, both the MMI and SAL dataset contain the offset segments, which can greatly help validate the occurrence of AUs in our HMM classification scheme.

5.4 Conclusions

In this Chapter we have proposed a method based on non-rigid registration using free form deformations to model dynamics of facial texture in near-frontal-view face image sequences for the purposes of automatic frame-by-frame recognition of AUs and their temporal dynamics. To the best of our knowledge, this is the first appearance-based approach to facial expression recognition that can detect all AUs and their temporal segments.

We have compared our approach to an extended version of the previously proposed approach based on Motion History Images. The FFD-based approach was shown to be far superior. On average, it achieved an F1-score of 65% on the MMI facial expression database, 72% on the Cohn-Kanade database and 76% on the SAL dataset (containing spontaneous expressions). For each correctly detected temporal segment transition, the mean of the offset between the actual and the predicted time of its occurrence is 2.46 frames (in the MMI database).

We have compared the proposed FFD-based method to several other state of the art works. Comparable results have been achieved for the Cohn-Kanade facial expression database, while our work performs best of all those presenting results on the MMI database.

Chapter 6

Fusion of facial and EEG modalities

Contents

6.1 Introduction

In this chapter, we utilize the methods described in Chapters 4 and 5 to investigate the possibilities for multi-modal fusion in affect recognition and implicit tagging. We use a dataset with recordings of participants watching video clips designed to elicit emotional responses. These responses, in the form of detected facial Action Units and EEG signals, are classified into arousal, valence, and dominance classes. Both feature-level fusion and decision-level fusion methods are explored and shown to improve upon the single modality results. In addition, we show how this method can be effectively

stimuli material	movie clips
stimuli selection	manual
number of videos	20
clip duration	variable (35-117s)
num. participants	30 (24 complete)
rating dimensions	arousal, valence, dominance
modalities classified	EEG, Facial expressions
modalities recorded	EEG, physiological, video (6 angles)
dataset publicly available	yes

Table 6.1: Summary of MAHNOB dataset used in this chapter.

used for the implicit tagging of videos by aggregating affect estimates from multiple participants. To the best of our knowledge, this is the first work to perform fusion of the facial expression and EEG modalities for affect recognition or for implicit tagging of videos.

Table 6.1 summarizes the MAHNOB dataset used for experiments in this chapter. This dataset was selected (over datasets used in Chapter 4 for two main reasons:

- Professional-grade, fully synchronized 60 fps face video is available.
- The project requirement of using music videos was no longer present and the movie clips used here may be more effective at eliciting strong emotional responses.

We describe first the used dataset in Section 6.2. Next, we detail the methodology used for feature extraction in each modality and the methods used for fusion in Section 6.3. We present the obtained results from single modalities and from both feature-level and decision-level fusion, as well as the results for implicit affective video tagging in Section 6.4. Section 6.5 concludes the chapter.

6.2 Dataset

We use the MAHNOB HCI[139] dataset in this experiment, which contains EEG, video, audio, gaze and peripheral physiological recordings of 30 participants. Each participant watched 20 clips extracted from hollywood movies and video websites such as YouTube.com and blip.tv. The stimuli were selected in order to elicit 5 emotions (disgust, amusement, joy, fear and sadness). In addition, various weather reports were included as neutral stimuli. The stimuli videos range in duration from 35 to 117 seconds. After watching each stimulus, the participants used SAM manikins to

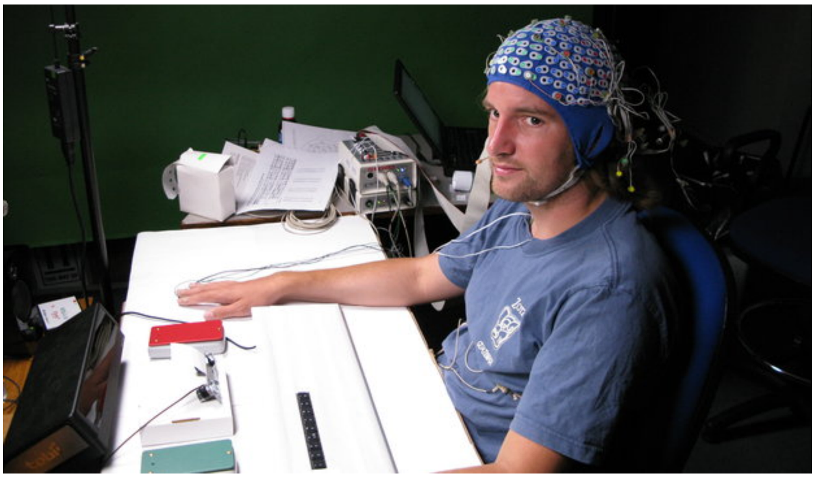


Figure 6.1: A participant in the MAHNOB HCI experiment.

rate their felt arousal, valence and dominance on a discrete scale of 1 to 9. Facial video was recorded by 6 cameras at 60 frames per second from different angles. In this study, we only use the frontal camera and the frame rate is down-sampled to 30 frames per second to more closely match the frame rate of earlier experiments (25 fps) and decrease processing time. 32 channel EEG was recorded at 256Hz using a BioSemi system, similar to experiments described earlier. However, in this case EOG was not recorded.

For 6 of the 30 participants, various problems such as technical failures occurred during the experiment. Here, only the 24 participants are used for which all data is available.

6.3 Affect recognition using EEG and Facial expressions

6.3.1 EEG features

The methodology used for EEG classification is similar to the methods explained in Chapter 4. The signal was down-sampled to 128Hz and a 4-45Hz bandpass filter used to reduce artefacts. As EOG was not recorded, eye movement artefacts were not suppressed. Power spectral density in the Theta (4 - 8Hz), slow alpha (8 - 10Hz), alpha (8 - 12Hz), beta (12 - 30Hz) and gamma (30 - 45Hz) bands was computed, as well as the lateralization for 14 left-right pairs. This gives a total of 230 features.

6.3.2 Facial expression features

We use the system described in Chapter 5 based on non-rigid registration to classify the Action Units in the face video. However, here the final classification target is not the AUs themselves, but rather the arousal, valence and dominance ratings. In addition, no AU ground truth is available for the dataset. Annotating the dataset of in total 480 videos in terms of AUs was not considered feasible. Therefore, we can not train the facial expression recognition system on this dataset. Instead, we use the system as trained on the (posed) MMI dataset, which earlier showed success in classifying AUs in the posed Cohn-Kanade and the spontaneous SAL datasets. The same 18 AUs are detected as before in the CK dataset (AUs 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 14, 15, 17, 20, 24, 25, 27 and 45).

The output of the frame-by-frame AU onset and AU offset gentleboost classifiers as well as the output of the HMM (see Chapter 5) are computed for each video. The window size of the gentleboost classifiers is optimized by cross-validation on the training set.

Several meta-features are extracted from these outputs and subsequently used to predict the arousal, valence and dominance ratings as given by the experiment participants. The 9 features specified in Table 6.3.2 are derived from the output of the onset/offset classifiers for each of the 18 AUs, giving a total of 162 features.

Feature	**Number**
Percentage of frames that the AU is detected in the onset/offset state	2
Number of onset/offset segments detected	2
Mean of the gentleboost classifier outputs	2
Max of the gentleboost classifier outputs	2
Number of complete AU activations detected (meaning neutral → onset → apex → offset → neutral)	1

Table 6.2: Meta-features extracted from face classifier output.

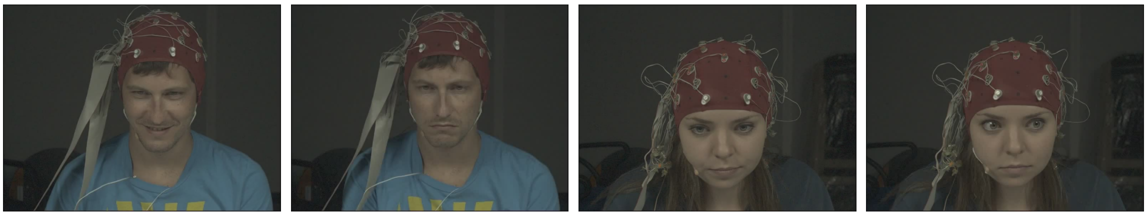


Figure 6.2: Example screenshots of face video in the MAHNOB HCI dataset.

Figure 6.2 shows several screenshots of face video in the MAHNOB HCI dataset. This dataset is quite challenging for facial expression analysis, partly due to the spontaneous nature of the expressions. In addition, unlike in the SAL dataset, there is no (virtual) conversation partner or in fact anyone else in the room. This means the number of expressions is quite sparse and expressions that are shown are generally quite subtle.

6.3.3 Classification

We perform binary classification on the arousal, valence and dominance ratings, which are thresholded into high and low classes. Those videos that were rated in the middle of the 9-point scale are grouped with the low class, since this gives the most balanced class distribution.

We use recursive feature elimination (RFE) to select features for classification. This is done by iteratively calculating the feature weights for a linear SVM classifier and then removing the 10% of features with lowest weights. This continues until the target number of features remains. The number of selected features is optimized by a 10-fold inner cross-validation on the training set.

For classification we use again the Gaussian Naïve Bayes (GNB) classifier as discussed in section 4.3.6. This classifier has the added advantage of providing probabilistic outputs, which can be used for decision-level fusion.

We perform two types of cross-validation. Firstly, we perform 10-fold cross-validation on the entire set of 480 trials. Second, we perform a per-subject leave-one-trial-out cross-validation, where the classifier is trained on 19 trials from the same subject and tested on the 20th.

6.3.4 Fusion

Approaches to fusion are generally divided into two categories; feature-level (or early) fusion and decision-level (or late) fusion. Some works have also attempted a combination of both methods in hybrid fusion[9]. In feature level fusion, the features derived from each modality are simple stacked together to form a single feature vector and a single classifier is used. The advantage of feature-level fusion is the ability to take advantage of correlations between features from different modalities. A disadvantage is that it can be hard to implement as the feature vectors from different modalities can for instance cover different time scales, exhibit large statistical differences or be otherwise different in nature (for instance discrete versus continuous features).

In decision-level fusion, the separate modalities are first classified individually and then the output of these classifications is merged. Decision-level fusion is straightforward to implement, as the output of different classifiers is usually easily converted to the same format. In addition, one can use classifiers for each modality that are most suited to it. However, one loses the ability to take advantage of inter-modality correlations between features. Various methods can be used for decision-level fusion, such as simple rule-based methods (for instance taking the sum or product of class probabilities) to classifier-based approaches, where a meta-classifier is trained taking the decisions from individual classifiers as its features.

The most common modalities to be merged are combinations of audio, video and text modalities[9]. In facial expression analysis, the main modalities to be used are aural and visual features[171]. The EEG modality is most often fused with fMRI , EMG, MEG or PET in clinical settings. For multimedia applications, EEG has been fused with peripheral physiological signals and gaze information in [139], with peripheral physiological signals and audio-visual features in [6, 1] and with audio-visual features in [70, 71]. To the best of our knowledge, the combination of EEG and facial expressions has not been attempted previously. For a more thorough review of previous methods for multi-modal fusion, the reader is referred to the recent surveys by Sebe et al.[136], Zeng et al.[171] and Atrey et al.[9].

We investigate both feature-level and decision-level fusion in this work. In feature-level fusion, the feature vectors from both modalities are stacked together and the total number of features becomes 392. Next, classification proceeds with RFE feature selection and GNB classification as before.

In decision-level fusion, we first classify the modalities individually as described above. Besides the probability p of each class as given by the GNB classifier, we additionally take into account the classifier's training set F1-score for each trial. We distinguish two different types of approaches.

Firstly, we present methods for estimating a per-sample weighting α for the different modalities where the final decision is a weighted sum of the outputs from classification of the individual modalities. For each trial, let p_e^x, p_f^x and $p_o^x \in [0, 1]$ denote the classifier probability for class $x \in \{1, 2\}$ for EEG, facial expressions and fusion respectively. Then the output class probabilities are given by

$$p_o^x = \alpha p_e^x + (1 - \alpha) p_f^x. \tag{6.1}$$

We can also consider the training set F1-scores from each modality for each sample. let $F_e, F_f \in [0, 1]$ denote the training set F1-score for EEG and facial expression classification respectively. We first normalize these scores to ensure the fusion probabilities for all classes sum up to 1. The normalized training set performance t_e and t_f are given by

$$t_e = \frac{F_e}{\alpha F_e + (1-\alpha)F_f}, t_f = \frac{F_f}{\alpha F_e + (1-\alpha)F_f}. \tag{6.2}$$

Then, the output class probabilities are given by

$$p_o^x = \alpha(p_e^x t_e) + (1-\alpha)(p_f^x t_f). \tag{6.3}$$

It can then be shown that $p_o^1 + p_o^2 = \alpha t_e + (1-\alpha)t_f = \frac{\alpha F_e + (1-\alpha)F_f}{\alpha F_e + (1-\alpha)F_f} = 1$.

We also present methods based on meta-classification, where a meta-classifier is trained on the outputs from classification of the individual modalities. The different methods investigated for decision-level fusion are each briefly explained below.

Equal weights fusion (W-EQ and W-EQ^T) **W-EQ** is the most straightforward method, where the output probabilities for each class are an equal weighting of the class probabilities from each single modality ($\alpha = 0.5$). That is,

$$p_o^x = 0.5 p_e^x + 0.5 p_f^x. \tag{6.4}$$

For **W-EQ**^{T}, the training performance is also considered. That is,

$$p_o^x = 0.5(p_e^x t_e) + 0.5(p_f^x t_f). \tag{6.5}$$

Estimated weights fusion (W-EST and W-EST^T) For a given sample, we numerically approximate the optimal decision weight α for the set of all other samples. This is done by varying α between 0 and 1 in steps of 0.01 and choosing that value which gives the highest F1-score on the other samples. The estimated weight is then applied to the current sample as in Equation 6.1. For **W-EST**^{T}, the training performance is also used as in Equation 6.3.

Regression-estimated weights fusion (W-REG and W-REG^T) Here we first train a linear support vector regressor (SVR) on the remaining samples with the class probabilities from the individual modalities as features and the optimal weight, determined as above, as the target. The SVR is then used to predict the optimal weight for the current sample and fusion class probabilities are computed as in Equation 6.1. In **W-REG**^{T}, the training set performance F_e and F_f are also included as features and fusion class probabilities are computed as in Equation 6.3.

Meta-classification of class label (M-CLASS and M-CLASS^T) Here we train a linear SVM classifier on the probabilistic outputs in order to directly predict the class of the current sample. In **M-CLASS**^{T}, the training set F1-scores F_e and F_f are included in the set of features.

Meta-classification of classifier correctness (M-CORR and M-CORR^T**)**
For **M-CORR**, we distinguish the following four cases in the classifier outputs of the remaining samples:

1. both modalities are correct
2. EEG classification is correct
3. Face classification is correct
4. neither is correct

Next, an SVM classifier is trained with these cases as the classes and the probabilistic outputs (and the training set F1-scores in **M-CORR**^{T}) as features. Then, depending on the predicted class, the following actions are taken:

1. equal weighting, $p_o^x = 0.5p_e^x + 0.5p_f^x$
2. $p_o^x = p_e^x$
3. $p_o^x = p_f^x$
4. invert the output of the least certain classifier. That is,
$$p_o^x = \begin{cases} 1 - p_e^x & \text{if } (p_e^1 p_e^2) > (p_f^1 p_f^2) \\ 1 - p_f^x & \text{otherwise} \end{cases}$$

A similar derivation holds for **M-CORR**^{T}.

6.4 Experimental Results

6.4.1 Single modalities

Table 6.3 gives the classification results for each single modality and rating target. Also included as a baseline are the results for completely random classification, majority class classification and classification with the class ratio (as explained in Section 4.3.6). These ratios were estimated from the training set ratings in each case, explaining the difference in the baseline results between per-subject cross-validation and cross-validation on the entire set. For per-subject classification the baseline results are somewhat better as the estimation of class ratio is more accurate.

Both modalities generally perform better when training and testing occur on the same participant, with the exception of face classification for valence ratings, which performs slightly better on the entire set. For arousal and dominance, results are all significantly above the baseline level. Somewhat surprisingly, valence turns out to be

		Arousal		Valence		Dominance	
Training	**Modality**	**CR**	**F1**	**CR**	**F1**	**CR**	**F1**
Entire set	**EEG**	63.1	58.6‡	54.6	52.5	60.0	59.7‡
	Face	53.8	53.4†	57.1	52.3	55.4	52.9†
	Random	50.0	47.2	50.0	48.6	50.0	47.5
	Majority	61.2	37.1	64.0	38.8	53.1	33.4
	Class ratio	52.2	47.7	53.8	49.5	53.8	47.4
Per subject	**EEG**	64.8	62.3‡	64.8	60.7‡	63.1	62.9‡
	Face	64.6	62.8‡	55.2	51.4	63.1	62.9‡
	Random	50.0	47.2	50.0	48.6	50.0	47.5
	Majority	66.5	39.4	64.4	39.0	67.3	39.9
	Class ratio	59.6	50.0	55.4	50.0	58.8	50.0

Table 6.3: Average classification rates (CR) and F1-scores (average of F1-score for each class) for the single modalities. Daggers indicate whether the F1-score distribution over subjects is significantly higher than 0.5 according to an independent one-sample t-test ($^\ddagger = p < .01$, $^\dagger = p < .05$). 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.

the hardest target to classify. In the case of facial expressions, one would for instance expect AU 12 (smile) to be a good indicator of valence. Upon visual inspection of the data however, it was found that the displayed smiles are firstly very subtle (and thus hard to detect), and do not only occur in high valence cases (e.g. awkward smiles can occur during disgusting videos). Another difficulty is the relatively low amount of facial expressions displayed throughout the dataset.

In general, EEG seems to outperform the facial expression analysis. As is the case for results in Chapter 4, arousal generally classifies better than valence. In this experiment however, results are better than those reported in Chapter 4, possibly due to higher effectiveness of the stimuli extracted from films versus the music videos used before. Facial expression analysis performs poorly when training and testing over the entire set, possibly indicating that the method is very subject-dependent. It may be possible to improve facial expression analysis results by training the AU detector on the MAHNOB HCI dataset, rather than the currently used MMI dataset. However, this would require the manual annotation of the entire set for AUs. Another possible improvement may be possible by directly classifying arousal and valence, rather than first classifying the AUs as an intermediary step. We leave these improvements as future work.

6.4.2 Feature-level fusion

		Arousal		Valence		Dominance	
Training	**Modality**	**CR**	**F1**	**CR**	**F1**	**CR**	**F1**
Entire set	**EEG**	63.1	58.6	54.6	52.5	60.0	59.7
	Face	53.8	53.4	57.1	52.3	55.4	52.9
	Fusion	64.6	**62.8**†	60.4	**57.2**†	60.8	**60.8**
Per subject	**EEG**	64.8	62.3	64.8	60.7	63.1	62.9
	Face	64.6	62.8	55.2	51.4	63.1	62.9
	Fusion	64.6	**63.3**	65.8	**64.1**†	65.8	**65.8**

Table 6.4: Average classification rates (CR) and F1-scores (average of F1-score for each class) for feature-level fusion. For comparison the results of single modalities are also shown. The best F1-scores per target and training type are shown in bold. Daggers indicate whether the fusion F1-score distribution over subjects is significantly higher than the score distribution of the best performing single modality according to a related two-sided t-test ($^\ddagger = p < .01$, $^\dagger = p < .05$).

Table 6.4 gives the result for feature-level fusion both per-subject classification and cross-validation on the entire set. In all cases, fusion outperforms the single modalities. The difference is statistically significant for arousal and valence on the entire set and valence on the per-subject basis. This strongly suggests complementary information is present in the two modalities.

6.4.3 Decision-level fusion

For decision-level fusion, we investigated the use of several different methods as described in Section 6.3.4. Results are shown in Table 6.5. While in almost half of the performed tests the fusion does outperform the single modalities, in none of these cases the difference is statistically significant. Feature-level fusion outperforms decision-level fusion for all proposed methods. Decision-level fusion seems to work best for the valence ratings.

The methods that in the most cases improve the results are **W-REG** and **W-EST** (and their counterparts with training set performance included). This suggests that trying to predict or approximate a weighting between the modalities generally works better than trying to directly classify the samples using the individual modalities' outputs as features. In addition, using the training set performance does not seem to lead to consistent improvement for any of the methods.

		Arousal		Valence		Dominance	
Training	**Modality**	**CR**	**F1**	**CR**	**F1**	**CR**	**F1**
Entire set	**EEG**	63.1	58.6	54.6	52.5	60.0	59.7
	Face	53.8	53.4	57.1	52.3	55.4	52.9
	W-EQ	58.3	56.9	55.8	**53.4**	59.6	59.3
	W-EQ^T	61.0	58.5	56.2	**53.9**	59.6	59.3
	W-EST	63.5	**60.9**	57.5	**55.1**	60.0	59.7
	W-EST^T	63.3	**59.9**	60.4	**57.1**	60.2	**59.9**
	W-REG	63.5	**60.9**	57.5	**55.1**	60.2	**59.9**
	W-REG^T	63.3	**59.9**	60.4	**57.1**	60.2	**59.9**
	M-CLASS	63.1	58.6	54.0	52.2	60.0	59.7
	M-CLASS^T	63.1	58.6	54.6	52.5	60.0	59.7
	M-CORR	63.1	58.6	60.0	49.4	55.2	54.4
	M-CORR^T	63.1	58.6	59.4	48.4	55.2	54.4
Per subject	**EEG**	64.8	62.3	64.8	60.7	63.1	62.9
	Face	64.6	62.8	55.2	51.4	63.1	62.9
	W-EQ	65.4	62.5	63.7	58.4	64.2	**63.7**
	W-EQ^T	65.8	**62.9**	65.4	**61.2**	62.9	62.7
	W-EST	64.8	62.6	65.0	**60.9**	64.8	**64.6**
	W-EST^T	65.6	62.8	66.5	**62.4**	61.9	**61.6**
	W-REG	64.8	62.6	65.4	**61.3**	64.8	**64.6**
	W-REG^T	65.8	**63.0**	66.7	**62.7**	64.2	**63.9**
	M-CLASS	64.8	62.3	64.8	60.7	63.1	62.9
	M-CLASS^T	64.8	62.3	64.8	60.7	63.1	62.9
	M-CORR	65.8	**62.9**	64.8	60.7	62.7	62.5
	M-CORR^T	65.8	**62.9**	64.8	60.7	62.7	62.5

Table 6.5: Average accuracies (CR) and F1-scores (average of F1-score for each class) for decision-level fusion. For comparison the results of single modalities are also shown. F1-scores higher than the best single-modality result are bold.

6.4.4 Implicit Video tagging

Table 6.6 shows the results for implicit tagging of the videos based on the classifier outputs of all participants. This is challenging since the classes here are not objectively defined. Arousal, valence, and dominance are subjective measures and participants frequently disagree on the affective content of videos. Nevertheless, in most cases there is a reasonable agreement between participants, which can be seen in the table. Here, we assign a class to each video clip based on the majority opinion of the participants. In the three cases where an even split of opinions was observed, the video was assigned to the negative class. Next, we estimate the class based on the outputs of the EEG, facial expression, and feature-level fusion classifiers. This is based on the per-subject

	Arousal					Valence					Dominance				
V	**H**	**C**	**E**	**Fa**	**Fu**	**H**	**C**	**E**	**F**	**Fu**	**H**	**C**	**E**	**Fa**	**Fu**
1	62	+	−	−	−	100	−	−	−	−	83	−	−	−	−
2	70	+	−	−	−	100	−	−	−	−	91	−	−	−	−
3	83	−	−	−	−	91	+	−	−	+	79	+	+	−	+
4	50	−	−	−	−	79	−	−	−	−	50	−	−	+	−
5	54	+	−	−	−	79	−	−	−	−	79	−	−	−	−
6	50	−	−	−	−	100	+	−	−	+	66	+	−	−	+
7	83	−	−	+	−	83	+	+	−	+	70	+	−	−	−
8	75	−	−	−	−	87	+	−	−	−	87	+	−	−	−
9	54	−	−	−	+	91	−	−	−	−	79	−	−	−	−
10	70	−	−	−	−	79	+	−	−	+	75	+	+	−	+
11	70	+	−	−	−	95	−	−	+	−	100	−	−	−	−
12	75	−	−	−	−	100	−	−	−	−	54	−	−	−	+
13	70	+	−	+	+	95	−	−	−	−	70	−	−	−	−
14	66	−	−	−	−	100	−	−	−	−	87	−	−	−	−
15	95	−	−	−	−	79	−	−	−	−	66	+	−	−	+
16	70	+	−	−	+	100	−	−	−	−	87	−	−	−	−
17	100	−	−	−	−	83	−	−	−	+	70	+	+	+	+
18	70	−	−	−	+	95	+	−	−	−	79	+	−	−	+
19	91	−	−	−	−	83	−	−	−	−	58	+	−	−	+
20	58	−	−	−	−	70	+	−	−	+	66	+	+	+	+
Mean human agreement and modality classification rates															
	.71		.70	.70	.70	.89		.70	.60	.85	.75		.70	.55	.85

Table 6.6: Video class labelling based on human annotation. The assigned class (**C**) of each video is determined by a majority vote among the 24 human raters. **H** stands for human agreement with the class label. **E**, **Fa**, and **Fu** indicate the assigned class labels from EEG, facial expressions and feature-level fusion classification respectively. **V** stands for video number. Also given is the mean human agreement with the class labels and the classification rate for each modality.

leave-one-video-out cross-validation. As can be seen from the table, the feature-level fusion performs much better on this task, frequently correcting the errors of one or both modalities. For both valence and dominance, an 85% classification rate is achieved. This is a strong indication that the gathering and analysis of implicit responses in large quantities can provide effective and reliable emotional tags.

Figure 6.3 shows the classification accuracy for different numbers of participants. This is the average classification accuracy over videos for the given number of participants, determined by averaging over all possible combinations of the 24 participants in the experiment. A clear advantage can be seen in aggregating the results from multiple participants. In the case of arousal and valence, it does appear the accuracy is levelling of after a certain number of participants has been reached. Naturally it is impossible to obtain a perfect classification as the ground truth is itself subjective and the human annotators themselves do not always agree on the appropriate class

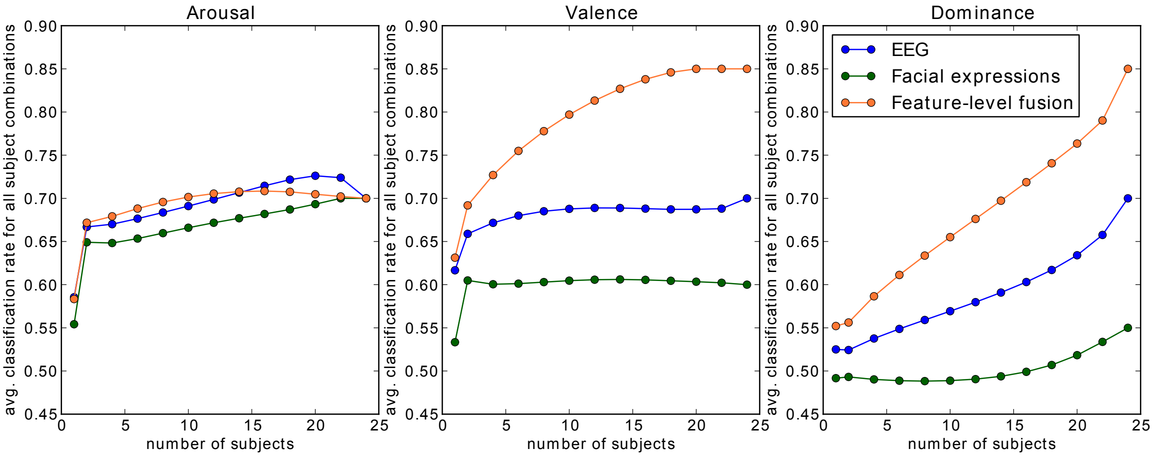


Figure 6.3: The number of participants plotted against the average classification rate for each rating scale.

label.

6.5 Conclusions

In this chapter, we explored several methods for the fusion of the EEG and facial expression modalities. A large dataset containing recordings of 24 subjects each watching 20 video clips was utilized for evaluation of these methods. In a binary classification of arousal, valence, and dominance affect dimensions, significant results were attained for both single modalities. A feature-level fusion approach is demonstrated to improve upon these single modality results. In addition, several methods were investigated for decision-level fusion. While the decision-level methods in nearly half of the cases outperform the single modalities, these improvements were not found to be significant.

Finally, we show the potential of this method of implicit tagging when aggregating the results of multiple participants. For arousal, valence, and dominance, video tag classification rates of 70%, 85% and 85% are attained respectively when aggregating across all 24 participants.

Chapter 7
Conclusions

7.1 Results and contributions

In this work, we have described our research on using two different modalities for implicit tagging and tag validation.

In Chapter 3 we proposed a method for the use of EEG signal analysis to validate implicit tags. To validate this method, we conducted an experiment with 17 subjects, each watching 98 videos displayed in conjunction with valid or invalid tags. Our hypothesis was that we would find an N400 event-related potential that would be present more strongly in the case of invalid tags. We found significant differences between EEG signals recorded during the display of valid and invalid tags. For single-trial classification, we have investigated three different approaches for features extraction: common spatial patterns, discriminative spatial patterns and the windowed means approach. For all three approaches, a linear SVM was used for classification. The discriminative spatial patterns approach gave the best results (55.8% mean classification rate), with results significantly higher than chance level for 12 out of 17 subjects. This indicates the proposed method could successfully be used for tag validation, especially if the validations were aggregated over a large number of participants.

In Chapter 4 our approach to assessing participants' emotional states based on EEG signals is described. We conducted a series of experiments culminating in the work done on the DEAP dataset. For this dataset, the EEG signals of 32 participants were recorded as each watched a collection of 40 music videos. Participants next rated their responses on the scales of arousal, valence, dominance and liking. We presented a novel semi-automatic stimuli selection method using affective tags. Single-trial classification was performed for the scales of arousal (F1: 58.3%), valence (F1: 56.3%) and liking (F1: 50.2%). The results for arousal and valence were shown to be significantly better than random classification. A method for regression based on the EEG signals was also presented, with significant results for arousal, valence, dominance and liking. This dataset was made publicly available and up the writing of this thesis researchers from 36 different institutions have requested access. Based on the experience gained from the analysis of this dataset, a real-time affective music video system was implemented and demonstrated.

In the research on facial expressions described in Chapter 5 we have proposed the FFD technique to model dynamics of facial expressions in near-frontal-view face image sequences for the purposes of automatic frame-by-frame recognition of all 27 AUs and their temporal dynamics. On average, our approach achieved an F_1-score of 65% on the MMI facial expression database, 72% on the Cohn-Kanade database and 76% on the SAL dataset (containing spontaneous expressions). We have compared the proposed FFD-based method to several other state of the art works. Comparable results have been achieved for the Cohn-Kanade facial expression database, while our work performs best of all those presenting results on the MMI database.

Finally, as described in Chapter 6, we explore methods for fusion of the EEG and facial expressions modalities. To evaluate these methods, a dataset containing recordings of facial expressions and EEG of 24 participants watching hollywood movies and other emotion-eliciting clips was used. We presented results for the single modalities that were significantly higher than chance level. Results of a feature-level fusion were consistently higher than the best single modality. We also investigated various methods for decision-level fusion, where results were less convincing than for feature-level fusion, but still higher than the single-modality results in many cases. We also demonstrate how even with the relatively low classification results, our approach to implicit tagging can be highly effective when aggregating results of many participants. When considering all 24 participants, the binary classification rate for arousal, valence and dominance per video reaches 70%, 85% and 85% respectively.

7.2 Lessons learned

During the completion of the work presented here, many mistakes were made by the author. Below is a list of several lessons learned (the hard way):

- Obtaining reliable ground truth Action Unit annotations was a major challenge in this work. It can take about an hour to annotate a minute of video with the Facial Action Coding System. And ideally, multiple annotations should be available in order to compute annotator agreement and ensure maximum reliability. As the average PhD student does not have the resources to hire expert annotators, it is advisable to use existing datasets with ground truth annotations as much as possible.

- Obtaining reliable self-reporting of affect is also challenging. The utmost care should be taken to ensure participants fully understand the used affect scales. Ideally, the same explanation of self-reports should be given to all participants.

Even so, the stimuli themselves can be ambiguous and many external, and hard to control for, factors influence the ratings. Examples include mood, time of day, familiarity with the stimulus, order of stimulus presentation, etc. etc.

- When doing EEG experiments, it is of critical importance to do everything possible to minimize the influence of external factors. This includes having the participant in a noise-isolated separate room during the experiment, insuring a constant level of lighting and temperature, minimizing subject fatigue (by including breaks and limiting the total experiment duration). Also, it is important to make sure the experimental system is well-tested to avoid nasty surprises (crashing experiments, data loss) with the first set of participants. Finally, care should also be taken with the choice of stimulus material, in order to maximize the chances of eliciting the desired affective responses.

- In single-trial analysis, we found substantial differences in performance for different participants. Also, training a one-fits-all classifier did not work nearly as well as subject-dependent classifiers. In addition, due to changes in external factors mentioned earlier, classifiers may not generalize well over time. To achieve the best performance, affective computing platforms should therefore attempt to model the context as well as adapt to the persons using them.

- When working with video material, it is easy to identify mistakes in labelling upon inspection of the data. For instance, when testing a facial expression classifier, it is easy to notice that the labels for Action Units 12 and 45 were accidentally swapped, just by looking at the video. This is not the case for EEG signal data, which all looks pretty much the same, and the utmost care should be taken, with checks along every step of the way, to ensure no mix-ups occur.

7.3 Directions for future research

The field of affective tagging, and affective computing in general, is still very young and many questions remain unanswered. Some possible directions for future research include:

- An interesting question is to investigate how well methods such as those presented here will generalize, both over time, and over participants. How do context, mood, etc. influence the performance of such methods?

- in this work, we considered each trial as a single data point. However, one may feel a wide range of emotion whilst viewing for instance a music video. It would be very interesting to consider the temporal variation of affective state. This would however, probably also require temporal ground truth annotation, which may be difficult to obtain without distracting participants too much.

- The modalities investigated in this work represent only a subset of possible modalities that can be used in affective computing. Modalities such as body pose, MEG, voice and non-vocal utterances, gaze, gestures, device interaction measures(e.g. typing speed), can all play in role in assessing a user's affective state. Fusing information from all of these modalities and identifying their relative strengths and weaknesses may increase affect estimation performance substantially.

- It has been shown in previous work on EEG that task-relevant events often elicit stronger responses than task-irrelevant events. In this thesis, participants were not given specific tasks to complete. In general, giving a participant a task goes against the concept of implicit tagging, where the goal is to assign tags by observing the user without any active participation from that user. However, one can envisage several scenarios in which users give themselves a task to complete, for instance while performing work tasks, or while playing a game. In these cases, the implicit tagging paradigm can still apply. It would be interesting to investigate whether in such cases, affective state estimation for task-relevant events performs better than for task-irrelevant events.

- Whilst the methods and apparatus used in this thesis work well in a laboratory setting, they are not directly applicable in real life. Recently, many new devices have been introduced that are unobtrusive and user-friendly, offering perspectives for use in everyday life (such as simple EEG devices with fewer (dry) electrodes and short setup times, physiological sensors embedded in every day devices, etc. etc.). Investigating if and how well methods for affective state estimation will translate when used with such devices is an interesting perspective.

Appendix A

A Real-time affective recommendation system

A.1 Introduction

As a proof of concept, a real-time affective music video recommendation system was implemented. This was a collaborative effort within the framework of the EU FP7 project PetaMedia[1]. Besides the EEG analysis method described in this work, modules were also created for affective state assessment based on multimedia content analysis and peripheral physiological signals. The output of all these modules was then merged into a single estimate of the user's affect. This estimate was then used as an input to a music recommendation module, yielding real-time affective music video recommendations.

The implementation of and subsequent experiment with this real-time recommendation system served two main goals:

1. Design and implement a real-time affective recommendation system to demonstrate the feasibility of this approach.
2. Evaluate what the added value is of affective feedback in this content recommendation systems.

Using the same techniques as described in Chapter 4, a dataset containing 300 music videos was collected. These videos were then all rated online by volunteers in terms of valence and arousal. An existing music recommendation system based on Last.fm taste profiles was adapted to accept affective inputs and used to deliver the recommendations.

The system operates as follows. First, the participant gives his Last.fm username. His most recent songs are retrieved from the Last.fm database and personalized clusters are generated for the 300-song dataset, based on the participants' taste profile as defined by the recent songs and the average affective ratings by online volunteers.

[1] PetaMedia: `http://www.petamedia.eu`. Collaborating researchers were: Mohammad Soleymani, Guillaume Chanel, Ashkan Yazdani, Eleni Kroupi and Engin Kurutepe.

Next, the participant watches and rates (in terms of valence and arousal) a series of 20 video clips as his EEG and physiological signals are recorded. This data is then used to train the binary valence and arousal classifiers as described in Chapter 4 and [1]. After the classifiers are trained, the system is started. The participant's valence and arousal is predicted every ten seconds by MCA, EEG and peripheral physiological classification. These estimates are then fused, weighted by the class probabilities. The final valence and arousal estimate is then used as an input to the recommendation HMM, which outputs the next song to play. Participants were also able to skip songs, which further influenced the recommendation HMM, making it more likely to transition to a different cluster of songs.

A.2 Design

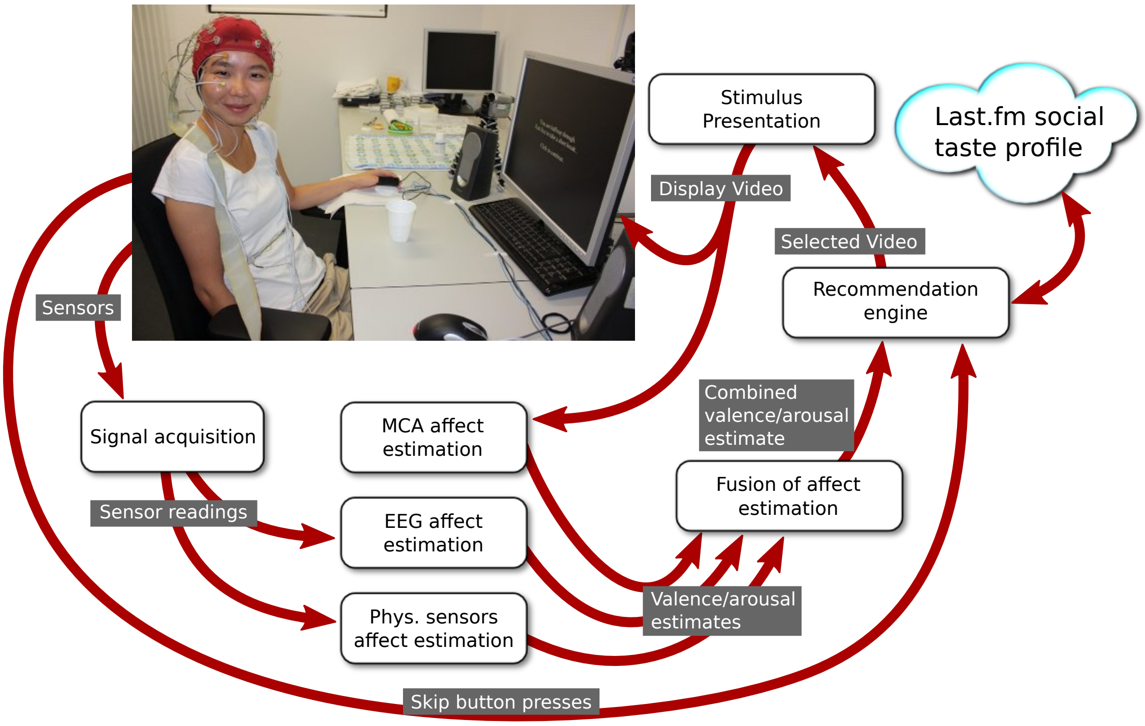


Figure A.1: Overview of the real-time music video recommendation system.

A high level overview of all the modules involved in the system is given in Fig. A.1. The modules were built mostly independently and then joined together in a Python-based framework. Each module exposes a socket to communicate with the other modules. This means modules can easily be run on different computers if need be and facilitates the independent development of each module, regardless of implementation details. The functionality of each module is described below.

MCA affect estimation module This module uses multimedia content analysis techniques to estimate valence and arousal values for a given video, based on features extracted from the audio and video tracks. See [1] for more details. While the extracted features are always the same for the 300 songs in the dataset, the ground truth is unique and so a new classifier is trained for each participant. Contrary to [1], here the songs are segmented into 10-second segments and each segment is classified separately, ensuring a new estimate is available every ten seconds.

EEG/Physiological sensor affect estimation module This module uses the developed techniques to estimate arousal and valence from the EEG and physiological signals. Classification based on peripheral physiological classification proceeds as described in [1]. Contrary to [1], here we classify the last 60 seconds of signal data every ten seconds.

Fusion of affect estimations The estimates of the three affect estimation modalities are fused by this module, using a straightforward equal weighting method that takes into account the class probabilities of each estimator. It returns the estimated classes for arousal and valence every ten seconds.

Recommendation module An existing music recommendation system was used to deliver the recommendations. Given a Last.fm account, This system generates song clusters based on tag similarity in the Last.fm database for the last 10000 songs listened to with the given account. Then, a Hidden Markov Model is used to traverse these clusters and output song recommendations. To adapt this system for affective recommendations, the clustering was limited to the 300 songs in our dataset and affect ratings were included as clustering parameters. Then, the real-time affective feedback is used as an input to the HMM with the goal of selecting those clusters containing songs with similar affective state ratings as the current affect estimate.

Stimulus presentation module This module displays the videos to the user. The user has the ability to skip a song if he/she does not like it. The interface to accomplish this and to log all the interaction is implemented here. When the user skips a song or it ends, this module communicates with the other modules to obtain the recommendation for the next song to play.

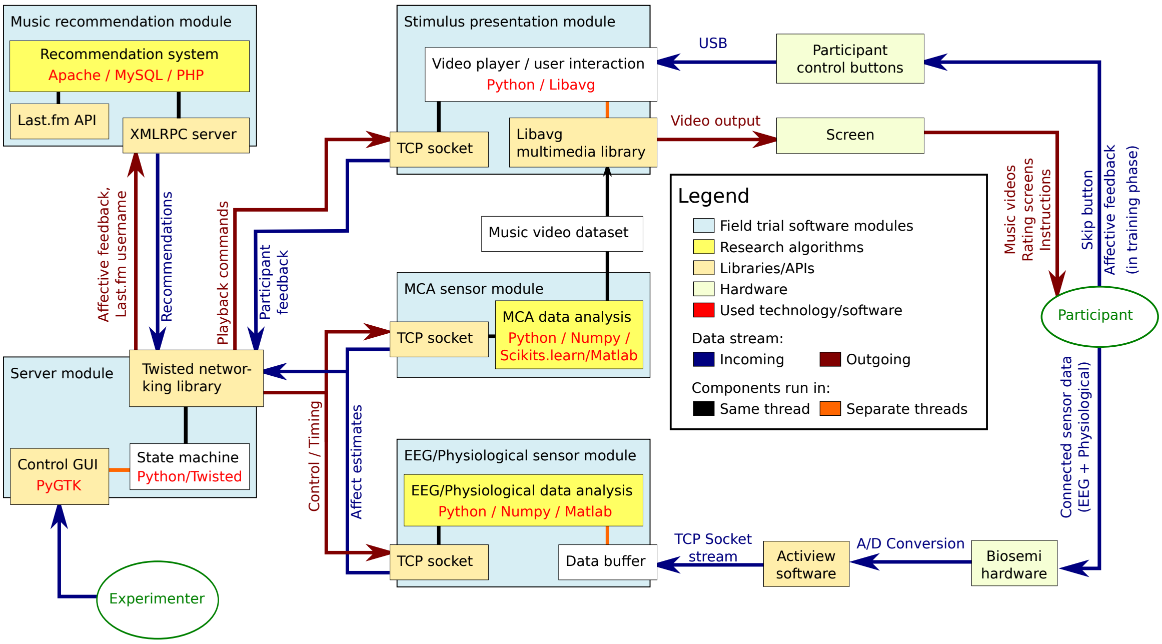


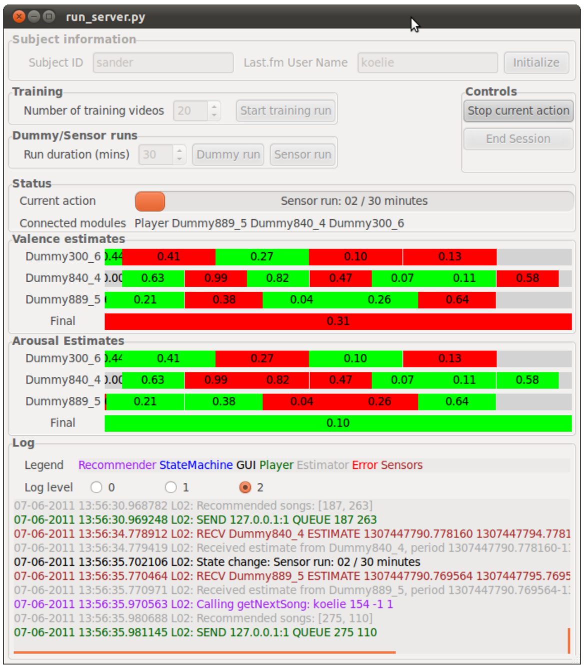
Figure A.2: Overview of system architecture.

A.3 Framework implementation

Figure A.2 gives a more low-level overview of the system architecture. As can be seen, several different programming languages (e.g. Python, Matlab, PHP) and libraries (Twisted, LibAVG, PyGTK, etc.) were used in the creation of the different modules. In addition, several modules were expected to be so demanding that not all modules would be able to run on a single PC and still deliver real-time performance. For these reasons, a framework was needed that would enable the modules to run independently on different systems and communicate over the network, as well as provide facilities to connect the different programming languages.

The chosen design was based on the Python[2] programming language, using the Twisted asynchronous communication framework[3]. This met the requirements set out above. The framework consists of a central server application for controlling and monitoring the system. A screenshot of the server GUI is displayed in Figure A.3. The experimenter can choose various experiment settings and monitor in real time the incoming affect estimates and other communication within the framework. A client application was implemented for each of the aforementioned modules, enabling the modules to communicate over TCP sockets using the Twisted framework. The client applications wrap the original implementations in their respective programming

[2] http://www.python.com
[3] http://twistedmatrix.com


run_server.py
Subject information
Initialize
Training
Start training run
Controls
Stop current action
End Session
Dummy/Sensor runs
Dummy run
Sensor run
Status
Valence estimates
Dummy300_6 0.44 0.41 0.27 0.10 0.13
Dummy840_4 0.00 0.63 0.99 0.82 0.47 0.07 0.11 0.58
Dummy889_5 0.21 0.38 0.04 0.26 0.64
Final 0.31
Arousal Estimates
Dummy300_6 0.44 0.41 0.27 0.10 0.13
Dummy840_4 0.00 0.63 0.99 0.82 0.47 0.07 0.11 0.58
Dummy889_5 0.21 0.38 0.04 0.26 0.64
Final 0.10
Log
Legend Recommender StateMachine GUI Player Estimator Error Sensors

Subject ID: sander
Last.fm User Name: koelie
Number of training videos: 20
Run duration (mins): 30
Current action: Sensor run: 02 / 30 minutes
Connected modules: Player Dummy889_5 Dummy840_4 Dummy300_6
Log level ○ 0 ○ 1 ● 2

```
07-06-2011 13:56:30.968782 L02: Recommended songs: [187, 263]
07-06-2011 13:56:30.969248 L02: SEND 127.0.0.1:1 QUEUE 187 263
07-06-2011 13:56:34.778912 L02: RECV Dummy840_4 ESTIMATE 1307447790.778160 1307447794.7781
07-06-2011 13:56:34.779419 L02: Received estimate from Dummy840_4, period 1307447790.778160-1:
07-06-2011 13:56:35.702106 L02: State change: Sensor run: 02 / 30 minutes
07-06-2011 13:56:35.770464 L02: RECV Dummy889_5 ESTIMATE 1307447790.769564 1307447795.7695
07-06-2011 13:56:35.770971 L02: Received estimate from Dummy889_5, period 1307447790.769564-1:
07-06-2011 13:56:35.970563 L02: Calling getNextSong: koelie 154 -1 1
07-06-2011 13:56:35.980688 L02: Recommended songs: [275, 110]
07-06-2011 13:56:35.981145 L02: SEND 127.0.0.1:1 QUEUE 275 110
```

Figure A.3: Screenshot of the server application.

languages (using Mlabwrap[4] for calls to Matlab, the Twisted framework for calls to PHP and the BioSemi recording hardware).

In previous experiments, we used the Presentation software to display the videos to the user and log their interaction with the system. However, it was quickly clear that this software would not be able to communicate with the rest of the framework in a convenient way. Therefore a substitute stimulus presentation module was implemented in Python using the Libavg toolkit[5] for video display and interaction.

A.4 Experiment

The field trial experiments were held over a two week period from the 13th to the 24th of June 2011 at the University of Geneva. A total of 35 participants took part in the experiment. An image of one of the participants is shown below.

The field trial experiment consisted of the following steps:

1. The participants first gave their last.fm account details in order to access their music taste profile, which is used in the recommendation module. For those participants with no last.fm accounts, the tastebuds.fm web-service was used to find a last.fm taste profile that closely matched their own interest.

[4] http://mlabwrap.sourceforge.net/

[5] http://www.libavg.de/

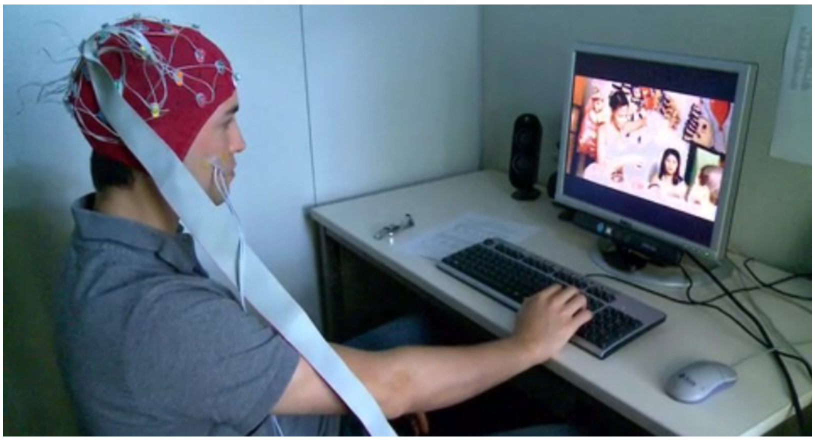


Figure A.4: A participant using the real-time music video recommendation system.

2. Next, the participants signed a consent form and read a set of instructions detailing the experiment protocol.

3. Participants were then led into the experiment room (with controlled lighting and temperature) and the EEG and peripheral physiological sensors were attached.

4. Participants then saw a set of 20 one-minute excerpts of music videos and rated their felt affect in terms of valence and arousal.

5. The signals recorded during this training period in combination with the provided affect ratings were then used to train personalized affect classifiers for the EEG, MCA, and peripheral physiological signals modalities.

6. Next followed two 30-minute sessions where participants were shown a series of music videos recommended to them by the recommendation module. In one of these sessions, the recommendation was solely based on the taste profile. In the other session, the recommendation was based on a combination of the taste profile and the output of the affect estimation modules. The order of the session was randomized per participant and the participants were not aware which session was which. During the sessions, participants were able to skip unwanted songs and after each session, they filled out a questionnaire regarding their experiences.

After the field trials, we analysed the responses to the questionnaires and the number of skipped songs for each session.

Measure	Baseline	Affective
Average and standard deviation of duration of skipped songs	63.04s (25.00s)	57.74s (24.27s)
Average and standard deviation of percentage of skipped songs	88.92% (16.43%)	85.74% (21.04%)

Table A.1: Comparison of song skipping in the baseline and affective systems.

A.5 Results

Of the two goals state in Section A.1, we can state the the first goal has clearly been achieved, namely the design and implementation of a real-time affective music video recommendation system. For this, all involved technologies have been integrated and adapted to work in real-time. The resulting prototype has proven to be a stable platform, which will be used for further experimentation in the future.

Our second goal was evaluating the added value of affective feedback in recommendation. This goal is much harder to evaluate. In the field trials, two systems were compared, the affective recommendation system and the baseline recommendation system (where it should be noted that the baseline system is actually a state of the art recommendation system in itself). Besides a challenge baseline, other difficulties include the relatively small number of subjects and the relatively short time spent with the system. At the same time we face a large amount of variables that are impossible to control (the participant's mood, familiarity with songs, large variations in song types etc.). All of this makes for a very challenging evaluation.

We have two main measures of system performance. First, we measured the percentage of songs skipped during each session and the average length of time the songs were listened to. It stands to reason that the system that recommends the best songs should also show the least percentage of skipped songs and the longest average listening time per skipped song.

As can be seen from Table A.1, participants listened on average for 6 more seconds to songs in the baseline system, but at the same time, skipped 3% more of the songs recommended by this system. Unfortunately, the large standard deviation of these measures renders the difference statistically insignificant. So, based on this measure, it is impossible to pass judgement on which system recommends the best songs.

The second measure of system performance is through analysis of the subjective questionnaires. The questionnaires were modelled after previously used questionnaires for user-centric evaluation (by Knijnenburg, Jones, Pu, and Novak). The questionnaire had 40 questions and answers were given on the 7-point Likert scale. The

Category	#	Example question
Acceptance	4	If an easy to use version of this system exists, I will use it to listen to music
Satisfaction	7	I enjoyed the presented clips
Personalization	5	The presented clips were tailored to my taste
Diversity	4	The clips were similar to each other
Interactivity	7	The system took my emotional disposition into consideration to present clips to me
Immersion	8	I forget about my immediate surroundings when I use the system
Novelty	4	The system helped me to discover new songs/clips
Preference of affective system over baseline system	1	The clips I saw during the first session better fit my interests than in the second session

Table A.2: Questionnaire categories, with number of questions per category and example question.

questions were drawn from 8 categories and can be considered rephrasings of each other. Table A.2 gives an overview of these categories, the number of questions per category and an example question from each category.

The responses from questions from each category were averaged in order to improve the reliability of the responses. Figure A.5 below shows the result averaged over all participants.

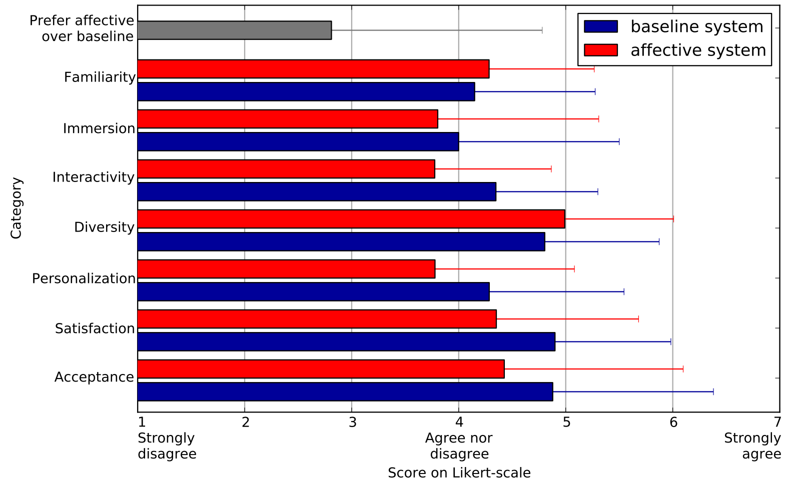


Figure A.5: Outcome of questionnaire comparing baseline and proposed systems.

A paired t-test was performed to determine the significance of the difference in results for both systems. For the categories novelty ($p = 0.42$), immersion ($p = 0.09$)

and diversity ($p = 0.27$) the score difference is not significant. However, for the categories of acceptance, interactivity ($p = 0.009$), satisfaction ($p = 0.03$), acceptance ($p = 0.02$) and personalization ($p = 0.04$) the baseline system significantly outperforms the proposed affective system. Furthermore, most participants clearly prefer the songs recommended by the baseline system over those recommended by the proposed affective system.

Interestingly, for either system, most participants agree more than disagree that the recommended songs were diverse, satisfying and that they would be interested in using such an affective recommendation system if it were easier to use. One can think of a system where a person just wears an unobtrusive EEG-sensing hat, or for instance a heart-rate sensing computer mouse (unlike the current system, where a 30 minutes setup time is still required).

In general, the results are disappointing. While our first goal of developing this system has been achieved, we were clearly not able to demonstrate the added value of affective feedback over current recommendation systems. Several factors may have contributed to this:

- The affective state estimation modules were adapted to work in real-time situations. This means that a shorter signal window is available for classification.

- We had only limited time available during each experiment. It is not recommended to use EEG equipment continuously for several hours. It is also not advisable to perform the different sessions of the experiment over different days, due to the significant differences in EEG signals that would occur. Lastly, there is clearly a limit to how long one can watch a set of music videos without becoming severely distracted. Because of this, we had to limit the number of training videos to 20. This may have lowered the accuracy of affective state estimation. Furthermore, the test session length of 30 minutes may be rather short to properly evaluate a music recommendation system.

- Our dataset size was limited to 300 music videos. This was mainly due to the requirement of multiple manual affect ratings. These ratings were needed to perform the affective clustering in the recommendation system. A limited dataset also means limited options for recommendation and limited personalization based on participants' taste profiles.

- Currently, the affective input is used in the recommendation system in such a way that the goal of the recommendation is to recommend those songs that match the current predicted affective state. However, does one always want to

	Arousal		Valence	
Modality	**CR**	**F1**	**CR**	**F1**
EEG/Physiological	56.6%	0.56	57.9%	0.53
Multimedia Content Analysis	55.7%	0.53	63.0%	0.60
Combined	58.7%	0.57	57.5%	0.53

Table A.3: Classification results using majority labels from online assessment. CR = Classification rate, F1 = F1-score.

hear happy songs when one is happy? The answer to this question currently eludes us. More research is clearly needed into determining successful strategies for navigation the song clusters based on the estimated affective input.

- More research is also needed into determining what is a proper 'weighting' for the affective feedback within the recommendation. The affective state estimation works only to a certain extent (we have reported up to 65% classification rates on binary classification of arousal and valence in earlier experiments). Further research will have to determine if and how we can use these weak (but nevertheless valuable) classifiers to improve rather than disrupt recommendation systems.

While we can not currently determine how much each of the above factors contributed to the achieved results, we can at least get some idea about the effectiveness of the affective state estimation in a real-time setting from the field trial experiments. To do this, we investigate the logged affective state estimates produced during the experiment. While we have no ground truth rating from the participants themselves, we can compare these estimates to the average arousal and valence ratings given to each of the songs by the volunteers in the online assessment. This is of course not a perfect solution, as the song ratings are quite subjective. However, for most songs, there is still a reasonable agreement between participants where it concerns binary labels of low/high arousal/valence. Thus, we threshold the labels from the online assessment and those gathered during the experiment into low and high classes. For each song, we take the average of all the arousal/valence estimates produced. Table A.3 gives the classification rates and F1-measures for the different modalities and rating scales.

Unsurprisingly these results are somewhat lower than those reported in previous work, due to the factors specified earlier and the lack of a personalized ground truth. Nevertheless, we can still see that the results are clearly better than random classification. This result is important, as it is the first validation of our methods in real-time settings. This also leads us to believe that the main problem requiring more research is the connection between affective state estimation and recommendation as explained above.

A.6 Dissemination of the prototype

The following dissemination activities have taken place:

- A short documentary about the system was recorded by the Swiss TV station TSR (Television Suisse Romande), which can be viewed at: http://www.nouvo.ch/2011/09/pimp-my-music.
- A live demonstration of the prototype took place at PetaMedia stand at the NEM Summit 2011.
- A talk discussing the field trials was given in the 'Search Computing and Social Media Workshop' at the NEM summit 2011. This talk was recorded and available online at: http://videolectures.net/petamediaworkshop2011_koelstra_trials/

A.7 Conclusion

In this Appendix, we described the design and development of a real-time affective music video recommendation prototype. While the prototype is fully functional, the goal of evaluating whether users perceive the addition of affective feedback in the recommendation system as useful was not successfully achieved. More research is required in order to effectively combine the affective state estimation with the recommendation system. However, the developed prototype is the first of its kind and breaks new ground in the research on real-time affective tagging based on user's implicit reactions. We demonstrated that the methods developed here can work in a real-time setting and provide meaningful tags. Another interesting output is that participants indicated that they are interested in and willing to accept such systems, provided they are more practical and easy to use. Overall, the prototype developed during the field trials lays an important foundation for further research in this area.

This proof-of-concept system has been tested with 36 participants over a two week period. The system ran distributed over 4 PC's and after some start-up problems, remained stable throughout the testing period. A short documentary about the system was recorded by the Swiss TV station TSR[6] (Television Suisse Romande). A

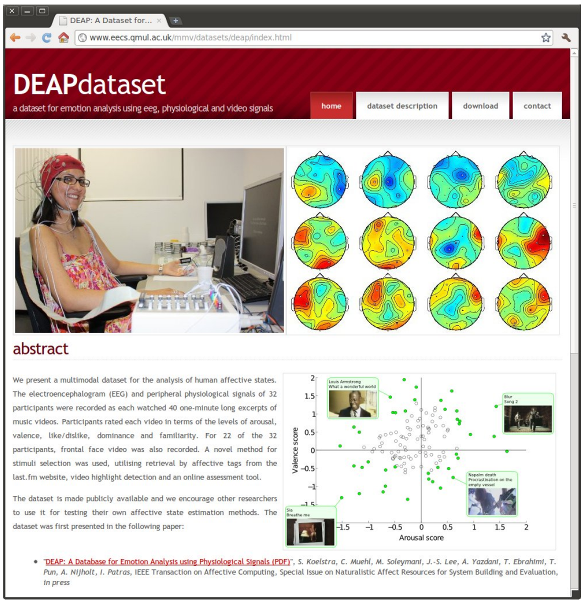
[6] http://www.nouvo.ch/2011/09/pimp-my-music

demonstration of the system took place at the NEM summit 2011[7], with more public demonstrations planned. The system is currently in preparation to be licensed as an open-source toolkit. Due to the modular construction of the system, it is highly adaptable and could be used in many scenarios involving real-time analysis of affective states.

[7] http://nem-summit.eu

Appendix B
The DEAP dataset

B.1 Introduction


DEAP: A Dataset for...
www.eecs.qmul.ac.uk/mmv/datasets/deap/index.html
DEAPdataset
a dataset for emotion analysis using eeg, physiological and video signals
home
dataset description
download
contact
Louis Armstrong What a wonderful world
Blur Song 2
Napalm death Procrastination on the empty vessel
Sia Breathe me
Valence score
Arousal score
2
1.5
1
0.5
0
−0.5
−1
−1.5
−1.5
−1
−0.5
0
0.5
1
1.5
2

abstract

We present a multimodal dataset for the analysis of human affective states. The electroencephalogram (EEG) and peripheral physiological signals of 32 participants were recorded as each watched 40 one-minute long excerpts of music videos. Participants rated each video in terms of the levels of arousal, valence, like/dislike, dominance and familiarity. For 22 of the 32 participants, frontal face video was also recorded. A novel method for stimuli selection was used, utilising retrieval by affective tags from the last.fm website, video highlight detection and an online assessment tool.

The dataset is made publicly available and we encourage other researchers to use it for testing their own affective state estimation methods. The dataset was first presented in the following paper:

- "DEAP: A Database for Emotion Analysis using Physiological Signals (PDF)", *S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, I. Patras*, IEEE Transaction on Affective Computing, Special Issue on Naturalistic Affect Resources for System Building and Evaluation, *in press*

Figure B.1: Screenshot of the DEAP dataset website.

The publicly available DEAP dataset is a direct result of the research described in Chapter 4. The data used in that Chapter was recorded in a collaborative effort in the framework of the EU FP7 project PetaMedia[1]. Data was recorded in two locations, at the University of Twente (Netherlands) and at the University of Geneva (Switzerland). The data was made available on a specially designed website at http://www.eecs.qmul.ac.uk/mmv/datasets/deap/ (see Fig. B.1). At the time of writing, 36 institutions have requested access to the data.

In the dataset, music video clips are used as the visual stimuli to elicit different emotions. To this end, a relatively large set of music video clips was gathered, partly using a novel semi-automatic stimuli selection method. A subjective online assessment was then performed to select the most appropriate test material. For each video, a

[1] PetaMedia: http://www.petamedia.eu. Collaborating authors were: Mohammad Soleymani, Christian Mühl, Ashkan Yazdani and Jong-Seok Lee.

one-minute highlight was selected automatically based on audio-visual features. 32 participants took part in the experiment and their EEG and peripheral physiological signals were recorded as they watched the 40 selected music videos. For 22 participants, frontal face video was also recorded. During the experiment, participants rated each video in terms of arousal, valence, like/dislike, dominance and familiarity.

To the best of our knowledge, this database has the highest number of participants in publicly available databases for analysis of spontaneous emotions from physiological signals. In addition, it is the only affective database that uses music videos as emotional stimuli.

B.2 Related works

Recent advances in emotion recognition have motivated the creation of novel databases containing emotional expressions in different modalities. These databases mostly cover speech, visual, or audiovisual data (e.g. [112, 37, 55, 46, 54]). The visual modality includes facial expressions and/or body gestures. The audio modality covers posed or genuine emotional speech in different languages. Many of the existing visual databases include only posed or deliberately expressed emotions.

Healey[60, 61] recorded one of the first affective physiological datasets. She recorded 24 participants driving around the Boston area and annotated the dataset by the drivers' stress level. 17 Of the 24 participant responses are publicly available[2]. Her recordings include electrocardiogram (ECG), galvanic skin response (GSR) recorded from hands and feet, electromyogram (EMG) from the right trapezius muscle and respiration patterns.

To the best of our knowledge, the only publicly available multi-modal emotional databases which includes both physiological responses and facial expressions are the enterface 2005 emotional database and MAHNOB HCI[131, 139]. The former was recorded by Savran et al.[131]. This database includes two sets. The first set has electroencephalogram (EEG), peripheral physiological signals, functional near infrared spectroscopy (fNIRS) and facial videos from 5 male participants. The second dataset only has fNIRS and facial videos from 16 participants of both genders. Both databases recorded spontaneous responses to emotional images from the international affective picture system (IAPS)[85]. The MAHNOB HCI database[139] consists of two experiments. The responses including EEG, physiological signals, eye gaze, audio and facial expressions of 30 people were recorded. The first experiment was watching 20 emotional videos selected from hollywood movies and social media websites. The

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

second experiment was a tag validation experiment in which images and short videos with human actions were shown to the participants first without a tag and then with a displayed tag. The tags were either correct or incorrect and participants' agreement with the displayed tag was assessed.

B.3 Database contents

Online subjective annotation	
Number of videos	120
Video duration	1 minute affective highlight (Section 4.3.1.2)
Selection method	60 via last.fm affective tags, 60 manually selected
No. of ratings per video	14 - 16
Rating scales	Arousal Valence Dominance
Rating values	Discrete scale of 1 - 9
Physiological Experiment	
Number of participants	32
Number of videos	40
Selection method	Subset of online annotated videos with clearest responses (Section 4.3.1.3)
Rating scales	Arousal Valence Dominance Liking (how much do you like the video?) Familiarity (how well do you know the video?)
Rating values	Familiarity: discrete scale of 1 - 5 Others: continuous scale of 1 - 9
Recorded signals	32-channel 512Hz EEG 512Hz Peripheral physiological signals DV quality Face video (for 22 participants)

Table B.1: Database content summary

The contents of the database consists of two parts. First, it contains all the data from the online subjective assessment, where 120 videos were rated for valence, arousal, and dominance by 14-16 participants. Second, it contains all data gathered during the experiment. This includes recorded signal data, frontal face video for a subset of the participants and subjective ratings from the participants. Table B.1 gives an overview of the database contents. The details of the distributed files are given in the next section.

B.3.1 File Listing

The dataset is made up of the files as described in Table B.2. Each file is explained in more detail below.

File name	Format	Part	Contents
Online-ratings	Spreadsheet	Online	All individual ratings from the online self-assessment.
Video-list	Spreadsheet	Both parts	Names/YouTube links of the music videos used in the online self-assessment and the experiment, as well as statistics of the individual ratings from the online self-assessment.
Participant-ratings	Spreadsheet	Experiment	All ratings participants gave to the videos during the experiment.
Participant-questionnaire	Spreadsheet	Experiment	The answers participants gave to the questionnaire before the experiment.
Face-video	Zip-file	Experiment	The frontal face video recordings from the experiment for participants 1-22.
Data-original	Zip-file	Experiment	The original unprocessed physiological data recordings from the experiment in BioSemi .bdf format
Data-preprocessed	Zip-file	Experiment	The preprocessed physiological data recordings from the experiment in Matlab and Python(numpy) formats.

Table B.2: Files included with the database

Online-ratings This file contains all the individual video ratings collected during the online self-assessment. The file is available in the following formats:

- Open-Office Calc (*online_ratings.ods*)
- Microsoft Excel (*online_ratings.xls*)
- Comma-separated values (*online_ratings.csv*)

The ratings were collected using an online self-assessment tool as described in Chapter 4. Participants rated arousal, valence and dominance using SAM manikins on a discrete 9-point scale. In addition, participants also rated the felt emotion using the Geneva Emotion Wheel[132]. The spreadsheet has one row per individual rating, where the columns are further explained in Table B.3.

Participant-ratings This file contains all the participant video ratings collected during the experiment. The file is available in the following formats:

Column name	Description
Online_id	The video id corresponding to the same column in the video_list file.
Valence	The valence rating (integer between 1 and 9).
Arousal	The arousal rating (integer between 1 and 9).
Dominance	The dominance rating (integer between 1 and 9).
Wheel_slice	The slice selected on the emotion wheel. For some participants the emotion wheel rating was not properly recorded. In these cases, the Wheel_slice value is 0. Otherwise, the mapping of emotions on the wheel to integers given here is: 1 Pride 5 Relief 9 Sadness 13 Envy 2 Elation 6 Hope 10 Fear 14 Disgust 3 Joy 7 Interest 11 Shame 15 Contempt 4 Satisfaction 8 Surprise 12 Guilt 16 Anger
Wheel_strength	The strength selected on the emotion wheel (integer between 0=weak and 4=strong).

Table B.3: Layout of Online-ratings file.

Column name	Column contents
Participant_id	The unique id of the participant (1-32).
Trial	The trial number (i.e. the presentation order).
Experiment_id	The video id corresponding to the same column in the video_list file.
Start_time	The starting time of the trial video playback in microseconds (relative to start of experiment).
Valence	The valence rating (float between 1 and 9).
Arousal	The arousal rating (float between 1 and 9).
Dominance	The dominance rating (float between 1 and 9).
Liking	The liking rating (float between 1 and 9).
Familiarity	The familiarity rating (integer between 1 and 5). Blank if missing.

Table B.4: Layout of the Participant-ratings file.

- Open-Office Calc (*participant_ratings.ods*)
- Microsoft Excel (*participant_ratings.xls*)
- Comma-separated values (*participant_ratings.csv*)

The spreadsheet has one row per video, where the columns are further explained in Table B.4. The start_time values were logged by the Presentation software. Valence, arousal, dominance and liking were rated directly after each trial on a continuous 9-point scale using a standard mouse. SAM Manikins were used to visualize the ratings for valence, arousal and dominance. For liking (i.e. how much did you like the video?), thumbs up and thumbs down icons were used. Familiarity was rated after the end of the experiment on a 5-point integer scale (from ’never heard it before’ to ’listen to it regularly'). Familiarity ratings are unfortunately missing for participants 2, 15 and 23.

Participant-questionnaire This file contains the participants' responses to the questionnaire filled in before the experiment. The file is available in the following formats:

- Open-Office Calc (*participant_questionnaire.ods*)
- Microsoft Excel (*participant_questionnaire.xls*)
- Comma-separated values (*participant_questionnaire.csv*)

Most questions in the questionnaire were multiple-choice and speak for themselves. Participant 26 unfortunately failed to fill in the questionnaire. This questionnaire also contains the answers to the questions on the consent forms (can the data be used for research, can your imagery be published?).

Face-video.zip Face-video.zip contains the frontal face videos recorded in the experiment for the first 22 participants, segmented into trials. `sXX/sXX_trial_YY.avi` corresponds to the video for trial `YY` of subject `XX`.

For participants 3, 5, 11 and 14, one or several of the last trials are missing due to technical difficulties. These videos are in the order of presentation, so the trial numbers do not correspond to the Experiment_id columns in the **Video-list** file. The mapping between trial numbers and Experiment_ids can be found in the **participant-ratings** file.

Videos were recorded from a tripod placed behind the screen in DV PAL format using a SONY DCR-HC27E camcorder. The videos were then segmented according to the trials and transcoded to a 50 fps deinterlaced video using the h264 codec. The transcoding was done using the mencoder software with the following command:

```
mencoder sXX.dv -ss trialYY-start-second -endpos 59.05 -nosound
-of avi -ovc x264 -fps 50 -vf yadif=1:1,hqdn3d -x264encopts
bitrate=50:subq=5:8x8dct:frameref=2:bframes=3 -noskip -ofps 50 -o
sXX_trialYY.avi
```

The synchronisation of the video is accurate to approximately 1/25 second. Synchronisation was achieved by displaying a red screen before and after the experiment at the same time as a marker sent to the EEG recording PC. The onset frame of this screen was then manually marked in the video recording. Individual trial starting times were then calculated from the trial starting markers in the EEG recording.

Data-original.zip These are the original data recordings. There are 32 .bdf files (BioSemi's data format generated by the Actiview recording software), each with 48 recorded channels at 512Hz. (32 EEG channels, 12 peripheral channels, 3 unused channels and 1 status channel). The .bdf files can be read by a variety of software toolkits, including EEGLAB for Matlab and the BIOSIG toolkit. The data was recorded in two separate locations. Participants 1-22 were recorded in Twente and participant 23-32 in Geneva. Due to a different revision of the hardware, there are some minor differences in the format. First, the order of EEG channels is different for the two locations. Second, the GSR measure is in a different format for each location.

Table B.5 gives the EEG channel names (according to the 10/20 system) for both locations and the indices that can be used to convert one ordering to the other. The remaining channel numbering is the same for both locations. However, please note the GSR measurement is in different units for the two locations. The Twente GSR measurement is skin resistance in nano-Siemens, whereas the Geneva GSR measurement is skin conductance in Ohm. The conversion is given by:

$$GSR_{Geneva} = 10^9/GSR_{Twente}$$

C	NT	NG	G>T	T>G	C	NT	NG	G>T	T>G
1	Fp1	Fp1	1	1	**17**	O2	Fp2	32	30
2	AF3	AF3	2	2	**18**	PO4	AF4	31	29
3	F7	F3	4	4	**19**	P4	Fz	29	31
4	F3	F7	3	3	**20**	P8	F4	30	27
5	FC1	FC5	6	6	**21**	CP6	F8	27	28
6	FC5	FC1	5	5	**22**	CP2	FC6	28	25
7	T7	C3	8	8	**23**	C4	FC2	25	26
8	C3	T7	7	7	**24**	T8	Cz	26	32
9	CP1	CP5	10	10	**25**	FC6	C4	22	23
10	CP5	CP1	9	9	**26**	FC2	T8	23	24
11	P7	P3	12	12	**27**	F4	CP6	20	21
12	P3	P7	11	11	**28**	F8	CP2	21	22
13	Pz	PO3	16	14	**29**	AF4	P4	18	19
14	PO3	O1	13	15	**30**	Fp2	P8	17	20
15	O1	Oz	14	16	**31**	Fz	PO4	19	18
16	Oz	Pz	15	13	**32**	Cz	O2	24	17

Table B.5: EEG channel names for both locations and mapping between them. C = Channel number, NT = channel name Twente, NG = channel name Geneva, G>T = mapping of Geneva channel number to Twente channel number, T<G = inverse mapping.

Table B.6 gives the meaning of the remaining channels. The status channel contains markers sent from the stimuli presentation PC, indicating when trials start and

C	N	Channel content
33	EXG1	hEOG_1 (to the left of left eye)
34	EXG2	hEOG_2 (to the right of right eye)
35	EXG3	vEOG_1 (above right eye)
36	EXG4	vEOG_4 (below right eye)
37	EXG5	zEMG_1 (Zygomaticus Major, +/- 1cm from left corner of mouth)
38	EXG6	zEMG_2 (Zygomaticus Major, +/- 1cm from zEMG_1)
39	EXG7	tEMG_1 (Trapezius, left shoulder blade)
40	EXG8	tEMG_2 (Trapezius, +/- 1cm below tEMG_1)
41	GSR1	Galvanic skin response, left middle and ring finger
42	GSR2	Unused
43	Erg1	Unused
44	Erg2	Unused
45	Resp	Respiration belt
46	Plet	Plethysmograph, left thumb
47	Temp	Temperature, left pinky
48	Status	Status channel containing markers

Table B.6: Channel names for auxiliary channels. C = channel number, N = channel name.

end. The employed status markers are explained in Table B.7.

Status code	Event length	Event Description
1 (First occurence)	N/A	start of experiment (participant pressed key to start)
1 (Second occurence)	120000 ms	start of baseline recording
1 (Further occurences)	N/A	start of a rating screen
2	1000 ms	Video synchronization screen (before first trial, before and after break, after last trial)
3	5000 ms	Fixation screen before beginning of trial
4	60000 ms	Start of music video playback
5	3000 ms	Fixation screen after music video playback
7	N/A	End of experiment

Table B.7: Status markers used during the experiment.

Data-preprocessed These files contain a downsampled (to 128Hz), preprocessed and segmented version of the data in the following formats:

- Matlab (*data_preprocessed_matlab.zip*)
- Python (pickled numpy arrays) (*data_preprocessed_python.zip*)

This version of the data is well-suited to those wishing to quickly test a classification or regression technique without the hassle of processing all the data first. Each zip file contains 32 .dat (python) or .mat (matlab) files, one per participant. Each participant file contains two arrays, the layout of which is explained in Table B.8. Some sample code to load a python datafile is below:

```
import cPickle
x = cPickle.load(open('s01.dat', 'rb'))
```

Array name	Array shape	Array contents
data	40 x 40 x 8064	video/trial x channel x data
labels	40 x 4	video/trial x label (valence, arousal, dominance, liking)

Table B.8: Format of data arrays.

The videos are in the order of Experiment_id, so not in the order of presentation. This means the first video is the same for each participant. Table B.9 describes the channel layout and the preprocessing steps performed.

C	Ch. content	Preprocessing
1	Fp1	1. The data was downsampled to 128Hz. 2. EOG artefacts were removed as in [1]. 3. A bandpass frequency filter from 4.0-45.0Hz was applied. 4. The data was averaged to the common reference. 5. The EEG channels were reordered so that they all follow the Geneva order as above. 6. The data was segmented into 60 second trials and a 3 second pre-trial baseline removed. 7. The trials were reordered from presentation order to video (Experiment_id) order.
2	AF3	
3	F3	
4	F7	
5	FC5	
6	FC1	
7	C3	
8	T7	
9	CP5	
10	CP1	
11	P3	
12	P7	
13	PO3	
14	O1	
15	Oz	
16	Pz	
17	Fp2	
18	AF4	
19	Fz	
20	F4	
21	F8	
22	FC6	
23	FC2	
24	Cz	
25	C4	
26	T8	
27	CP6	
28	CP2	
29	P4	
30	P8	
31	PO4	
32	O2	
33	hEOG	1. The data was downsampled to 128Hz. 2. The data was segmented into 60 second trials and a 3 second pre-trial baseline removed. 3. The trials were reordered from presentation order to video (Experiment_id) order.
34	vEOG	
35	zEMG	
36	tEMG	
37	GSR	
38	Respiration belt	
39	Plethysmograph	
40	Temperature	

Table B.9: Applied preprocessing steps. For GSR, values from Twente were converted to Geneva format (Ohm). C = channel number, hEOG = horizontal EOG (hEOG$_1$ - hEOG$_2$), vEOG = vertical EOG (vEOG$_1$ - vEOG$_2$), zEMG = Zygomaticus Major EMG (zEMG$_1$ - zEMG$_2$), tEMG = Trapezius EMG (tEMG$_1$ - tEMG$_2$).

Column name	Description
Online_id	The unique id used in the online self-assessment.
Experiment_id	If this video was selected for the experiment, this lists the unique id used in the experiment. Blank if not selected.
Lastfm_tag	If this video was selected via last.fm affective tags, this lists the affective tag. Blank otherwise.
Artist	The artist that recorded the song.
Title	Title of the song.
Youtube_link	The original youtube link where the video was downloaded. Note that due to copyright restrictions we are unable to provide the videos we used.
Highlight_start	The time in seconds where the extracted one-minute highlight begins as determined by MCA analysis. For some videos, the highlight was manually overridden (for instance when a section of the song is particularly well-known).
Num_ratings	The number of volunteers that rated this video in the online self-assessment
VAQ_Estimate	The valence/arousal quadrant this video was selected for by the experimenters. For each quadrant, 15 videos were selected by last.fm and 15 by manual selection. The quadrants are: 1. high arousal, high valence. 2. low arousal, high valence. 3. low arousal, low valence. 4. high arousal, low valence.
VAQ_Online	The valence/arousal quadrant as determined by the average ratings of the volunteers in the online self-assessment. Note that these can and sometimes do differ from the estimated quadrants.
AVG_*x*, STD_*x*, Q1_*x*, Q2_*x*, Q3_*x*	Average, standard deviation and first, second and third quartile of ratings x (Valence/Arousal/Dominance) by volunteers in the online self-assessment.

Table B.10: Layout of Video-list file.

Video-list This file lists all the videos used in the online self-assessment and in the experiment in a spreadsheet. The file is available in the following formats:

- Open-Office Calc (*video_list.ods*)
- Microsoft Excel (*video_list.xls*)
- Comma-separated values (*video_list.csv*)

The spreadsheet has one row per video, where the columns are further explained in Table B.10.

Publications

[1] S. Koelstra, J.-s. Lee, A. Yazdani, 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.

[2] S. Koelstra, C. Mühl, and I. Patras. Eeg analysis for implicit tagging of video data. In *Proc. Workshop on Affective Brain-Computer Interfaces*, pages 27–32, 2009.

[3] S. Koelstra, M. Pantic, and I. Patras. A dynamic texture-based approach to recognition of facial actions and their temporal models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(11):1940–54, Nov. 2010.

[4] S. Koelstra and I. Patras. The FAST-3D Spatio-Temporal Interest Region Detector. In *Workshop on Image Analysis for Multimedia Interactive Services*, pages 242–245, 2009.

[5] S. Koelstra, A. Yazdani, M. Soleymani, C. Mühl, J. 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. *Brain Informatics*, pages 89–100, 2010.

[6] M. Soleymani, S. Koelstra, I. Patras, and T. Pun. Continuous Emotion Detection in Response to Music Videos. In *Int'l Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE (EmoSPACE) In conjunction with the IEEE FG 2011*, pages 803–808, 2011.

Bibliography

[7] K. Anderson and P. W. McOwan. A Real-Time Automated System for Recognition of Human Facial Expressions. *IEEE Trans. Systems, Man and Cybernetics*, 36(1):96–105, 2006. (Cited on pages 21 and 24.)

[8] I. Arapakis, I. Konstas, and J. M. Jose. Using facial expressions and peripheral physiological signals as implicit indicators of topical relevance. In *Proc. ACM Int'l Conf. Multimedia*, pages 461–470, 2009. (Cited on page 24.)

[9] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 16(6):345–379, 2010. (Cited on pages 118 and 119.)

[10] T. Bänziger, V. Tran, and K. Scherer. The Geneva Emotion Wheel: A tool for the verbal report of emotional reactions. *Poster presented at ISRE*, 2005. (Cited on page 12.)

[11] R. J. Barry, A. R. Clarke, S. J. Johnstone, and C. R. Brown. EEG differences in children between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, 120(10):1806–1811, 2009. (Cited on page 73.)

[12] R. J. Barry, A. R. Clarke, S. J. Johnstone, C. A. Magee, and J. A. Rushby. EEG differences between eyes-closed and eyes-open resting conditions. *Clinical Neurophysiology*, 118(12):2765–2773, 2007. (Cited on page 73.)

[13] M. Bartlett, B. Braathen, G. Littlewort-Ford, J. Hershey, I. Fasel, T. Marks, E. Smith, T. Sejnowski, and J. R. Movellan. Automatic analysis of spontaneous facial behavior: A final project report. Technical report, Machine Perception Lab, Institute for Neural Computation, University of California, San Diego, 2001. (Cited on page 22.)

[14] M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski. Measuring facial expressions by computer image analysis. *Psychophysiology*, 36(2):253–263, 1999. (Cited on pages 20 and 21.)

[15] M. S. Bartlett, G. Littlewort-Ford, M. G. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *IEEE Conf. Comp. Vision and Pattern Recognition*, pages 568–573, 2005. (Cited on pages 19, 21, 24, 111 and 112.)

[16] M. S. Bartlett, G. Littlewort-Ford, M. G. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Fully automatic facial action recognition in spontaneous behavior. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 223–230, 2006. (Cited on pages 19, 21, 24, 111 and 112.)

[17] S. Bentin, T. Allison, A. Puce, E. Perez, and G. McCarthy. Electrophysiological studies of face perception in humans. *Journal of cognitive neuroscience*, 8(6):551–565, 1996. (Cited on page 16.)

[18] B. Blankertz, G. Dornhege, C. Schäfer, R. Krepki, J. Kohlmorgen, K.-R. Müller, V. Kunzmann, F. Losch, and G. Curio. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial EEG

analysis. *IEEE transactions on neural systems and rehabilitation engineering : a publication of the IEEE Engineering in Medicine and Biology Society*, 11(2):127–31, June 2003. (Cited on page 38.)

[19] B. Blankertz, R. Tomioka, S. Lemm, M. Kawanabe, and K. Muller. Optimizing spatial filters for robust EEG single-trial analysis. *IEEE Signal Processing Magazine*, 25(1):581–607, 2008. (Cited on page 17.)

[20] P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345, 2001. (Cited on page 78.)

[21] V. Bostanov and B. Kotchoubey. The t-CWT: a new ERP detection and quantification method based on the continuous wavelet transform and Student's t-statistics. *Clinical neurophysiology*, 117(12):26–44, 2006. (Cited on pages 15, 29 and 35.)

[22] M. Bradley and P. Lang. International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings. Technical Report B-2, The Center for Research in Psychophysiology, University of Florida, Gainesville, Florida, US, 1999. (Cited on page 68.)

[23] M. M. Bradley and P. J. Lang. Measuring emotion: the Self-Assessment Manikin and the Semantic Differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, 1994. (Cited on pages 11 and 12.)

[24] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007. (Cited on page 31.)

[25] 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*, 67(8):607–627, 2009. (Cited on pages 18, 57 and 76.)

[26] Y. Chang, C. Hu, R. Feris, and M. Turk. Manifold-based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006. (Cited on pages 20, 111 and 112.)

[27] L. Chen, S. Gunduz, and M. T. Ozsu. Mixed Type Audio Classification with Support Vector Machine. In *Proc. Int. Conf. Multimedia and Expo*, pages 781–784, 2006. (Cited on page 79.)

[28] D. Chetverikov and R. Péteri. A Brief Survey of Dynamic Texture Description and Recognition. *Proc. Conf. Computer Recognition Systems*, pages 17–26, 2005. (Cited on pages 22 and 23.)

[29] I. Cohen, N. Sebe, F. Cozman, M. Cirelo, and and T. Huang. Learning Bayesian network classifiers for facial expression recognition both labeled and unlabeled data. In *IEEE Conf. Comp. Vision and Pattern Recognition*, pages 595–601, 2003. (Cited on pages 20 and 24.)

[30] H. Cole and W. J. Ray. EEG correlates of emotional tasks related to attentional demands. *International Journal of Psychophysiology*, 3(1):33–41, 1985. (Cited on page 73.)

[31] R. R. Cornelius. *The Science of Emotion. Research and Tradition in the Psychology of Emotion.* Prentice-Hall, Upper Saddle River, NJ, 1996. (Cited on page 75.)

[32] A. J. Cowell, K. Hale, C. Berka, S. Fuchs, A. Baskin, D. Jones, G. Davis, R. Johnson, R. Patch, and E. Marshall. Brainwave-Based Imagery Analysis. *Digital Human Modeling: Trends in Human Algorithms*, pages 17–27, 2008. (Cited on pages 14 and 17.)

[33] J. W. Davis and A. Bobick. The representation and recognition of human movement using temporal templates. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 928–934, 1997. (Cited on pages 89 and 90.)

[34] H. A. Demaree, E. D. Everhart, E. A. Youngstrom, and D. W. Harrison. Brain Lateralization of Emotional Processing: Historical Roots and a Future Incorporating "Dominance". *Behavioral and Cognitive Neuroscience Reviews*, 4(1):3–20, Mar. 2005. (Cited on page 54.)

[35] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005. (Cited on page 31.)

[36] G. Dornhege, B. Blankertz, and G. Curio. Speeding up classification of multichannel brain-computer interfaces: common spatial patterns for slow cortical potentials. *IEEE EMBS Conf. Neural Engineering*, pages 595–598, 2003. (Cited on page 17.)

[37] E. Douglas-Cowie, R. Cowie, and M. Schröder. A New Emotion Database: Considerations, Sources and Scope. In *Proc. ISCA Workshop on Speech and Emotion*, pages 39–44, 2000. (Cited on page 144.)

[38] 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. *Affective Computing and Intelligent Interaction*, pages 488–500, 2007. (Cited on page 102.)

[39] P. Ekman. Basic Emotions. In T. Dalgleish and M. J. Power, editors, *Handbook of Cognition and Emotion*, pages 45–61. John Wiley & Sons, Ltd, Chichester, UK, 1999. (Cited on page 10.)

[40] P. Ekman, W. V. Friesen, and J. C. Hager. *The Facial Action Coding System: A Technique for the Measurement of Facial Movement.* A Human Face, Salt Lake City, UT, 2002. (Cited on pages 18, 19, 20, 25 and 85.)

[41] 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*, 53(4):712–717, 1987. (Cited on pages 10, 18 and 75.)

[42] P. Ekman and E. L. Rosenberg. *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS).* Oxford University Press, 2005. (Cited on pages 5, 20, 25, 85 and 100.)

[43] I. Essa and A. Pentland. Coding, Analysis, interpretation, and recognition of facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):757–763, 1997. (Cited on pages 21, 22 and 24.)

[44] M. Fabiani, G. Gratton, and M. Coles. Event-related brain potentials - methods, theory, and applications. In J. Cacioppo, L. Tassinary, and G. Berntson, editors, *Handbook of psychophysiology*, pages 55 – 83. Cambridge University Press, New York, 2000. (Cited on page 15.)

[45] M. Fabiani, G. Gratton, D. Karis, and E. Donchin. The definition, identification, and reliability of measurement of the P300 component of the event-related brain potential. In P. K. Ackles, J. R. Jennings, and M. G. H. Coles, editors, *Psychophysiology*, volume 1, pages 1–78. JAI Press, Greenwich, CT, 1987. (Cited on page 15.)

[46] G. Fanelli, J. Gall, H. Romsdorfer, T. Weise, and L. Van Gool. A 3-D Audio-Visual Corpus of Affective Communication. *IEEE Trans. Multimedia*, 12(6):591–598, 2010. (Cited on page 144.)

[47] L. A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988. (Cited on pages 13 and 15.)

[48] J. R. J. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth. The World of Emotions is not Two-Dimensional. *Psychological Science*, 18(12):1050–1057, 2007. (Cited on pages 11 and 62.)

[49] J. Friedman, T. Hastie, and R. Tibshirani. Additive Logistic Regression: a Statistical View of Boosting. *The Annals of Statistics*, 28(2):337–407, 2000. (Cited on page 97.)

[50] A. Gabrielsson. Emotion perceived and emotion felt: same or different? *Musicae Scientiae*, (Special Issue 2001-2002):123–147, 2002. (Cited on page 12.)

[51] A. D. Gerson, L. C. Parra, and P. Sajda. Cortically coupled computer vision for rapid image search. *IEEE Trans. neural systems and rehabilitation engineering*, 14(2):174–179, June 2006. (Cited on pages 14, 15, 16, 17 and 29.)

[52] S. B. Gokturk, J. Y. Bouguet, C. Tomasi, and B. Girod. Model-based face tracking for viewindependent facial expression recognition. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 272–278, 2002. (Cited on pages 20 and 24.)

[53] I. Goncharova, D. J. McFarland, J. R. Vaughan, and J. R. Wolpaw. EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology*, 114(9):1580–1593, 2003. (Cited on page 73.)

[54] M. Grimm, K. Kroschel, and S. Narayanan. The Vera am Mittag German audio-visual emotional speech database. In *Proc. Int'l Conf. Multimedia and Expo*, pages 865–868, 2008. (Cited on page 144.)

[55] H. Gunes and M. Piccardi. A Bimodal Face and Body Gesture Database for Automatic Analysis of Human Nonverbal Affective Behavior. In *Proc. Int'l Conf. Pattern Recognition*, pages 1148–1153, 2006. (Cited on page 144.)

[56] H. Gunes, B. Schuller, M. Pantic, and R. Cowie. Emotion representation, analysis and synthesis in continuous space: A survey. *Proc. IEEE Conf. Face and Gesture Recognition*, pages 827–834, 2011. (Cited on page 25.)

[57] G. Guo and C. R. Dyer. Learning From Examples in the Small Sample Case - Face Expression Recognition. *IEEE Trans. Systems, Man and Cybernetics*, 35(3):477–488, 2005. (Cited on pages 21 and 24.)

[58] A. Hanjalic and L.-Q. Xu. Affective video content representation and modeling. *IEEE Trans. Multimedia*, 7(1):143–154, 2005. (Cited on page 78.)

[59] E. Harmon-Jones. Clarifying the emotive functions of asymmetrical frontal cortical activity. *Psychophysiology*, 40(6):838–848, 2003. (Cited on page 73.)

[60] J. A. Healey. *Wearable and automotive systems for affect recognition from physiology*. PhD thesis, MIT, 2000. (Cited on pages 18 and 144.)

[61] J. A. Healey and R. W. Picard. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Trans. Intelligent Transportation Systems*, 6(2):156–166, 2005. (Cited on page 144.)

[62] N. J. Hill, T. N. Lal, M. Schröder, T. Hinterberger, B. Wilhelm, F. Nijboer, U. Mochty, G. Widman, C. Elger, B. Schölkopf, A. Kübler, and N. Birbaumer. Classifying EEG and ECoG signals without subject training for fast BCI implementation: comparison of nonparalyzed and completely paralyzed subjects. *IEEE Trans. Neural Systems and Rehabilitation Engineering*, 14(2):183–186, June 2006. (Cited on page 17.)

[63] S. Hjelm and C. Browall. Brainball-Using brain activity for cool competition. In *Proceedings of NordiCHI*, pages 177–188, 2000. (Cited on page 13.)

[64] M. Hussein and T. Elsayed. Studying Facial Expressions as an Implicit Feedback in Information Retrieval Systems, 2008. (Cited on page 24.)

[65] A. Hyvärinen. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, 1999. (Cited on page 32.)

[66] J. Jiao and M. Pantic. Implicit image tagging via facial information. In *Proc. Int'l Workshop on Social Signal Processing*, pages 59–64, 2010. (Cited on page 24.)

[67] K. Kallinen and N. Ravaja. Emotion perceived and emotion felt: Same and different. *Musicae Scientiae*, 10(2):191–213, Sept. 2006. (Cited on page 12.)

[68] B. Kamousi, Z. Liu, and B. He. Classification of motor imagery tasks for brain-computer interface applications by means of two equivalent dipoles analysis. *IEEE Trans. Neural Systems and Rehabilitation Engineering*, 13(2):166–171, 2005. (Cited on page 32.)

[69] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 46–53, 2000. (Cited on pages 101 and 108.)

[70] A. Kapoor, P. Shenoy, and D. Tan. Combining Brain Computer Interfaces with Vision for Object Categorization. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 1–8, 2008. (Cited on pages 14, 17, 31 and 119.)

[71] A. Kapoor, D. Tan, P. Shenoy, and E. Horvitz. Complementary computing for visual tasks: Meshing computer vision with human visual processing. *Proc. IEEE Conf. Face and Gesture Recognition*, 18(4):1–7, Sept. 2008. (Cited on pages 14, 17 and 119.)

[72] S. Kastner and L. G. Ungerleider. Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, 23(1):315–41, Jan. 2000. (Cited on page 35.)

[73] 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*, pages 25–28, May 2009. (Cited on page 78.)

[74] J. Kierkels, M. Soleymani, and T. Pun. Queries and tags in affect-based multimedia retrieval. *IEEE Int'l Conf. Multimedia and Expo*, pages 1436–1439, 2009. (Cited on pages 14 and 18.)

[75] J. Kim and E. André. Emotion Recognition Based on Physiological Changes in Music Listening. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(12):2067–2083, 2008. (Cited on pages 18 and 76.)

[76] Y. Kim, E. Schmidt, and L. Emelle. Moodswings: A collaborative game for music mood label collection. In *Proc. Int'l Conf. Music Information Retrieval*, pages 231–236, 2008. (Cited on page 3.)

[77] W. Klimesch, P. Sauseng, and S. Hanslmayr. EEG alpha oscillations: the inhibition-timing hypothesis. *Brain Research Reviews*, 53(1):63–88, 2007. (Cited on page 73.)

[78] G. Knyazev. Motivation, emotion, and their inhibitory control mirrored in brain oscillations. *Neuroscience & Biobehavioral Reviews*, 31(3):377–395, 2007. (Cited on page 54.)

[79] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A. D. Friederici. Music, language and meaning: brain signatures of semantic processing. *Nature neuroscience*, 7(3):302–307, Mar. 2004. (Cited on page 29.)

[80] J. Kohlmorgen, G. Dornhege, M. Braun, B. Blankertz, K. R. Müller, G. Curio, K. Hagemann, A. Bruns, M. Schrauf, and W. Kincses. *Improving human performance in a real operating environment through real-time mental workload detection*, chapter 24, pages 409–422. Toward Brain-Computer Interfacing. MIT Press , Cambridge, MA, 2007. (Cited on pages 5 and 16.)

[81] Z. Koles. The quantitative extraction and topographic mapping of the abnormal components in the clinical EEG. *Electroencephalography and Clinical Neurophysiology*, 79(6):440–447, 1991. (Cited on page 38.)

[82] I. Kotsia and I. Pitas. Facial Expression Recognition in Image Sequences Using Geometric Deformation Features and Support Vector Machines. *IEEE Trans. Image Processing*, 16(1):172–187, 2007. (Cited on pages 19 and 20.)

[83] S. D. Kreibig. Autonomic nervous system activity in emotion: A review. *Biological Psychology*, 84(3):394–421, 2010. (Cited on page 76.)

[84] M. Kutas and S. Hillyard. Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, 207(4427):203–205, 1980. (Cited on page 16.)

[85] P. Lang, M. Bradley, and B. Cuthbert. International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8, University of Florida, USA, 2008. (Cited on pages 68 and 144.)

[86] P. J. Lang, M. K. Greenwald, M. M. Bradley, A. O. Hamm, and M. M. Bradely. Looking at pictures: Affective, facial, visceral, and behavioral reactions. *Psychophysiology*, 30(3):261–273, May 1993. (Cited on pages 18, 56, 57 and 75.)

[87] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions From Movies. In *IEEE Conf. Comp. Vision and Pattern Recognition*, pages 1–8, 2008. (Cited on page 31.)

[88] Lazar, N. Combining Brains: A Survey of Methods for Statistical Pooling of Information. *NeuroImage*, 16(2):538–550, June 2002. (Cited on page 72.)

[89] J.-S. Lee and C. H. Park. Robust audio-visual speech recognition based on late integration. *IEEE Trans. Multimedia*, 10(5):767–779, 2008. (Cited on page 78.)

[90] R. Leeb, R. Scherer, C. Keinrath, G. Pfurtscheller, I. K. Discovery, D. Friedman, G. Street, F. Y. Lee, H. Bischof, and M. Slater. *Combining BCI and Virtual Reality: Scouting Virtual Worlds*, chapter 23, pages 394–408. Toward Brain-Computer Interfacing. MIT Press , Cambridge, MA, 2007. (Cited on page 16.)

[91] D. Li, I. K. Sethi, N. Dimitrova, and T. McGee. Classification of general audio data for content-based retrieval. *Pattern Recognition letters*, 22(5):533–544, 2001. (Cited on pages 78 and 79.)

[92] X. Liao, D. Yao, D. Wu, and C. Li. Combining spatial filters for the classification of single-trial EEG in a finger movement task. *IEEE Trans. Bio-medical Engineering*, 54(5):821–31, May 2007. (Cited on pages 17 and 40.)

[93] C. L. Lisetti and F. Nasoz. Using noninvasive wearable computers to recognize human emotions from physiological signals. *EURASIP Journal on Applied Signal Processing*, 11(1):1672–1687, 2004. (Cited on page 18.)

[94] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan. Dynamics of facial expression extracted automatically from video. *Image and Vision Computing*, 24(6):615–625, 2006. (Cited on pages 19, 21, 24, 111 and 112.)

[95] T. M. Loughin. A systematic comparison of methods for combining p-values from independent tests. *Computational Statistics & Data Analysis*, 47(3):467–485, 2004. (Cited on pages 53 and 72.)

[96] L. Lu, H. Jiang, and H. Zhang. A robust audio classification and segmentation method. In *Proc. ACM Int. Conf. Multimedia*, pages 203–211, 2001. (Cited on pages 78 and 79.)

[97] Z. Lu, W. Xie, J. Pei, and J. Huang. Dynamic Texture Recognition by Spatio-Temporal Multiresolution Histograms. In *IEEE Workshop on Motion and Video Computing*, volume 2, pages 241–246, 2005. (Cited on page 22.)

[98] S. Lucey, A. B. Ashraf, and J. F. Cohn. Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face. In K. Delac and M. Grgic, editors, *Face Recognition*, pages 275–286. I-Tech Education and Publishing, Vienna, 2007. (Cited on pages 21, 111 and 112.)

[99] S. Luck. *An introduction to the event-related potential technique.* MIT Press Cambridge, MA, USA, 2005. (Cited on page 15.)

[100] M. Mandel and D. Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165, 2008. (Cited on page 3.)

[101] 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*, 76(14):1089–1093, 1995. (Cited on pages 57 and 75.)

[102] D. McDuff and R. E. Kaliouby. Affect valence inference from facial action unit spectrograms. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 17–24, 2010. (Cited on page 26.)

[103] J. D. Morris. Observations: SAM: The Self-Assessment Manikin; An Efficient Cross-Cultural Measurement of Emotional Response. *Journal of Advertising Research*, 35(8):38–63, 1995. (Cited on pages 11 and 12.)

[104] E. Niedermeyer and F. Da Silva. *Electroencephalography: basic principles, clinical applications, and related fields.* Lippincott Williams & Wilkins, 2005. (Cited on pages 13 and 16.)

[105] J. Onton and S. Makeig. High-frequency Broadband Modulations of Electroencephalographic Spectra. *Frontiers in Human Neuroscience*, 3(61):1–18, 2009. (Cited on page 73.)

[106] G. Orgs, K. Lange, J.-H. Dombrowski, M. H. Is, and M. Heil. Is conceptual priming for environmental sounds obligatory? *Int'l Journal of Psychophysiology*, 65(2):162–166, 2007. (Cited on pages 16, 29 and 35.)

[107] M. Pantic and M. S. Bartlett. *Machine Analysis of Facial Expressions.* I-Tech Education and Publishing, Vienna, Austria, 2007. (Cited on pages 6, 18 and 24.)

[108] M. Pantic and I. Patras. Detecting Facial Actions and their Temporal Segments in Nearly Frontal-View Face Image Sequences. In *Proc. IEEE Conf. Systems, Man and Cybernetics*, volume 4, pages 3358–3363, 2005. (Cited on pages 19, 20, 24, 111 and 112.)

[109] M. Pantic and I. Patras. Dynamics of Facial Expressions - Recognition of Facial Actions and their Temporal Segments from Face Profile Image Sequences. *IEEE Trans. Systems, Man and Cybernetics*, 36(2):433–449, 2006. (Cited on pages 19, 20, 24, 100 and 110.)

[110] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions - the state of the art. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000. (Cited on pages 18 and 24.)

[111] M. Pantic and L. J. M. Rothkrantz. Facial Action Recognition for Facial Expression Analysis from Static Face Images. *IEEE Trans. Systems, Man and Cybernetics*, 34(3):1449–1461, 2004. (Cited on pages 19, 20, 24, 98 and 110.)

[112] M. Pantic, M. F. Valstar, R. Rademaker, and L. Maat. Web-based Database for Facial Expression Analysis. In *Proc. IEEE Conf. Multimedia and Expo*, pages 317–321, 2005. (Cited on pages 31, 101 and 144.)

[113] L. C. Parra, S. Member, C. Christoforou, A. D. Gerson, M. Dyrholm, A. Luo, M. Wagner, M. G. Philiastides, and P. Sajda. Spatio-temporal linear decoding of brain state: Application to performance augmentation in high-throughput tasks. *IEEE Signal Processing Magazine*, 25(1):95–115, 2008. (Cited on page 16.)

[114] W. G. Parrott. *Emotions in Social Psychology: Essential Readings*. Psychology Press, Philadelphia, 2001. (Cited on pages 10 and 60.)

[115] A. Pfefferbaum and J. Ford. ERPs to stimuli requiring response production and inhibition: effects of age, probability and visual noise. *Electroencephalography and Clinical Neurophysiology*, 71(1):55–63, 1988. (Cited on page 15.)

[116] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. da Silva. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage*, 31(1):153–159, 2006. (Cited on page 17.)

[117] R. W. Picard. *Affective Computing*. MIT Press Cambridge, MA, USA, 1997. (Cited on page 78.)

[118] D. Plass-Oude Bos, B. Reuderink, B. Laar, H. Gürkök, C. Mühl, M. Poel, A. Nijholt, D. Heylen, and D. S. Tan. Brain-Computer Interfacing and Games. In D. S. Tan and A. Nijholt, editors, *Brain-Computer Interfaces*, Human-Computer Interaction Series, pages 149–178. Springer London, London, 2010. (Cited on page 14.)

[119] R. Plutchik. The nature of emotions. *American Scientist*, 89(4):344–350, 2001. (Cited on page 10.)

[120] R. Polana and R. Nelson. Temporal texture and activity recognition. *Computational Image and Vision*, 9(1):87–124, 1997. (Cited on page 22.)

[121] F. Popescu, S. Fazli, Y. Badower, B. Blankertz, and K. R. Müller. Single trial classification of motor imagination using 6 dry EEG electrodes. *PLoS ONE*, 2(7):e637, 2007. (Cited on page 17.)

[122] R Kaliouby. Real-time inference of complex mental states from facial expressions and head gestures. In B. Kisačanin, V. Pavlović, and T. Huang, editors, *Real-time vision for human-computer interaction*, pages 181–200. Springer US, 2005. (Cited on page 25.)

[123] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio. Basic emotions are associated with distinct patterns of cardiorespiratory activity. *International Journal of Psychophysiology*, 61(1):5–18, 2006. (Cited on page 76.)

[124] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Trans. Circuits, Systems and Video Technology*, 15(1):52–64, 2005. (Cited on page 78.)

[125] F. Renkens and J. Millán. Brain-actuated control of a mobile platform. *Int. Conf. Simulation of Adaptive Behavior*, 2002. (Cited on page 5.)

[126] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid Registration Using Free-Form Deformations: Application to Breast MR Images. *IEEE Trans. medical imaging*, 18(8):712–721, 1999. (Cited on pages 90, 91 and 92.)

[127] J. Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178, 1980. (Cited on pages 11 and 26.)

[128] P. Saisan, G. Doretto, Y. Wu, and S. Soatto. Dynamic texture recognition. In *IEEE Conf. Comp. Vision and Pattern Recognition*, volume 2, pages 58–63, 2001. (Cited on page 22.)

[129] D. Sammler, M. Grigutsch, T. Fritz, and S. Koelsch. Music and emotion: Electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, 44(2):293–304, 2007. (Cited on page 46.)

[130] D. Sander, D. Grandjean, and K. R. Scherer. A systems approach to appraisal mechanisms in emotion. *Neural Networks*, 18(4):317–352, 2005. (Cited on page 75.)

[131] 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 *Proc. eNTERFACE 2006 Workshop*, pages 69–80, 2006. (Cited on page 144.)

[132] K. R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, 2005. (Cited on pages 11, 13 and 146.)

[133] A. Schlögl, C. Keinrath, D. Zimmermann, R. Scherer, R. Leeb, and G. Pfurtscheller. A fully automated correction method of EOG artifacts in EEG recordings. *Clinical Neurophysiology*, 118(1):98–104, 2007. (Cited on page 76.)

[134] L. Schmidt and L. Trainor. brain electrical activity (EEG) distinguishes valence and intensity of musical emotions. *Cognition and emotion*, 15(4):487–500, 2001. (Cited on page 46.)

[135] E. Schubert. The influence of emotion, locus of emotion and familiarity upon preference in music. *Psychology of Music*, 35(3):499–515, July 2007. (Cited on page 62.)

[136] N. Sebe, I. Cohen, and T. Huang. Multimodal emotion recognition. *Handbook of Pattern Recognition and Computer Vision*, pages 981–256, 2005. (Cited on page 119.)

[137] 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*, 3(2):235–254, 2009. (Cited on pages 18 and 76.)

[138] M. Soleymani, J. J. M. Kierkels, G. Chanel, and T. Pun. A Bayesian framework for video affective representation. In *Proc. Int'l Conf. Affective Computing and Intelligent interaction*, pages 1–7, 2009. (Cited on page 61.)

[139] 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, in press.* (Cited on pages 115, 119 and 144.)

[140] T. Solis-Escalante, G. Müller-Putz, and G. Pfurtscheller. Overt foot movement detection in one single Laplacian EEG derivation. *Journal of Neuroscience Methods*, 175(1):148–153, 2008. (Cited on page 54.)

[141] H. Tao and T. Huang. Explanation-based facial motion tracking using a piecewise bezier volume deformation model. In *IEEE Conf. Comp. Vision and Pattern Recognition*, volume 1, pages 611–617, 1999. (Cited on pages 20 and 21.)

[142] Y. L. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(2):97–115, 2001. (Cited on pages 20, 22 and 24.)

[143] Y. L. Tian, T. Kanade, and J. F. Cohn. Evaluation of gabor-wavelet-based facial action unit recognition in image sequences of increasing complexity. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 218–223, 2002. (Cited on pages 20, 22 and 24.)

[144] Y. L. Tian, T. Kanade, and J. F. Cohn. Facial Expression Analysis. In S. Z. Li and A. K. Jain, editors, *Handbook of Face Recognition*, pages 247–276. Springer, New York, USA, 2005. (Cited on pages 20, 22 and 24.)

[145] M. E. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*, 1:211–244, 2001. (Cited on pages 61 and 82.)

[146] R. Tomioka, K. Aihara, and K. Muller. Logistic regression for single trial EEG classification. *Advances in Neural Information Processing Systems*, 19:1377–1384, 2007. (Cited on page 17.)

[147] Y. Tong, W. Liao, and Q. Ji. Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(10):1683–1699, 2007. (Cited on pages 19, 20, 24, 111 and 112.)

[148] H. Touyama. Photo Data Retrieval via P300 Evoked Potentials. *IEICE Trans. Information and Systems*, 91(8):2212—-2213, 2008. (Cited on pages 15 and 17.)

[149] P. Valdez and A. Mehrabian. Effects of color on emotions. *J. Exp. Psychol. Gen.*, 123(4):394–409, 1994. (Cited on page 78.)

[150] M. Valstar, I. Patras, and M. Pantic. Facial Action Unit Detection using Probabilistic Actively Learned Support Vector Machines on Tracked Facial Point Data. In *IEEE Conf. Comp. Vision and Pattern Recognition*, volume 3, page 76, 2005. (Cited on page 24.)

[151] M. F. Valstar and M. Pantic. Fully automatic facial action unit detection and temporal analysis. *IEEE Conf. Comp. Vision and Pattern Recognition*, 3:149, 2006. (Cited on pages 24, 101, 109, 110, 111 and 112.)

[152] M. F. Valstar and M. Pantic. Combined Support Vector Machines and Hidden Markov Models for Modeling Facial Action Temporal Dynamics. *Lecture Notes in Computer Science*, 4796:118–127, 2007. (Cited on pages 19, 20, 24, 111 and 112.)

[153] M. F. Valstar, M. Pantic, and I. Patras. Motion history for facial action detection from face video. In *Proc. IEEE Conf. Systems, Man and Cybernetics*, pages 635–640, 2004. (Cited on pages 19, 21, 22, 23, 24, 86, 111 and 112.)

[154] C. van Petten and H. Rheinfelder. Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, 33(4):485–508, 1995. (Cited on page 29.)

[155] A. Vinciarelli, N. Suditu, and M. Pantic. Implicit Human-centered Tagging. *IEEE Int'l Conf. Multimedia and Expo*, pages 1428–1431, 2009. (Cited on pages 2 and 28.)

[156] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. Conf. Human factors in computing systems*, pages 319–326. ACM Press, Apr. 2004. (Cited on page 3.)

[157] L. von Ahn, R. Liu, and M. Blum. Peekaboom. In *Proc. Conf. Human factors in computing systems*, pages 55–64. ACM Press, Apr. 2006. (Cited on page 3.)

[158] D. Vukandinovic and M. Pantic. Fully automatic facial feature point detection using Gabor feature based boosted classifiers. In *Proc. IEEE Conf. Systems, Man and Cybernetics*, volume 2, pages 1692–1698, 2005. (Cited on pages 20, 21 and 87.)

[159] H. L. Wang and L.-F. Cheong. Affective understanding in film. *IEEE Trans. Circuits, Systems and Video Technology*, 16(6):689–704, 2006. (Cited on pages 61 and 78.)

[160] J. Wang and Y. Gong. Recognition of multiple drivers' emotional state. In *Proc. Int'l Conf. Pattern Recognition*, pages 1–4, 2008. (Cited on pages 18, 57 and 76.)

[161] Z. Wen and T. Huang. Capturing subtle facial motions in 3d face tracking. In *Proc. Int'l. Conf. Computer Vision*, volume 2, pages 1343–1350, 2003. (Cited on pages 20, 21 and 24.)

[162] J. Whitehill and C. W. Omlin. Haar features for FACS AU recognition. In *Proc. IEEE Conf. Face and Gesture Recognition*, pages 97–101, 2006. (Cited on pages 111 and 112.)

[163] A. C. Williams. Facial expression of pain: An evolutionary account. *Behavioral and Brain Sciences*, 25(4):439–488, 2002. (Cited on pages 25 and 85.)

[164] C. K. I. Williams and C. E. Rasmussen. *Gaussian Processes for Machine Learning*. MIT Press, 2007. (Cited on page 82.)

[165] J. R. Wolpaw and D. J. McFarland. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proc. National Academy of Sciences*, 101(51):17849–17854, 2004. (Cited on pages 5, 13 and 16.)

[166] C. D. Woody. Characterization of an adaptive filter for the analysis of variable latency neuroelectric signals. *Medical and biological engineering and computing*, 5(6):539–554, 1967. (Cited on page 15.)

[167] J. Xiao, S. Baker, I. Matthews, and T. Kanade. Real-time Combined 2D+3D Active appearance models. In *IEEE Conf. Comp. Vision and Pattern Recognition*, volume 2, pages 535–542, 2004. (Cited on pages 20 and 21.)

[168] Q. Xu, H. Zhou, Y. Wang, and J. Huang. Fuzzy support vector machine for classification of EEG signals using wavelet-based features. *Medical engineering & physics*, 31(7):858–65, 2009. (Cited on page 17.)

[169] A. Yazdani, J.-S. Lee, and T. Ebrahimi. Implicit emotional tagging of multimedia using EEG signals and brain computer interface. In *Proceedings of the first SIGMM workshop on Social media - WSM '09*, pages 81–88, Oct. 2009. (Cited on pages 14 and 18.)

[170] L. Yu and H. Liu. Efficient Feature Selection via Analysis of Relevance and Redundancy. *Journal of Machine Learning Research*, 5:1205–1224, 2004. (Cited on page 58.)

[171] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009. (Cited on pages 18 and 119.)

[172] Y. Zhang and Q. Ji. Active and Dynamic Information Fusion for Facial Expression Understanding from Image Sequence. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 27(5):699–714, 2005. (Cited on pages 20, 21 and 24.)

[173] G. Zhao and M. Pietikäinen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(6):915–928, 2007. (Cited on pages 19, 23, 24 and 94.)