UNIVERSITE DE GENEVE FACULTE DES SCIENCES

Département d'Informatique Professeur Thierry Pun

Emotion assessment for affective computing based on brain and peripheral signals

THESE

présenté à la faculté des sciences de l'Université de Genève

pour obtenir le grade de Docteur ès sciences, mention informatique

par

Guillaume Chanel

de

Crozet (France)

Thèse N° XXXX

GENEVE

2009

Résumé

Les Interfaces Homme-Machine actuelles manquent « d'intelligence émotionnelle » : elles ne sont pas capables d'identifier les émotions humaines et de prendre cette information en compte pour choisir les actions à exécuter. Le but de l'informatique affective ou *affective computing* est de combler ce manque en détectant les indices émotionnels se produisant durant l'interaction avec la machine et en synthétisant les réponses émotionnelles adéquates. Durant ces dernières années, la plupart des études s'intéressant à la détection d'émotions se sont concentrées sur l'analyse des expressions faciales et de la parole pour déterminer l'état émotionnel d'une personne. Toutefois, il existe d'autres sources d'information émotionnelle provenant en particulier de l'analyse des signaux physiologiques. Comparés à l'analyse des expressions faciales, ces signaux n'ont été que peu étudiés bien qu'ils aient plusieurs avantages; par exemple, il est plus difficile de simuler une réponse physiologique qu'une expression faciale.

Cette thèse traite de l'utilisation de deux types d'activités physiologiques pour détecter les émotions dans le cadre de *l'affective computing* : l'activité du système nerveux central (le cerveau) et l'activité du système nerveux périphérique. L'activité centrale est mesurée par l'utilisation d'électro-encéphalogrammes (EEGs). L'activité périphérique est mesurée au moyen des capteurs suivants : un capteur de réponse cutané galvanique (GSR) permettant de mesurer la transpiration ; une ceinture de respiration mesurant les mouvements de l'abdomen ; un pléthysmographe enregistrant les variations de volume sanguin (*blood volume pulse -BVP*), et un capteur de température servant à mesurer la température cutanée du doigt.

L'espace *valence-arousal* (hedonicité-excitation) est choisi pour représenter les émotions car, issu de la théorie cognitive des émotions, il est suffisamment général pour être utilisable pour plusieurs applications. Certaines régions de l'espace valence-arousal sont utilisées pour définir des classes émotionnelles. Afin de reconnaître ces classes à partir des signaux physiologiques, il est nécessaire de trouver un modèle informatique qui associe une classe à une activité physiologique donnée. Les algorithmes de reconnaissance de forme et d'apprentissage artificiel sont utilisés pour déterminer ce modèle.

Trois protocoles sont conçus pour enregistrer les signaux physiologiques lors de stimulations émotionnelles de différents types: regarder des images, se remémorer des épisodes émotionnels passés, et jouer à un jeu vidéo à différents niveaux de difficulté. Pour chaque stimulation, des caractéristiques sont extraites des signaux EEG et périphériques. Les signaux EEG sont caractérisés par leur énergie dans des bandes de fréquences en relation avec les processus émotionnels. De plus, l'information mutuelle calculée entre chaque paire d'électrodes est proposée comme un nouvel ensemble de caractéristiques pour détecter les émotions. Pour ce qui est des signaux périphériques, les caractéristiques sont calculées sur la base de la littérature

existante et discutées par rapport aux aspects temporels. Plusieurs classifieurs (*naïve Bayes*, analyse discriminante, machines à vecteurs de support, *relevance vector machines*) sont ensuite entraînés sur la base de données des caractéristiques. La performance des ces algorithmes est évaluée par rapport à leur précision pour des classifications intra- (deux premiers protocoles) et inter- (troisième protocole) participants. Des méthodes de sélection de caractéristiques sont employées pour trouver les caractéristiques les plus intéressantes en vue de la détection d'émotions ainsi que pour réduire la taille de l'espace des caractéristiques. Finalement, la fusion des informations du système nerveux central et périphérique est analysée tant au niveau des caractéristiques (avant classification) qu'au niveau décisionnel (après classification).

Les résultats montrent que les signaux EEG sont utilisables pour détecter les émotions car la précision obtenue par les algorithmes de classification sur les données EEG est bien supérieure à la précision aléatoire. La meilleure précision moyenne obtenue pour reconnaître trois classes est de 68% et autour de 80% pour deux classes (moyennes de la précision pour plusieurs participants). Les caractéristiques calculées en utilisant l'information mutuelle entre les paires d'électrodes sont également utilisables pour détecter les émotions comme le montrent les résultats obtenus (par exemple 62% de précision pour reconnaître trois états émotionnels). De plus la précision de classification en utilisant les signaux EEG est supérieure à celle obtenue avec les signaux périphériques lorsque leurs caractéristiques sont calculées sur des fenêtres temporelles relativement courtes (6 à 30 secondes). La fusion des informations périphériques et centrales au niveau décisionnel augmente significativement la précision de détection (de 3% à 7%), ce qui encourage la fusion avec d'autres sources d'information émotionnelle comme les expressions faciales ou la parole. Une application des méthodes développées est proposée comme exemple d'une interface affective. Celle-ci adapterait automatiquement le niveau de difficulté d'un jeu vidéo en fonction de l'état émotionnel du joueur.

Summary

Current Human-Machine Interfaces (HMI) lack of "emotional intelligence", i.e. they are not able to identify human emotional states and take this information into account to decide on the proper actions to execute. The goal of affective computing is to fill this lack by detecting emotional cues occurring during Human-Computer Interaction (HCI) and synthesizing emotional responses. In the last decades, most of the studies on emotion assessment have focused on the analysis of facial expressions and speech to determine the emotional state of a person. Physiological activity also includes emotional information that can be used for emotion assessment but has received less attention despite of its advantages (for instance it can be less easily faked than facial expressions).

This thesis reports on the use of two types of physiological activities to assess emotions in the context of affective computing: the activity of the central nervous system (i.e. the brain) and the activity of the peripheral nervous system. The central activity is monitored by recording electro-encephalograms (EEGs). The peripheral activity is assessed by using the following sensors: a Galvanic Skin Response (GSR) sensor to measure sudation; a respiration belt to measure abdomen expansion; a plethysmograph to record blood volume pulse (BVP); and a temperature sensor to record finger temperature.

The valence-arousal space is chosen to represent emotions since it originates from the cognitive theory of emotions and is general enough to be usable for several applications. Several areas in the valence-arousal space are used as ground-truth classes. In order to automatically detect those classes from physiological signals, it is necessary to find a computational model that maps a given physiological pattern to one of the chosen classes. Pattern recognition and machine learning algorithms are employed to infer such a model.

Three protocols that use different emotion elicitation methods (images, recall of past emotional episodes and playing a video game at different difficulty levels) are designed to gather physiological signals during emotional stimulations. For each emotional stimulations, features are extracted from the EEG and peripheral signals. For EEG signals, energy features that are known to be related to emotional processes are computed. Moreover, the Mutual Information (MI) computed between all pairs of electrodes is proposed as a new set of features. For peripheral signals, the computed features are chosen based on the review of the literature and are discussed relatively to temporal aspects. Several classifiers (Naïve Bayes, Discriminant Analysis, Support Vector Machines – SVM and Relevance Vector Machines - RVM) are then trained on the resulting databases. The performance of these algorithms is evaluated according to the obtained accuracy for both intra (two first protocols) and inter (third protocol) participant classification. Feature selection methods are also employed to find the most relevant features for emotion

assessment and reduce the size of the original feature spaces. Finally, fusion of the central and peripheral information at the decision and feature level is analyzed.

The results show that EEG signals are usable for emotion assessment since the classification accuracy obtained with EEG features is much higher than the random level. The best accuracy for the detection of three emotional classes is 68% and around 80% for two classes, when averaged across participants. The effectiveness of the MI feature set for emotion assessment is also demonstrated by the classification accuracies (for instance 62% to detect three emotional classes). Moreover, the accuracy obtained for classification based on EEG features is found to be higher than based on peripheral features when the features are computed on short time periods (6 to 30 seconds). Fusion of the peripheral and EEG features at the decision level significantly increased the accuracy (by an amount of 3% to 7%), encouraging further fusion with other sources of emotional information (facial expressions, speech, etc.). An application of the developed methods which automatically adapts the difficulty of games based on emotion assessment is proposed as an example of affective HMI.

Table of contents

Table of symbols

$x(n)$	Value of the signal x at discrete time n
μ_x	Mean of signal x
σ_x	Standard deviation of signal x
δ_x	Mean of the derivative of signal x
Min_x, Max_x	Minimum and maximum values of signal x
N	Number of trials and number of instances in the feature set **F**
N_e	Number of EEG electrodes
N_s	Number of samples in one of the signals of a trial
T	Duration of a trial
$f_{<1>}^{<2>}$	Feature having name <2> and computed from signal <1>
F	Size of the feature space, number of features in a feature vector $\mathbf{f}_i$
$\mathbf{f}_i$	Feature vector containing the values of some $f_{<1>}^{<2>}$ features for the trial i (line of the feature set **F**)
$\tilde{\mathbf{f}}_i$	Vector containing the values of the feature i for several trials (column of the feature set **F**)
F	Matrix containing all the $\mathbf{f}_i$ instances
C	Number of classes
ω_i	Class i
y_i	True label associated to trial i, $y_i \in \{\omega_1, \ldots, \omega_C\}$
$\hat{y}_i$	Estimated label of trial i, $\hat{y}_i \in \{\omega_1, \ldots, \omega_C\}$
y	Column vector of true labels y_i
$\hat{\mathbf{y}}$	Column vector of estimated labels $\hat{y}_i$

Acronyms

ANOVA	ANalysis Of VAriance
ANS	Autonomic Nervous System
BCI	Brain-Computer Interface
BDI	Belief, Desire and Intensions
BP	Blood Pressure
BVP	Blood Volume Pulse
CNS	Central Nervous System
DBP	Diastolic Blood Pressure
ECoG	Electro-Cortico-Graphy
EDA	Electrodermal Activity
EDR	Electrodermal Response
EEG	Electro-encephalogram
EMG	Electromyogram
FCBF	Fast Correlation Based Filter
FFT	Fast Fourier Transform
fMRI	functional Magnetic Resonance Imagery
fNIRS	functional Near InfraRed Spectroscopy
GSR	Galvanic Skin Response
HCI	Human-Computer Interaction
HF	High Frequency (frequency band of interest in the HR power spectrum)
HMI	Human-Machine Interfaces
HR	Heart Rate
HRV	Heart Rate Variability
IADS	International Affective Digitized Sound system
IAPS	International Affective Picture System
KNN	K-Nearest-Neighboor
LDA	Linear Discriminant Analysis
LF	Low Frequency (frequency band of interest in the HR power spectrum)
MEG	Magneto-encephalogram
MI	Mutual Information
MSE	Mean Square Error

OCC	Ortony, Clore and Collins
PET	Positron Emission Tomography
PNS	Peripheral Nervous System
QDA	Quadratic Discriminant Analysis
RVM	Relevance Vector Machine
SAM	Self Assessment Manikin
SBP	Systolic Blood Pressure
SEC	Sequential Evaluation Check
SFFS	Sequential Floating Forward Selection
SNS	Somatic Nervous System
SPECT	Single Photon Emission Computed Tomography
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
VLF	Very Low Frequency (frequency band of interest in the HR power spectrum)

Chapter 1 Introduction

1.1 Emotions and machines

Emotions are part of any natural communication involving humans. They can be expressed either verbally through emotional vocabulary, or by expressing non-verbal cues such as intonation of voice, facial expressions and gestures. In this context, decoding emotional cues is essential to correctly interpret a message. For instance, aggressively shouting “Shut up” at someone does not have the same meaning as saying it while laughing and smiling. In the second case, one could interpret that the speaker does not really want his interlocutor to shut up but rather that he found his remark funny and at the same time at the limit of social acceptance. However, disambiguating and enriching communication is not the only role of emotions. It has also been found that they play a key role in decision making and that they are strongly related to tendencies to action. This discovery is certainly a reason for the importance they have nowadays in neuromarketing that aim at uncovering the cerebral mechanisms leading to the purchase decision. In his bestseller book, D. Goleman [1] does not hesitate to talk about “emotional intelligence”, i.e. the ability of someone to understand, communicate and manage his / her own emotions as well as his / her capacity to understand and feel emotions of the others. According to him those competences are as important as logic and verbal skills to “succeed in life”. Of course the definition of what is “success in life” could be discussed, but nevertheless, the success of Goleman’s book and of many other books on emotions clearly demonstrates the importance that people attach to this topic in our society.

Despite the importance of emotions in communication, many Human-Machine Interfaces (HMI) completely lack “emotional intelligence”. The reaction of users in response to this shortage is either to be more and more frustrated by the machine or to invent new ways to communicate their emotions using the limited interfaces at their disposition. The number of internet videos showing a person getting upset in front of his / her computer is a clear indicator of the first reaction, while the use of emoticons in chatting sessions and mails is an example of the second. In the last decades, researchers and manufacturers started to take into account the emotions of users in the design of products, adding the concept of user experience to the one of usability. For computers this is part of what is called “human-centered computing”, where the goal is to have interfaces that are made for the human: they should be intuitive, easily handled, and use as most as possible the standard communication channels used by humans (speech, gestures, etc.). Current interfaces are far from this objective since they are complex and require training to be used. Even if including the user experience in the design of software is a step toward the improvement of HMI, it is still far from having machines that can react properly to a sudden emotional reaction of the user.

In order to obtain such an “emotional machine” it is necessary to include emotions in the human-machine communication loop (see Figure 1.1). This is what is defined as “affective computing”. According to Picard [2], affective computing “proposes to give computers the ability to recognize [and] express […] emotions”. Synthetic expression of emotions can be achieved by enabling avatars or simpler agents to have facial expressions, different tones of voice, and empathic behaviors [3, 4]. Detection of human emotions can be realized by monitoring and interpreting the different cues that are given in both verbal and non-verbal communication. However, a system that is “emotionally intelligent” should not only detect and express emotions but it should also take the proper action in response to the detected emotion. Expressing the adequate emotion is thus one of the outputs of this decision making. The proper action to take is dependent on the application but examples of reactions could be to provide help in the case the user feels helpless or to lower task demand in the case he / she is highly stressed. Many applications are detailed in chapter 1.2.2. Accurately assessing emotions is thus a critical step toward affective computing since this will determine the reaction of the system. The present work will focus on this first step of affective computing by trying to reliably assess emotion from several emotional cues.

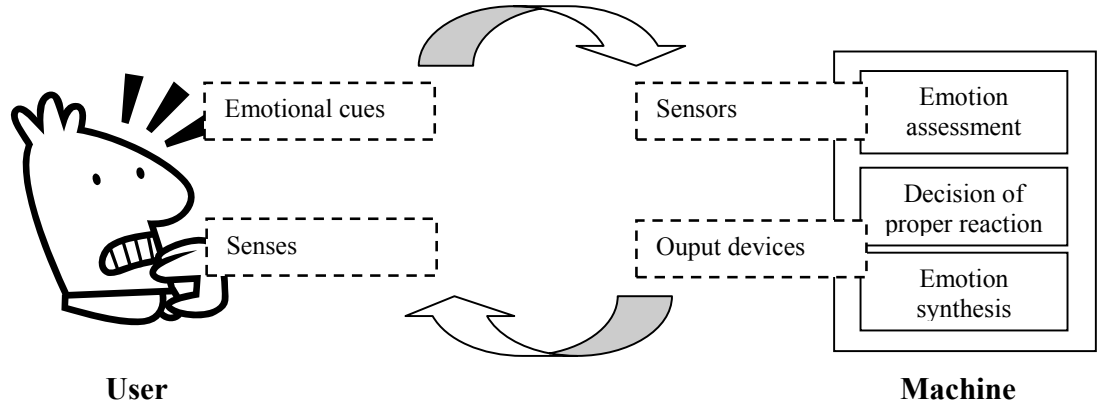


Figure 1.1. Including emotions in the human-machine loop

1.2 Emotion assessment

1.2.1 Multimodal expression of emotion

A modality is defined as a path used to carry information for the purpose of interaction. In HMI, there are two possible approaches to define a modality: from the machine point of view and from the user point of view. On the machine side, a modality refers to the processes that generate information to the physical world and interpret information from it. It is thus possible to distinguish between input and output modalities and to associate them to the corresponding communication device. A keyboard, a mouse and a pad with the associated information processes are typical input modalities, while a text, an image and music presented on screens and speakers

are typical output modalities. From the user perspective, a modality corresponds to one of the senses used for interpreting (resp. transmitting) messages from (resp. to) the machine. In psychology a modality is similarly defined as one of the five sensorial categories: touch, vision, hearing, taste and smell. In this thesis, we choose to use the word modality in a wide sense, including both of those aspects.

Emotions can be expressed trough several channels and modalities. Facial expressions, gestures, postures, speech and intonation of voice have already been cited in the introduction and are certainly those that are the most obvious. However, emotional information can also be found in many other modalities. For instance it has been shown that there are different physiological activations of the body corresponding to different emotions [5-7]. Examples of those activations and inactivation are paralysis of muscles in case of fear, increase of heart rate and sudation for aroused emotions. They are generally less perceivable by people, unless an observer is close enough to the person that feels the emotion. However these reactions could be easily recorded using specific sensors. Some of those physiological changes can also be directly perceived such as a strong increase of blood pressure that would lead to a blush of the cheeks.

All those modalities could be used for emotion recognition, but so far, most of the studies concerning machine based emotion assessment have used facial expressions and/or speech. However, we believe that physiological signals have several advantages compared to video and sound:

- the sensors used to record those signals are closer to the body since they are generally directly placed on the user, reducing potential sources of noise and problems due to the unavailability of the signal (user not turning the head in front of the camera or not speaking);
- they have very good time responses, for instance muscle activity can be detected earlier by using electromyography (EMG) than by using a camera;
- they are harder to control, it is thus harder to fake an emotion;
- in the case of impaired users that cannot move facial muscles or express themselves, many physiological signals, such as brain waves, are still usable for emotion assessment.

Since emotions are clearly multimodal processes it is necessary to perform fusion of different modalities to reliably assess them. This could be done by fusing different physiological signals but also by fusion with other modalities such as postures, gestures and speech. Fusing those different sources could help to increase the accuracy of emotion assessment; it is also advantageous when some signals are not available because of technical problems like

disconnected or damaged sensors. Moreover, the synchronization between the activation of different modalities could also be a potential source of emotional information.

The goal of this thesis is to investigate the usability of physiological modalities to improve affective computing methods with the long-term goal of enhancing HMI. For this purpose, this work will focus on the assessment of a user's emotional states from physiological signals of different nature. More specifically we are interested in the use of brain signals in conjunction with more classical physiological signals as will be detailed in Chapter 2.

1.2.2 Emotion assessment as a component of HCI

Figure 1.2 presents a framework describing how emotion assessment could be integrated as a component of Human-Computer Interaction (HCI). As proposed by Norman [8] the interaction with a machine, from the point of view of the user, can be decomposed in execution / evaluation cycles. After identifying his / her goals, the user starts an execution stage. It consists in formulating his / her intentions, specifying the necessary sequence of actions and executing those actions. Next, the computer executes the given commands and output results through available modalities. The second stage is the evaluation which is realized by: perceiving computer outputs, interpreting them and evaluating the outcome (*i.e.* are the goals satisfied?).

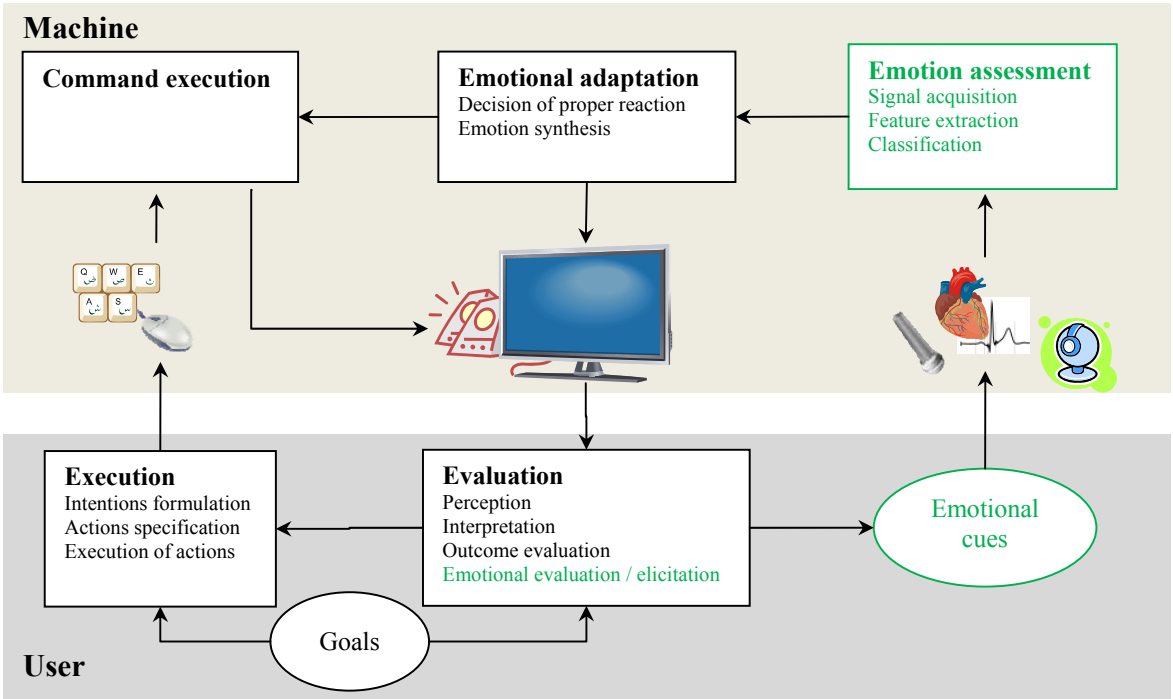


Figure 1.2. Emotion assessment in human computer interfaces, adapted from the execution / evaluation model [8].

According to the cognitive theory of emotions, emotions are issued from a cognitive process called appraisal that evaluates a stimulus according to several criteria such as goal relevance and consequences of the event [9-11]. For this reason, an emotional evaluation step, corresponding to the appraisal process, was added in Figure 1.2 at the evaluation stage. Elicitation of emotions is known to be related to changes in several components of the organism such as physiological, motor and behavioral components [10]. It is thus possible to consider those changes as emotional cues that can be used to automatically detect the elicited emotion after being recorded by the adequate sensors. The detected emotion can then be used to adapt the interaction by modifying command execution. The information presented on the output modalities can also be influenced directly by the emotional adaptation, for instance by synthesizing an emotional response on screens and speakers.

This thesis will focus on the green parts of Figure 1.2 to evaluate the performance of physiological signals for emotion assessment. For this purpose physiological signals should be acquired during emotional stimulation. This is often done by designing protocols to elicit emotions. Three protocols were designed during this thesis with one of them corresponding to a HCI context. Once the signals are acquired, some of their features that are known to be related to emotional reactions are extracted. The computed features are then used to train a classifier that finds a computational model mapping physiological features values to a given emotional state. This model (classifier) can then be used to recover the emotional state corresponding to a new instance of the features.

Several applications can be derived from the presented framework, some of them going beyond human-computer interfaces to reach human-machine interfaces in general and even human-human interfaces.

1.2.3 Applications of emotion assessment

When thinking about emotion recognition one application that generally comes to mind is the lie detector. This is certainly due to the fact that it was the first tool able to estimate the true internal state of someone without him / her wanting it, which led to the debates that we all know. However, this section will show that the lie detector is just the top of the iceberg and that many more applications can be targeted by research on emotion assessment. We believe that most of them are more ethical and less subject to debate, essentially because their goal is not to reveal information that could be seen as private, but rather to improve the way machines react to human feelings, prevent accidents and rehabilitate impaired people. The following list of applications is of course not exhaustive but is rather made to give some insights into the possibilities and advantages of emotion assessment technologies as well as to provide a better understanding of the context and stakes of this thesis.

a. Human machine interfaces

Emotion assessment can be used to automatically evaluate the user experience of software. User experience design aims at integrating the perception of the system by the user in the development of applications and devices. This mainly consists of asking users to fill in surveys about their emotional state and monitoring their facial expressions and physiological signals while they manipulate different interfaces [12]. In this case, offline automatic assessment of emotions helps to avoid the manual and visual processing of large amount of gathered data. But certainly one of the main interests for HMI is to adjust online the behavior of the machine to the emotional state of the user (i.e. directly after or during the feeling of the emotion).

As stated in Section 1.1, the inclusion of emotions in the human-machine loop can help to improve and disambiguate communication. For instance, when the user repeats a command for the second time and agitation or stress is detected as the current emotion, the system could infer that it should try to change its previous response because the result was not satisfying. There are at least two particular applications that could be cited to illustrate adaptation of HMI to human emotions: learning platforms and games.

In a learning situation many affective states can be elicited. Discouragement, anxiety, excitation, curiosity, are just some examples of the whole space of possible emotions. Some of those emotions are favorable to knowledge acquisition (i.e. curiosity) while some others are not (i.e. discouragement), some could also be useful as long as their intensity does not exceed a certain threshold (i.e. anxiety and excitation). With the development of e-learning, serious games and other learning platforms [13-15], it is now more and more common to learn through the use of computer interfaces. In a standard class-room the teacher is generally able to react according to the feeling of participants but this is not the case with standard computer interfaces. In this situation, automatically adapting the learning strategy of the system to the emotional state of the user could foster the acquisition of knowledge. Adaptation could be done trough the regulation of standards criteria like the speed of the information flow, the difficulty of exercises and questions but it could also be done by providing advices and tips to the learner. Another possibility is to have a virtual agent that could show artificial empathy by providing encouragements and displaying appropriate emotional cues [16]. In all those cases, the idea is to try to motivate the learner toward a state of mind that will increase his / her performance and learning outcomes.

Games are also interesting from a HMI point of view because they are an ideal ground for the design of new ways to communicate with the machine. One of the main goals of games is to provide emotional experiences such as fun and excitement which generally occurs when the player is strongly involved in the course of the game action. However, a loss of involvement can occur if the game does not correspond to the player's expectations and competences; he might

then feel emotions like boredom, distress and frustration [17, 18]. In this situation, the content of the game could be adapted to better fit the player and game designer's expectations. For instance, in a role playing game, if the player seems to have more fun by exchanging information with non-players characters and solving puzzles than fighting monsters, the game could automatically increase this type of events. Some games are also purposely inducing particular emotional states such as fear in the case of a horror game. Emotion detection could then be used to control the effectiveness of emotion elicitation and adapt the elicitation methods for better success. Emotional states could also be used to modulate movements of the player character such as having a worse precision at shooting in case of stress and limited movements in the case of fear to simulate paralysis. Finally, another possibility is to adapt the difficulty of the game according to the felt emotion. This can avoid falling in the extreme cases of too easy and too hard games that would respectively elicit boredom and frustration. This later approach will be discussed in Chapter 7.

b. Behavior prediction and monitoring of critical states

Emotions are known to modulate tendencies to action [10, 19, 20]. For instance when someone is feeling a positive emotion he / she will rather tend to approach the stimuli, while fear will generally result in withdrawal from the situation. Based on this fact, emotion detection is an indicator for behavior prediction. It is thus possible to monitor critical emotional states that could lead to potential harmful or dangerous behaviors.

Examples of such applications could be to monitor stress while driving to avoid accidents by encouraging the driver to stop and calm down. On a similar basis, some operators have to supervise critical operations and a loss of task engagement could induce severe damage. Monitoring of this cognitive state could then be of interest to re-engage the supervisor in his task. Emotions can also be monitored while performing dangerous operations, for instance to forbid the use of potentially harmful commands in the case of too high stress. Finally, monitoring of emotional states could also provide information to determine when an elderly or disabled person needs help, without having that person perform usual actions like giving a phone call to ask for help.

c. Information indexing and retrieval

The quantity of multimedia data (images, videos, sounds, music …) that is available on the internet is increasing. Moreover, availability of easy to use multimedia recorders embedded in low cost mobile devices leads to an incredible amount of self-collected data. This will be further encouraged by the coming of ubiquitous computing: the embedding of information processing tools in objects [21]. Recording would then be possible at any moment and any place.

For this reason, there is a real need of new technology to index and intelligently retrieve information. According to Hanjalic [22] retrieving information could be done by either highlighting particular items of interest or by summarizing the content of a complete recording. This could be done by finding information that contains strong emotional content [23, 24]. In this context emotion detection could be used to construct individual profiles that would store the relations between multimedia features and emotional states. In this way the search of emotional content would be adapted to each person according to his / her profile and his / her subjective interpretation of multimedia features.

Events that people would like to record generally coincide with strong emotional states like pleasure and joy. Thus, emotion detection could be used, in a ubiquitous environment, to detect those events and automatically trigger the recording of relevant scene of the every-day life. At the same time, this data could be indexed emotionally and retrieved later by using emotional tags. Automatic emotional tagging could also be performed at a later stage by presenting previously recorded data to participants while monitoring their emotional states. This last point is an active topic of research being conducted within the European Network of Excellence PetaMedia[1] to which we participate.

d. Health and rehabilitation applications

Empirical data analysis has shown that emotions are factors that can influence the speed and efficiency of treatments that are provided during a cure [1]. Thus, emotion assessment can be used as any other monitoring device to control patient's signs of rapid and effective recovery. However its usability and effectiveness would certainly be more important to help for the healing of disorders implying emotional and social impairments such as depression and autism. An example for autism is given by [25] where an emotional robot is designed to play with autistic children. The idea behind the use of a robot is that autistic children would prefer to interact with a machine that has simplified rules of communication and is thus less intimidating. In this application, emotion assessment is used to adapt the behavior of the robot (speed and direction of a basketball basket, background music and game objective) according to the preferences of the child. The objective is to have a robot that is able to change its behavior along time like a real therapist would do.

Some accidents and diseases, like amyotrophic lateral sclerosis, have as a consequence the complete paralysis of the body. Disabled persons that suffer from this type of trauma are not able to communicate through standard channels since this generally requires the activation of several muscles. In order to give them the possibility to communicate, one solution is to use interfaces

[1] http://www.petamedia.eu/ (Retrieved on 30 April 2009)

that do not rely on standard peripheral nerves and muscles but directly on the brain neuronal activations. Such interfaces are named Brain-Computer Interfaces (BCI) [26-28]. Current BCI aim to detect brain activity that corresponds to complex tasks (mental calculus, imagination of finger taping, etc.) that are traduced in commands like moving a mouse cursor and choosing a letter from the alphabet [27, 29, 30]. Generally the user needs training before using such systems. In case the objective of the user is to express an emotion, classical BCI tasks (e.g., imagination of finger tapping) seem to be really far from this objective and it is more appropriate to use mental tasks such as the remembering of a similar emotional event (see Chapter 6). This emotion elicitation task can then be regarded as a mental task that the user tries to perform in order to communicate his / her feelings. The objective is then to detect the resulting emotional states from brain signals and output it in order to give back to disabled people the ability to express their feelings [31].

1.2.4 Related questions / problems

Emotions have been studied from several perspectives and for many centuries starting with discussions of philosophers such as Aristotle [1, 32]. Darwin was also a pioneer who tried to explain the origin of emotions and their importance for the survival of species [33, 34]. But it is only recently (1950's - 1960's), with the emergence of cognitive sciences in psychology, that emotions gathered a real interest. Thus the study of emotions in psychology is rather a new field of research and there are still a lot of issues that remain unsolved. Those issues have a strong importance for HMI researchers since the definition, elicitation and characterization of emotions is the basis for proper emotion assessment, synthesis and adaptation. This section will emphasize the issues that are most relevant to emotion assessment.

When designing a system for emotion assessment one of the critical steps is to specify the emotions that will be detected. This question is directly related to the issue of the definition of emotions: what is an emotion, how many emotions exist, are there emotions more important than others? The representation and modeling of emotions is a subject of debate in psychology. As a consequence several models that account for different components of emotions were developed. The representation and models of emotions that were found to be of interest in the specific case of emotion assessment will be presented in Chapter 2. The easiest way to represent an emotion is certainly by using words like fear, anger and joy. However, there is an incredible amount of vocabulary to refer to our affective states. In the dictionary of affect in language Whissel [35] registered up to 3000 English words with affective connotations. Those words do not always have the same meaning from one individual to another and it is sometimes difficult to determine if a word refers to an emotion or not (think to "disobedient" for instance that is part of Whissel dictionary). Words are also culture and language dependent and the difference between two words is sometimes difficult to catch such as for joy and happiness. From an HMI point of view,

the question of the relative importance of some emotion labels to others is strongly dependent on the application. For instance, detecting disgust is not really interesting to adapt the learning strategy of a virtual teacher but it could be very useful to index horror movies. It is thus really difficult to choose the appropriate representation of emotions for the general purpose of emotion assessment while there is a strong importance of designing and choosing emotion models that are consistent with the targeted application.

Another point of interest is the question of subjectivity that can occur at several levels. If two persons are stimulated by the same event, they can feel very different emotions depending on several criteria such as their past experience and their current goals. For instance if a person is hungry, presenting him or her with chocolate can elicit pleasure, while it would elicit disgust in the case the person would be sick of chocolate. It is thus really difficult to infer the emotional state of someone just by having information about the event that elicited the emotion. Since drawing a complete profile of someone goals and past experiences is quite unfeasible, using emotional outputs is a promising solution to automatically infer emotions. Associating a stimulus to an emotional label is thus difficult, raising the question of the construction of emotional databases that need to associate recorded data with a specific emotion. A solution is to ask the user to report about his / her feelings exactly as if an expert was asked to give labels to objects of interest. This supposes that a person is an expert in evaluating his / her feelings. But is it really the case? Not always, according to psychologists, since they separate the objectively felt emotion (change of the organism state) from the resulting subjective feeling (the self-perceived emotion). All those remarks have strong implications on the design of protocols to gather emotional data and raise several questions: how to elicit an emotion with maximum certainty and how to annotate the collected data? There are no clear answers to those questions and assumptions will have to be made for the recording of emotional collections.

The use of physiological signals also raises a lot of questions. What type of physiological signals should be used? Is the physiological response to emotional stimuli always the same? There are many physiological signals that have been shown to correlate with emotions. Chapter 2 discusses most of these signals. However, it has been shown that there are differences in their activation for a given emotion elicited in different contexts. Thus an emotion detection system should be designed for a particular application or take into account contextual information. Simultaneously measuring various physiological signals requires the use of several sensors that are sometimes quite obtrusive since they can monopolize the use of one hand and are not comfortable. The price of those sensors should also to be taken into account. For these reasons the issue of determining the most useful sensors is of importance. Finally, there is also variability in physiological signals, from person to person but also from day to day, that yield to difficulty in designing a system that

will detect emotions accurately for everyone and everyday. Methods to cope with this variability have to be developed.

1.3 Contributions

The main objective of this thesis is to evaluate the usability of physiological signals as emotional cues for assessing emotions in the context of affective computing. More precisely, performance is analyzed for signals from the central nervous system and the peripheral nervous system separately as well as for the fusion of these two types of physiological measures.

The contributions of this thesis are:

- Identification of key points in computerized emotion assessment: issues such as the definition of emotional classes of interest, the choice of a ground truth for emotion classification and the duration of emotional epochs are investigated.
- Setting up acquisition protocols for different elicitation contexts: three different emotional stimuli were employed, namely images, recall of past emotional events and playing a game at several difficulty levels.
- Evaluation of classification algorithms for emotion assessment: comparison of different classification (Naive-Bayes, LDA, SVM, RVM), feature selection (ANOVA, FCBF, SFFS, Fisher) and fusion (feature and classifier level) techniques was performed.
- Demonstration of usefulness of EEGs (ElectroEncephaloGrams) for emotion assessment: the accuracy of emotion assessment from EEG features was compared to the accuracy obtained with peripheral features on different emotional classes and contexts. It was shown that EEG features can perform better than peripheral features especially for short-term assessment.
- Demonstration of the usefulness of fusion between peripheral and EEG modalities: the fusion was done at different levels (features and decision levels) with results showing the interest of fusion.

Those contributions were included in a number of publications in international conferences [36-38] and a journal [39]. These publications are also listed in Appendix E together with other contributions and collaborations related to the study of physiological signals for emotion assessment.

1.4 Thesis overview

This thesis is divided in 8 chapters.

Chapter 2 presents a state of the art regrouping the important notions concerning emotion assessment from physiological signals. It addresses several models and definitions of emotions proposed in the fields of psycho-physiology and sociology. The different physiological signals related to emotional processes are reviewed together with the existing methods and sensors usable for the acquisition of those signals. Finally, the results of studies using physiological signals to assess emotions are reported and discussed according to several criteria.

Chapter 3 concerns the acquisition of physiological signals. It first presents the material used for the acquisition of physiological signals. It also provides some recommendations regarding how to position the sensors and control the quality of the signals. The features extracted from the signals to characterize the physiological activity are then presented.

Chapter 4 describes the methods employed to assess emotions from the extracted features. It first addresses the possible methods available to construct a ground-truth, i.e. the true emotional state associated to a physiological activity. In the second part the algorithms used for the classification of those emotional states are presented together with the measure chosen to compare their performances. Finally, the feature selection methods used to reduce the amount of features are detailed.

Chapter 5, 6 and 7 present the results of emotion assessment from physiological signals for different emotion elicitation strategies and classification methods. In Chapter 5, the emotions were elicited by using pictures and different ground-truths were compared. In Chapter 6, a self-induction method was employed to elicit emotions belonging to three categories (pleasant, unpleasant and calm). The performances of different physiological signals were compared and the advantages of fusion of these signals were investigated. In Chapter 7 a gaming paradigm was proposed to get closer to real HCI applications. The performance of emotion assessment from different signals was investigated for several time scales. The changes in physiological activity following game-over events were also analyzed.

Chapter 8 presents the conclusions of this work and provides some suggestions for future work and improvements.

Chapter 2 State of the art

2.1 Emotion representations and models

Before designing a system able to represent and recognize emotions, it is first necessary to define what an emotion is. At the moment, there is no common framework that can be used to answer this question. This is mainly due to the fact that emotions are complex phenomena that influence many aspects of everyday life: they help us to avoid danger, take decisions and they also have some important social values. Over the past century, four main theories of emotions have developed [33].

From a **social constructivist** point of view, emotions are only products of social interactions and cultural rules [33]. This statement is often considered too restrictive but nowadays most of researchers in affective sciences admit that social and cultural rules at least regulate emotions: one will not have the same emotional behavior in front of one's senior than in front of one's friend. On the one hand, this view emphasizes the importance that emotions have in society and uncovers applications for emotion assessment in social environment, for instance in computerized social networks. On the other hand, it also suggests that emotions should be assessed with models that are cultural and environmental specific, making this assessment more complex.

Darwinians consider emotions as phenomena that are selected by nature according to their survival value, i.e. fear exist because it helps us avoid danger. This is one of the oldest views concerning emotions and has initially been proposed by Darwin in 1872 [33, 34]. The main implication is that emotions should have identical constructs across individuals and thus common emotional expressions and behaviors should be found across cultures. This point of view is clearly opposite to social constructivism and supports the idea that a single model for assessing and representing emotions universally is conceivable.

For **Jamesians**, emotions uniquely corresponds to the feeling of bodily changes, such as modifications of heart rate and blood pressure, which follow the emotional stimuli ("I am afraid because I shiver") [33, 40]. Thus if perceptual mechanisms are impaired there is no possibility to feel emotions anymore. Although controversial, this later approach emphasizes the important role of physiological responses in the study of emotions. More specifically this implies that each emotion corresponds to a unique pattern of physiological activity, which is strongly encouraging to go toward emotion assessment from physiological signals.

In the **cognitive** theory, emotions are issued from a cognitive process called appraisal. This process is supposed to be "direct and non-reflective" and evaluates a stimulus according to several criterions such as relevance and consequences of an event [9, 10, 41]. This approach supports the idea that the brain is the main organ implied in emotions through organization of

emotional processes and triggering of emotional responses. As a consequence, brain signals should not be neglected for the analysis and detection of emotions. An advantage of this theory, from a computer-science point of view, is that it allows for computationally tractable models of emotions (see Section 2.1.3).

From all those theories, different models and representations of emotions have emerged. There were also some attempt to combine and harmonize the different views but they are generally still anchored in one of the above theories. In the following sections two types of representations will be emphasized: the so-called **basic emotions** and the **continuous representations**. They are clearly the most used in the emotion assessment community. Models of emotions are discussed in section 2.1.3 but they are generally more complex and less usable for the purpose of emotion assessment from physiological signals alone.

2.1.1 Emotion categories and basic emotions

As stated in the introduction, dealing with the emotional vocabularies is difficult because of the high number of emotional words, the fact that they can be interpreted differently from one individual to another and also because they are culture dependent. In order to organize and reduce this large vocabulary one can rely on the basic emotions. Unfortunately, basic emotions have been introduced by several researchers [34, 41, 42] so that their definition, their number and their identity can be different depending on the studies and the theory in which researchers believe. According to Ortony [34] there are two main approaches to the definition of basic emotions: the **biological view** that is strongly anchored in the Darwinian and the Jamesian theories, and the **psychological view**.

For the biological point of view, basic emotions are those that are solely evolved by nature and thus have important survival functions. There are several criteria that an emotion should meet in order to be basic, the most commons being:

- distinctive physiology: the emotion must be associated to a unique pattern of physiological activity,
- universality: the emotion must be found in all cultures,
- unlearning: the emotion must not have been learned from previous experiences but is naturally present and hardwired in the brain.

Table 2.1 presents several lists of basic emotions depending on the criterion that was used to construct them. As can be seen from this table the length of the proposed lists does not exceed 10 different emotions, while general lists of emotional terms and emotional dictionaries can contain

up to thousands of words [35]. What about the remaining emotions? The psychological point of view can help to give an answer to that question.

Reference	Basic emotions	Criteria
Ekman [42]	Anger, disgust, fear, joy, sadness, surprise	Universal facial expressions
Gray [43]	Rage and terror, anxiety, joy	Hardwired
Panksepp [44]	Expectancy, fear, rage, panic	Hardwired
McDougall [45]	Anger, disgust, elation, fear, subjection, tender-emotion, wonder	Relation to instincts
Mowrer [46]	Pain, pleasure	Unlearned emotional states
Plutchik [47]	Acceptance, joy, anticipation, anger, disgust, sadness, surprise, fear.	Relation to adaptive biological process
James [40]	Fear, grief, love, rage	Bodily involvement

Table 2.1. Lists of basic emotions from a biological point of view (from [34]).

For several psychologists, an emotion is considered as basic if it is an "elementary" one that can be used to construct, in combination with others basic emotions, a large number (if not unlimited) of non basic emotions [34]. The word "elementary" signifies that a basic emotion is one that cannot be decomposed into a combination of other emotions. One of the main advantages of this definition of basic emotions is that it enables to virtually construct any emotion as a combination of basic emotions like it is possible to mix primary colors to obtain secondary ones. However, the way the combination is done differs among researchers and is not really clear. An example of this view is the wheel of emotions proposed by Plutchik [48, 49] that is shown in Figure 2.1. This wheel is composed of several quadrants containing the proposed basic emotions organized according to their closeness: emotions that are facing each other on the wheel are considered to be opposite while emotions in adjacent quadrants have common properties (especially if they have similar colors). Two adjacent emotions can then be combined to form a new non-basic emotion. For instance anticipation and joy will constitute optimism. In his wheel, Plutchik also added the concept of intensity as can be seen from Figure 2.1.

Both the psychological and the biological views of basic emotions are challenged by social constructivism. Since in this theory emotions are considered to be a product of nurture rather than nature, there is no reason for the existence of basic emotions in the biological sense. Concerning the psychological view, social constructivists argue that the choice of the basic emotions is biased by cultural influence and that the so called "elementary" emotions are rather emotions that are prominent in the culture. For this reason, Shaver et al. [50] performed clustering of several emotional words using similarity between words evaluated by many participants. The result is a hierarchical tree structure where words are organized in three groups' levels. The names of the

groups were chosen as the most common emotions or as a close basic emotion. The main advantage of this representation is that it allows for the taxonomy of several emotional words.

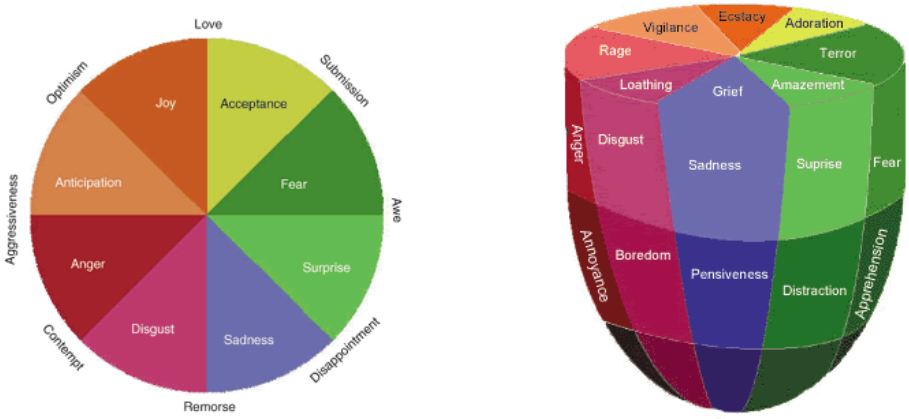


Figure 2.1. Plutchik's wheel of emotions. (Left) The basic emotion represented as quadrants and possible combinations of basic emotions[2]. (Right) The same wheel with the added concept of intensity[2].

While representing emotions with words has the advantage of being instinctive, the questions of cultural differences, misunderstanding and choice of words are still not resolved. Moreover, labels are discrete and hence cannot fully represent some aspects of emotions like the continuous scale of emotional strength. This is why several researchers focused on the search of a continuous space that would represent emotions with less ambiguity.

2.1.2 Continuous representations

The idea of searching for a continuous space able to represent emotions comes from the cognitive theory that assumes that any person possesses an internal representation of emotions [51]. The goal is to find what the dimensions of this representation are and how emotions are mapped into it. Several researches (see [51] for a review) have generally addressed those questions by analyzing participant-reported measures of similarity between either verbal or facial emotional expressions. In those analysis different methods were applied to find a space that minimize (resp. maximize) the distance between similar (resp. different) expressions. Most of the studies obtained different spaces but, on a closer look, some of the dimensions were redundant and similar across studies.

As a result there is nowadays an agreement on the first two bipolar and most important dimensions: valence and arousal (Figure 2.2). Valence represents the pleasantness, ranging from unpleasant to pleasant (it is also called evaluation); while arousal represents the awakeness,

[2] http://library.thinkquest.org/25500/index2.htm (Retrieved on 4 May 2009)

alertness and activation of the emotion (this dimension is sometimes called sleep-tension and activation) [20, 52]. There is currently no consensus on a potentially third dimension. Most of the studies emphasize that it accounts for a low part of the variance in participants judgments, generally leading to its neglect. Otherwise, this third dimension could be related to concepts such as potency, control, and dominance in order to distinguish between emotions that are related to approach and withdrawal reactions.

The two-dimensional valence-arousal space has several advantages. First, it is possible to represent emotions without using labels but just a coordinate system that includes emotional meaning. As a consequence, any emotion could be represented as a point in this space, even if there is no particular label or expression to define it.

Secondly, since this space was created from the analysis of emotional expressions (verbal and non verbal), it is possible to associate areas of this space to emotional labels if necessary (Figure 2.2). In [51], the direct mapping of verbal expressions on the space gave the result shown in Figure 2.2.a. As can be seen, labels tend to form a circle within the space. However there are some evidences that this mapping is not exactly the same from a person to another. In consequence, there are no exact boundaries in the valence-arousal space that define emotional expressions. A solution could be to represent this variability by defining expressions as areas that can overlap (Figure 2.2.b). Using probabilistic models that define the probability of having a given expression knowing the position in the valence-arousal space could help determine such boundaries.

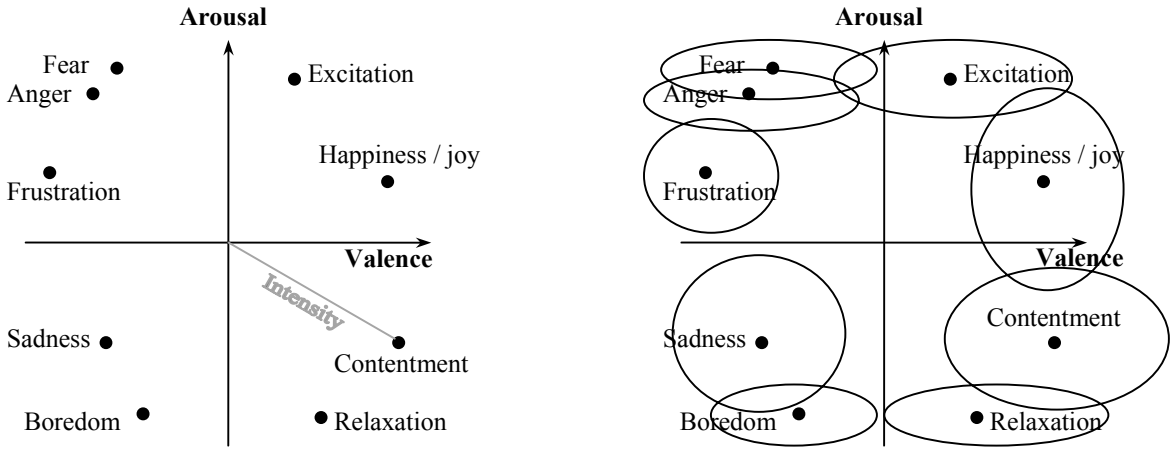


Figure 2.2. Valence arousal space with associated labels as (a) points (adapted from [Russell], (adjectives have been changed to nouns and only some of the words are displayed for clarity) and (b) areas.

Thirdly, this space can also represent the intensity of the emotion [20, 51]: a point in the middle of the space would represent a low intensive and neutral feeling while a point on the periphery

would indicate a full-blown emotion. Finally, it has been shown [53] that the mapping of emotional words on this space does not vary significantly between four Europeans languages (English, Estonian, Greek and Polish). This encourages the use of this space as a way to represent emotions interculturally.

As discussed, emotional labels can be projected from the discrete space of emotional labels to the continuous valence-arousal space. Unfortunately there is nothing that guarantees this projection to be injective so that two different labels could have the same coordinate in the valence-arousal space. In the representation above (Figure 2.2), this is merely the case for fear and anger, two highly negative and excited emotions. In that case the clear distinction between the two emotions could be achieved by adding the third dimension of control: when one is feeling fear one does not have control over the situation while control is present for anger.

The valence-arousal space has been shown to be effective not only to represent emotional expressions but also for self-assessment of emotions and moods [51, 52, 54, 55]. However, while Russel [52] argues for independent dimensions of valence and arousal, there are some evidences that this is not true for self-reported feelings. Lang [54] observed that when people assess their own emotion while watching pictures of the IAPS (International Affective Picture System), their judgment tends to follow a U-shape centered in the space (Figure 2.3). This shape has also been observed using film clips as stimulations in [24]. This distribution of self-assessments is not surprising since it is difficult to elicit emotions that have low arousal and high valence simultaneously as well as emotion of high arousal and no valence. This also demonstrates that areas of this space are more important than others and that a system performing emotion assessment should focus on them.

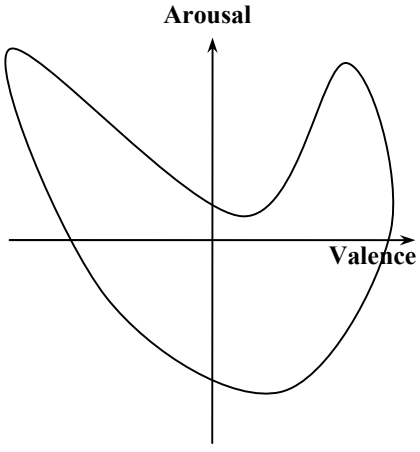


Figure 2.3. Self-assessments distribution obtained when eliciting emotions with images: most of the self-assessed images lie inside the U-shape.

2.1.3 Models of emotions

Sections 2.1.1 and 2.1.2 describe different representations of emotions. While they are useful for taxonomy, categorization and differentiation of emotions they do not provide much information about the process that gives rise to emotions. Several models, essentially from the cognitive view of emotions, have been proposed to answer this question [9, 10, 32, 56-58]. Most of them are based on the central concept of appraisal which has been defined by Arnold as the cognitive process that constantly evaluates the environment and elicit emotions [41]. It is generally considered as an unconscious process that is direct and non-reflective. To our view, the main differences between models of appraisal are the different evaluation criteria they propose.

Two of the most famous models are presented in the following sections: the OCC (Ortony Clore and Collins) typology [57], which has been used to define computational models of emotions, and Scherer's SECs (Stimulus Evaluation Checks) [9, 10], which can be viewed as an attempt to unify the different emotion theories.

a. OCC typology

According to Ortony [34], "the defining feature [of an emotion] that we consider most reasonable and least contentious is that the appraisal underlying the emotion be valenced, either positively or negatively". It is thus not surprising to find the valence concept at the heart of his model. In the OCC (Ortony, Clore and Collins) typology [57], an emotion is viewed as a valenced reaction to a stimulus. More precisely, the elicitation of an emotion relies on the evaluation of three main criteria: the **type of the stimulus** (agent, object, and event), the **concerned entity** (self or another) and finally the **valence** of the emotion (positive or negative). Since each of these criteria can only be discretely evaluated, it is possible to represent the model as a tree with the resulting emotions as the leaves (Figure 2.4).

Two examples are given bellow that respectively corresponds to the green and blue dotted line of Figure 2.4. As a first example, imagine that a student just sat for an exam and is rather pleased by his / her performance. In this case the type of stimuli is an action (sitting for an exam), the agent of interest is the self, and since the valence is positive the resulting emotion is pride (green dotted line in Figure 2.4). As a second example, our student receives a letter concerning the result of the exam and reads it. Here the stimulus is an event (reading of the letter) and the concerned entity is again the self. Since prospects are relevant in this case and the student has opened the letter, the elicited emotions will be either satisfaction or disappointment depending of his / her results (blue dotted line in Figure 2.4). Notice that as long as the letter wasn't read the elicited emotion was hope (he or she was happy about his / her performance).

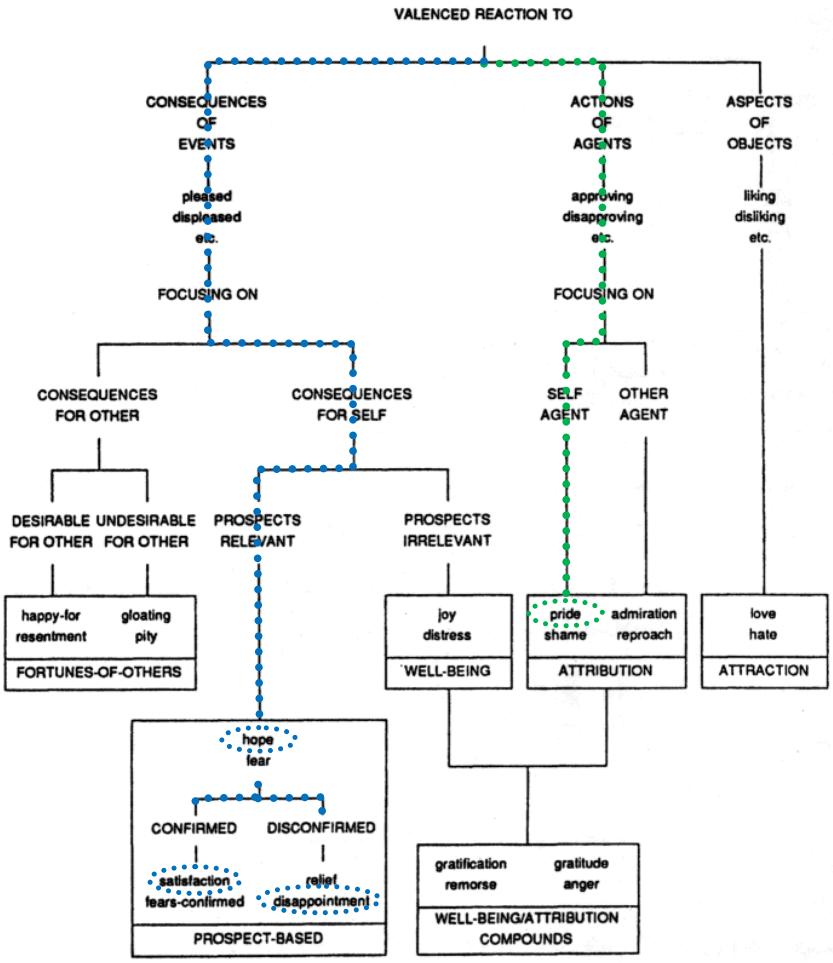


Figure 2.4. The OCC typology (from [57]). Green and blue dotted lines correspond to the examples above.

This model has the advantage of being computationally tractable. Consequently, it has been used for the purpose of emotion synthesis. Elliot extended it in [59] by adding few emotions, and implemented it in a computational affective reasoner. In [60], the authors converted this model in a BDI (Belief, Desire and Intensions) architecture and also demonstrated the effectiveness of this approach.

This model can also be used for the purpose of user-emotion assessment. In order for this model to be applicable in this case, it is mandatory to have a complete knowledge or control of the environment in which the user is evolving to be able to follow a path in the tree. This is an interesting fact since it strongly emphasizes the importance of context in emotion assessment. However, even with a strong knowledge of context, valence is still a hard criterion to evaluate. Let's take the example of a learning application since, as explained in the introduction, there are several advantages to include emotion assessment in this type of environment. Imagine that the user is informed that he failed in a task then the valence could be guessed as being negative. Now

imagine that he just finished the task but he has not been informed about his / her performance then the valence will depend on the user interpretation of his / her success. Since this information is not directly available to the system, it is not possible to infer any emotional state. In this case, detecting the valence from the measures of user's emotional cues can solve the issue.

b. SECs theory

In his theory, Scherer [9, 10] proposed that appraisal can be decomposed into several checks that are cognitively evaluated. They are called the Stimulus Evaluation Checks (SECs, Table 2.2). Each of these processes is supposed to occur in a specific time span from the elicitation event with the possibility of parallel processing (time span that overlap). Those SECs can be grouped in the four appraisal objectives listed below.

- **Relevance detection**: the objective of this group of SECs is to determine if the stimulus requires attention and further processing by analyzing its relevance for the person; the goal is to answer the question "Are there any possible implications to me or my direct environment?".

- **Implication assessment**: the next step is then to determine what the implications of the stimuli are in terms of consequences for the self. This objective regroups the central processes of appraisal and its function is mainly the protection and the progress of the organism: "is the stimulus dangerous or appealing?", "does it correspond to my goals and needs?". Notice that this appraisal objective is very close to what Darwin defines as the function of emotions as a whole.

- **Coping potential determination**: the concept of coping is well known in psychology and is defined as the cognitive mechanisms that are implemented to control and reduce the impact of stressful and emotional events. According to Lazarus [32], there are two types of coping: problem oriented (determination of an action to solve the problem that gave rise to the stressful situation) and emotion oriented (cognitive regulation of the stress for instance by reconsidering the situation). The current appraisal objective is to deal with both of those aspects.

- **Normative significance evaluation**: the objective of this SECs group is to evaluate whether the stimulus is compatible with one's own principles and standards as well as with social norms and values.

The order in which the SECs are presented in Table 2.2 is significant as it represents the sequence in which the checks are supposed to be completely evaluated. The appraisal decomposition in the present SECs as well as the temporal aspects have been validated in [61, 62].

Appraisal objectives	SECs	Description / Answer the question
Relevance detection	Novelty	Is stimulus novel and does it require attention? Is it familiar and/or predictable?
	Intrinsic pleasantness	Will the stimulus lead to pleasure or pain? This is a property of the stimulus and is not related to the current state of the organism (it does not depend on the current goal and objectives of the subject but could have been learned in the past).
	Goal relevance check	Does the stimulus have some consequences on my current goals?
Implication assessment	Causal attribution	What or who is the cause of the stimulus and why? Often divided in two categories: the responsible agent and the motive.
	Outcome probability	What are the consequences of the stimulus? What is the probability of each consequence?
	Discrepancy from expectation	To which extent the stimulus is different from what was expected.
	Goal / need conduciveness	Will the consequences of the stimulus help me to accomplish my goals or will they obstruct my goals?
	Urgency	Should the stimulus be handled quickly?
Coping potential determination	Control	To which extent the stimulus can be influenced and controlled.
	Power	If control is possible, check the available resources (physical, knowledge, etc.) for a potential action. It is related to the problem oriented coping of Lazarus.
	Adjustment	If it is not possible to influence the situation (because of a non controllable stimulus or lack of resources), check how well the organism can adjust with the situation. It is related to the emotion oriented coping of Lazarus.
Normative significance	Internal standards	Is the stimulus in accordance with one's own principles, ideals and moral code?
	External standards	Is the stimulus in accordance with social norm and values?

Table 2.2. List and description of the different Stimulus Evaluation Checks (SECs) grouped by appraisal objective and temporally ordered.

Contrarily to the OCC model, where the different criteria are evaluated discretely, Scherer proposed that most of the SECs are evaluated on continuous scales (for instance the outcome probability SEC is evaluated for each event on a continuous scale ranging from 0 to 1). Moreover, 14 emotions such as happiness, disgust, anxiety, fear, pride, guilt and boredom were profiled by giving for each of them the possible associated evaluation of the SECs [9] (for instance Happiness is associated to high intrinsic pleasantness, medium goal/need relevance, very high outcome probability, very low urgency, etc.). It was shown that these profiles can be used to correctly identify an emotion associated with situations described according to the SECs.

However, the emotional profiles are described with words like “low”, “high”, “medium” so that the computational implementation of this model could not be directly done with continuous evaluations of the SECs. A first step should be taken to find evaluation thresholds (for each SEC) that define the boundaries between emotions. Moreover, it still remains that this model is very complex and requires the evaluation of 13 SECs confirming the remark made for the OCC model about the difficulty of components evaluation.

In his appraisal theory, Scherer does not limit his analysis to the different components of the cognitive appraisal but also studied the relation between the systems (or components) of the organism taking part in the emotion elicitation and differentiation [10]:

- the **cognitive** system, that is responsible for the evaluation of the emotional event (appraisal of the stimuli);
- the **autonomic** system, that provides support for other components (for instance the quantity of blood in hands can increase in case of anger to prepare for action);
- the **motivational** system, related to action tendencies, urges and desires (i.e. will one withdraw from the stimuli or approach it);
- the **motor** system, to execute the (re)-action but also to show facial expressions as well as emotional gestures and behaviors;
- the **monitor** system, which gives rise to the subjective feeling one can experience after being confronted to an emotional situation.

This list clearly demonstrates the multimodal aspects of emotions, since emotional activation is reflected in the activity of all those systems. Notice that this activity could be monitored in order to assess emotions; actually it is even possible that monitoring the activity of ALL those systems is necessary to reliably and fully assess emotions.

This last statement goes along with the componential patterning theory. According to this theory, a given evaluation of the SECs leads to a precise activity pattern in the five organism components and this pattern can be used to predict emotion. Another important statement is that the different components are not independent but rather interrelated. Consequently, they exchange information for the evaluation of the different SECs. For instance the evaluation of a SEC (cognitive component) will give rise to activity in another component (for instance in the autonomic system); this activity could then be feedbacked for the evaluation of another SEC. Since the SECs are evaluated in sequence, this theory implies that the pattern of emotional activity changes during the appraisal process and that synchronization between the different systems is necessary.

The collaboration between different components has also been suggested in [63] while the dynamic and synchronization aspects are detailed in [64].

2.1.4 Finding an adequate representation

In the emotion assessment field, many studies are classifying emotions into categories such as anger, fear and disgust. In order to determine the categories of interest most of the studies focus on basic emotions since they are the most relevant because of their social or biological value but also because they can be used to determine more complex non-basic emotions. However, several lists of basic emotions are available (Table 2.1 is just a sample of those lists) and it is thus necessary to make choices. For instance, Ekman's six basic emotions, also known as the big six, have been widely used as the target classes to be recognized from the analysis of facial expressions [20, 65-68]. Those target classes were chosen because Ekman's definition of basic emotions was based on the universality of facial expressions. Having a system able to detect the big six from facial expressions would thus be universal. But what happens if one wants to assess emotions based on others modalities such as signals of the body? There is evidence that some of Ekman's basic emotions have specific patterns of physiological activity [5] but it is still not clear if this is true for all of them. If brain activity is monitored one could rely on James's basic emotions or on Gray and Panksepp sets (Table 2.1). But does it really make sense to change the set of targeted emotions according to the monitored activity? A solution could be to have a look at all lists of basic emotions and retrieve the most frequent ones to construct our own set of emotion to recognize. Unfortunately this is a difficult task because of the high number of such lists and the ambiguity of emotional words (ie. are rage and anger different emotions or do they refer to the same state?).

One interesting point coming from the psychologist view of basic emotions is the idea of mixture of emotions. This idea states that emotions do not always appear in isolation but that they can also co-occur on a relatively short time scale (order of the second). The co-occurred emotions could be similar but also opposite, and imply the activation of two parallel, or at least two quickly consecutive appraisal processes. For instance, I used to visit my grand-parents at the hospital and each time I entered the room, the elicited feeling was a mixture of pleasure and sadness. Pleasure because I was happy to see them again, sadness because I could see that they were not as healthy as I left them from the previous visit. There are two possible ways to account for such a phenomenon in emotion assessment: the first is to consider that the mixture is a new emotion (for instance by using Plutchik mixture model) and the second is simply to allow for multiple labeling of emotion episodes. In both cases it is necessary to first identify several emotions from one epoch's data. Several models have been proposed to perform multiple emotion labeling [69, 70]; of interest is the fact that some of those also allow for the characterization of intermediary states,

occurring at the transition of an emotion to another, which can be regarded as mixed emotional states [71].

In any case, the definition of a set of emotional labels should be strongly related to the targeted application. For instance, if the goal of an application is to monitor critical states while driving, there is actually no sense in detecting an emotional state such as disgust just because emotions are detected from facial expression. First, because there are very few chances that this emotion arises during driving. Secondly, because there is no clear action that the system (the "emotionally intelligent" car) should take to deal with the situation. On the other hand detecting boredom, which is not part of Ekman's basic emotions, could be of interest.

It is also important to check if the application can gather contextual information that could be useful for emotion assessment. For instance, the OCC model could be used together with contextual information to determine a path in the OCC tree while a binary assessment of valence (positive vs. negative) would then be enough to finally determine the elicited emotion.

More than detecting emotion categories this approach could be useful to obtain information concerning the intensity of the felt emotion which discrete labels generally fail to do. The labels proposed by Plutchik in his wheel of emotions are related to intensity but this representation has the disadvantage of increasing the number of emotions to detect (Figure 2.1). Another possibility is to associate to each emotional label a continuous level that will represent the intensity of each emotion. This is what Rani [70, 72] proposed after observing that emotional physiological reactions where different for low and high anger. Rani's representation has the advantage that mixtures of emotions can also be represented. However, in this representation it is important to account for the relations between the different emotions in order to avoid inconsistent state such as high anger and high boredom which is not the case in Rani's proposal.

As stated before, the continuous representation directly handles the notion of intensity. But it suffers for a major drawback: it is not intuitive compared to discrete emotional labels leading to difficult manual tagging of emotional data. While this is not a very important issue if this task is done by an expert, most of tagging is actually done by people who do not have any knowledge about this type of representations. To alleviate this problem, a description of how to use of a valence-arousal grid for self-assessment has been proposed in [52]. Alternatively, the SAM (Self Assessment Manikin) has been proposed in [73]. The SAM is composed of three nine-point scales where the different values of valence, arousal and dominance are associated with pictograms (Figure 2.5). Even if pictograms help the user to better understand the definition of the different axes they should be introduced to him / her prior to self-assessment. Another disadvantage of the valence-arousal space is that mixtures of negative and positive emotions cannot be represented. It is then necessary to decompose the emotion into two points in the

continuous space or choose the average response. This often confuses users when they try to self-assess their emotions on a long time scale where there are probably several emotions that can be elicited.

One of the main advantages of continuous spaces is that it can be seen as a general representation of any emotion. As a result it is possible to use them in any application, with only the area of interest varying from an application to another. In the case of the valence-arousal space it seems that only some areas are of real importance [23, 54]: negative-excited, positive-excited and calm neutral (Figure 2.3). Another advantage is that if a point in this space is determined (by automatic emotion assessment for instance), it is possible to associate an emotional label to it. It is also possible to assess only one axis of the valence-arousal space which could be enough for certain applications (for instance, monitoring of arousal for driving). Because of its generality but also for the other reasons mentioned above, we believe that it is preferable to assess valence-arousal space dimensions or areas than discrete labels if there is no particular targeted application for emotion assessment.

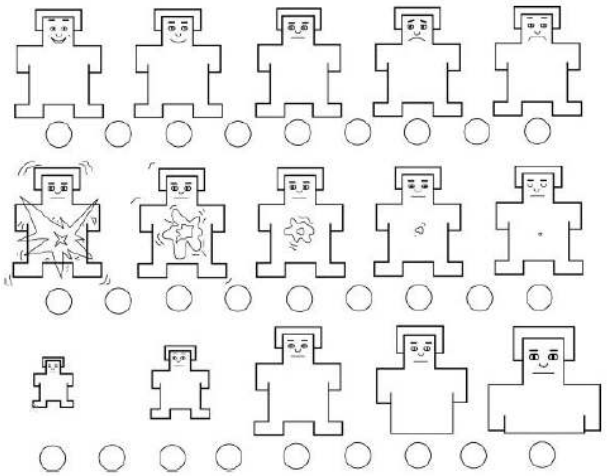


Figure 2.5. Picture of the SAM scales (from [73]).The first line evaluates valence from positive (left) to negative (right), the second arousal from excited to calm and the third dominance from submissive to powerful.

2.2 Physiological signals

As sustained by the SECs theory, emotions should be assessed from the monitoring of the five systems involved in emotion elicitation (see Section 2.1.3.b). In this thesis, the performances of two of those systems are analyzed for emotion assessment: the cognitive and the autonomic systems. Several signals can be gathered to relate activities of the five systems (for instance a camera can be used to partially measure the motor activity) but the present work focuses on signals recorded by non-invasive sensors placed directly on the body: the physiological signals.

Physiological signals can be defined as signals that quantify physical and chemical phenomena occurring in organs and tissues. This definition emphasizes the existence of a device able to quantify a physical and chemical phenomenon: the sensor. The complete apparatus allowing the visualization and recording of (many) physiological signals is called an acquisition system and is detailed in Figure 2.6.

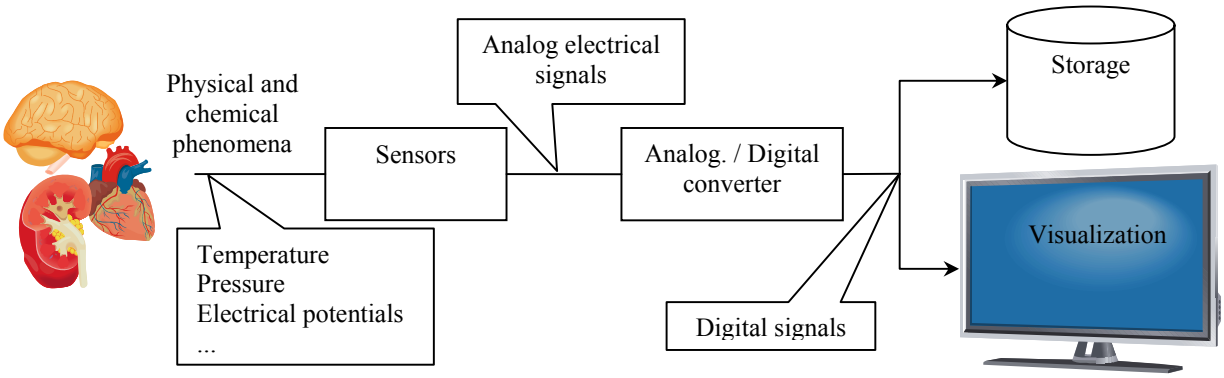


Figure 2.6. An acquisition system for visualization and storage of physiological data.

There is a wide variety of physiological signals depending on their source (the organ under consideration) and the phenomenon that is measured. For instance heart activity can be measured by recording the electrical potentials on particular positions of the body, the heart sound and the blood pressure in the coronary arteries. In this thesis, since all organs are connected to and controlled by the nervous system, we chose to divide them according to the following taxonomy: the central nervous system (CNS) and the peripheral nervous system (PNS). Thus, activities of organs that are controlled by the PNS will be referred to as peripheral activities while the activity of other organs (in this study only the brain) will be named the central activity. Notice that this taxonomy enables us to separate the two systems that are supposed to be at the center of the cognitive and Jamesian theories of emotions.

2.2.1 Central nervous system (CNS)

The CNS is composed of the brain, the cerebellum, the brain stem and the spinal cord. Since in this study only brain activity was analyzed, this description will focus on the brain structures and activities related to emotions as well as on the devices that could be used to measure those activities.

a. Brain structure

To better understand the functioning of the different devices used to monitor brain activity and the analysis of the resulting signals, it is first necessary to briefly discuss the anatomy of the brain. The human brain is made of around 100 billions of neurons that are interconnected through synapses, constituting neuronal networks. Neurons are specialized cells that vary in size and

shapes but are always constituted of three main parts: the dendrites, the soma and the axon (Figure 2.7.a). The soma is the cell body of the neuron and contains the same elements as typical cells (nucleus, ribosomes, etc.). The dendrites receive the output of one or many incidental neurons as inputs while the axon propagate the output of its neuron. The electrical signal transmitted from one neuron to another is composed of action potentials. The main task of a neuron is to integrate input action potentials to determine if an action potential should be generated on the axon (Figure 2.7.b). The activity of a neuron is proportional to its firing rate: the more a neuron is active the higher the firing rate.

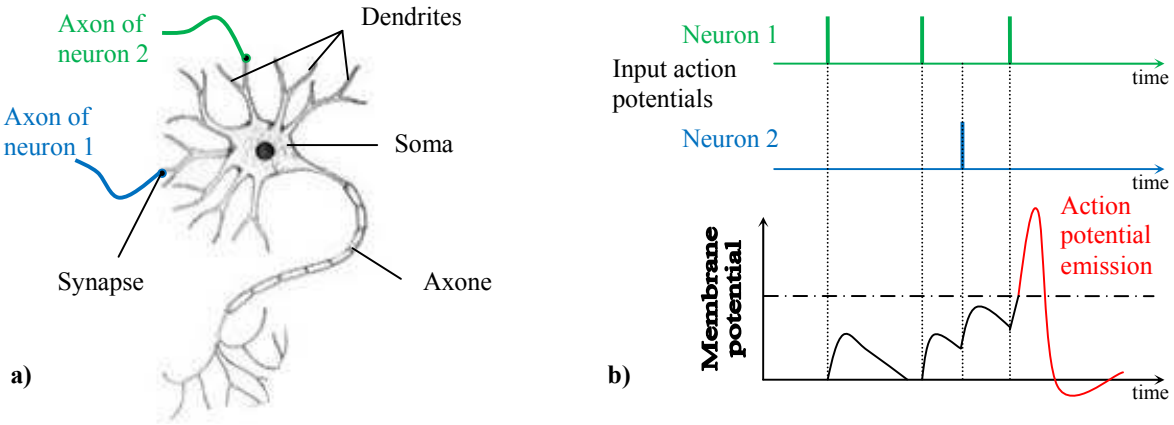


Figure 2.7. (a) Figure of a neuron[3] connected with two input neurons (named 1 and 2). (b) Representation of the integration of input action potentials; the neuron fires only if its membrane potential exceeds a given threshold.

The brain is divided into four lobes according to the names of the bones that surround them and the sulcus that separate the different lobes (Figure 2.8): the frontal, parietal, temporal and occipital lobes. Each of those lobes has been considered to be specialized in particular cognitive tasks. For instance, the frontal lobe is known to be implied in planning tasks while the occipital lobe is considered to be the vision center. Even if this view is superseded by new approaches that emphasize the cooperation between different areas of the brain during a single cognitive task, this nomenclature is still used to describe areas of the brain.

[3] http://www.wiredtowinthemovie.com/mindtrip_xml.html (retrieved on 27 April 2009)

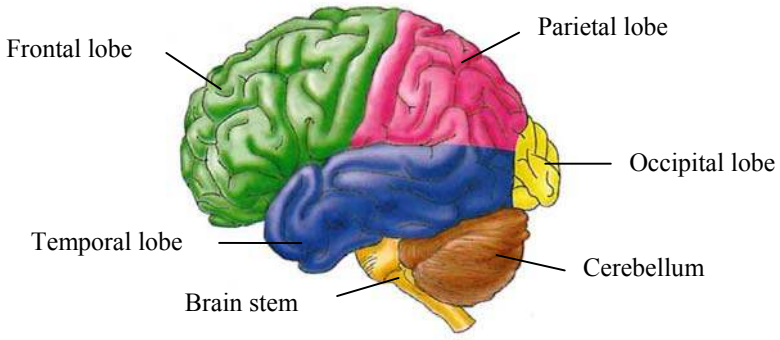


Figure 2.8. Image of the brain, the brain stem and the cerebellum with the different lobes highlighted (from [74]).

b. Measuring brain activity

Since neurons communicate trough action potentials it is possible to monitor brain activity by measuring electro-magnetic fields. Those methods are known as **direct methods** since they directly measure the activity produced by the neurons and can be divided into two types:

- electroencephalography (EEG) measures electrical potentials by placing electrodes either on the scalp (surface EEG) or, by incising the skull, on the surface of the brain (Electro-Cortico-Graphic electrodes - ECoG) or directly in neurons of interests (intra-cortical electrodes);
- magnetoencephalography (MEG) measures the magnetic activity of neurons by using specialized devices (Superconducting QUantum Interference Devices - SQUID) able to detect small changes in magnetic fields.

Indirect methods monitor other parameters that are related to neuronal activity such as artificial tracers injected in the body and resources consumed by neurons (oxygen for instance):

- Positron Emission Tomography (PET) measures the positron emission of a slightly radioactive tracer;
- Single Photon Emission Computed Tomography (SPECT), similarly to the PET, requires the injection of a radioactive tracer emitting gamma radiations that are measured by a gamma ray camera;
- functional Magnetic Resonance Imagery (fMRI) detects the changes of oxyhemoglobin and desoxyhemoglobin by applying successive magnetic fields on the head that force protons to release energy at a particular frequency that is detectable by the system;

- functional Near InfraRed Spectroscopy (fNIRS) also monitors the changes of oxy-desoxyhemoglobin by measuring the reflectance of a near infrared diode (0.7-1.5 µm wave-length) placed on the scalp using a near infrared sensor placed next to the diode.

All those methods can be evaluated according to several criteria: spatial resolution, time resolution, financial cost, needed machinery size / weight, invasiveness, chirurgical operation and/or technical preparation. Table 2.3 evaluates the different methods above according to those criteria. High spatial and temporal resolutions are important for a precise analysis of signals but current methods do not allow for both high spatial and temporal resolution. However, recent studies are trying to perform multimodal recording of brain activity in order to combine resolution advantages of direct and indirect methods.

Methods		Spatial resolution	Time resolution	Financial cost	Machinery size / weight	Invasive.	Preparation
Direct methods	Surface EEG	Low	High	Low	Low	None	Apply gel on the scalp
	ECoG	Low	High	High	Low	High	Surgical incision of the skull
	Intra-cortical	High	High	High	Low	High	Surgical incision of the skull and insertion of the electrode in the brain
	MEG	Low	High	High	High	None	
Indirect methods	PET	High	Low	High	High	Low	Injection of a radioactive tracer
	SPECT	Low	Low	High	High	Low	Injection of a radioactive tracer
	fMRI	High	Low	High	High	None	
	fNIRS	Low	Low	Low	Low	None	

Table 2.3. Comparison of the different methods for the monitoring of brain activity.

The criteria can also be viewed as constraints for the use of devices in HMI. For instance, a relatively low financial cost is mandatory for a commercial use of an emotion detection system. Having devices that are easily wearable and non invasive is also essential since it is unlikely that users will accept to receive injections or have a heavy chirurgical operation just for using a new HMI. For this reason, as can be seen from Table 2.3, surface EEG is certainly the most applicable method for use in HMI systems. As a consequence, this sensor was chosen for emotion detection in this study. However EEG still requires quite some preparation, such as the application of a conductive gel on the surface of the scalp to insure contact between the skin and electrodes, and the plugging of each electrode in a tightly fitting headcap. This can be seen as a high constraint to go toward commercial applications but there are now wireless dry-electrodes systems available

on the market which do not require as much preparation and are quite cheap (for instance the NeuroSky[4] and Emotiv[5] devices).

Since in this thesis surface EEGs were used to monitor brain activity, the term EEG will now on refer to this type of measurement and not to EcoG and intra-cortical measurements. The electrical activity of a neuron has first to go trough several layers of different matter to reach an EEG electrode (gray matter, white matter, cerebrospinal fluid, bone and skin). The signal measured by an EEG electrode is thus the integration of the signals emitted by all the neurons that are close enough to be recorded. However, researchers have found that there are particular frequency bands (often called rhythms in the literature) of interest to interpret EEG signals: delta (2-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz) and gamma (> 30 Hz) bands. Notice that the frequency bands associated to each rhythm can vary from a study to another and from one person to another. Moreover, different names can be given to frequency bands according to the function that is associated to it, for instance the alpha band is also referred to as the mu-rhythm when considering the activity in the motor cortex. The reason why those rhythms are observed and fluctuate is still unclear but it is supposed that synchronization of several populations of neurons could be at the source of those phenomena [75].

Currently, EEG systems can record signals with more than 250 electrodes. Increasing the number of electrodes can improve the spatial resolution but increases the time necessary to apply gel and plug electrodes as well as the number of variables to analyze. In order for the EEG community to have some standards concerning the positioning and the naming of electrodes the 10-20 system was proposed [76]. Its name comes from the fact that the front and back median electrodes are positioned at 10% of the inion-nasion distance while the other electrodes are separated from each other by 20% of the same distance. The inion is a bone on the back of the skull and the nasion is the intersection of bones just above the nose. Latter, other systems extended the 10-20 system to define locations and names of more electrodes: the 10-10 system (74 locations) and the 10-5 system (345 locations) [76].

Noise is a critical issue in the measurement of EEG. According to Kronegg [28], there are at least three potential sources of noises. Environment noises concern all the electrical noises that surround the recording, for instance the 50Hz power line noise. Physiological noises refer to noises from the body such as muscles contractions. The last source of noise is the background activity of the brain and is defined as all brain activities that are recorded but are not related to the particular brain function under study.

[4] http://www.neurosky.com/ (retrieved on 29 April 2009)

[5] http://emotiv.com/ (retrieved on 29 April 2009)

c. The brain and emotions

The cognitive view of emotions supports the idea that brain structures are involved in emotional processes. To gather evidences supporting this hypothesis, researchers have first analyzed the behavior of patients suffering from brain damages and animals with segmented brain areas. In this way they could identify structures that regulate emotional expressions. The invention of new methods to monitor brain activity such as EEG and fMRI allowed for a more precise and easier identification of those structures as well as for the determination of their functions in emotional processes.

One of the first structures that was shown to be strongly involved in emotional expression is the **limbic system**. The elements composing this system are part of the temporal lobes and are placed under the cortex, deep in the brain. Papez was the first to suspect this system to be partially dedicated to emotions and found that the elements of the limbic system form a circuit, called the Papez circuit, that interconnect the cortex with the hypothalamus through structures such as the dorsal thalamus, the cingular gyrus, the hippocampus and the fornix [74]. Since the hypothalamus was known to coordinate emotional behaviors and expressions [74] this discovery supports the existence of a higher level of emotional processing occurring in the cortex. Later, an important structure was added as a part of this circuit: the **amygdala** [77]. It has been found that the amygdala is highly interconnected with sensory cortical areas suggesting that it could be at the source of the emotional attributions to sensorial stimuli (Figure 2.9). This hypothesis has been supported by many studies that shows the activation of this region at least for negative emotions such as fear and anger [78]. It was also proposed that the amygdala is involved in reinforcement learning [78, 79] and recognition of emotions [80].

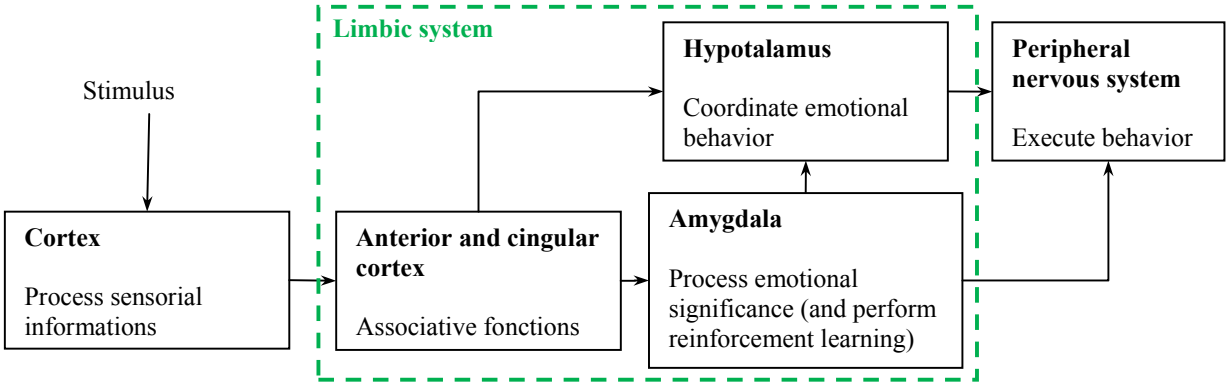


Figure 2.9. Principal structures of the limbic system together with their functions.

The other brain area that is assumed to play a key role in the elicitation of emotions is the **prefrontal cortex** [74, 78-81]. This area is located at the front of the frontal lobe and contains the orbito-frontal cortex (above the orbits of the eyes) that is known to be involved in cognitive

processes such as decision making. Davidson [19, 81] has shown that the prefrontal cortex is involved in approach and withdrawal reactions, which are closely linked with emotions, while Rolls [79] argues that the orbito-frontal cortex also plays a role in reinforcement learning of emotions.

Lateralization of those brain structures is also of interest. The most consistent findings concerning this hypothesis concern the prefrontal cortex. At least two types of lateralization have been observed. First Davidson [19, 81] has noticed that the left prefrontal lobe is more involved in emotional reactions corresponding to approach reactions while the right prefrontal lobe is more involved in withdrawal reactions. This was concluded by observing the power changes in the alpha band from EEG signals of participants subject to the two different types of stimuli. Some researchers also suggest that this lateralization corresponds to positive and negative emotions. However, this theory is discussed since the relation between approach-withdrawal reactions and positive-negative emotions is not direct. For instance, some negative stimuli can lead to approach reactions (anger for instance). Secondly, evidences suggest that the right hemisphere is more involved in emotional processing than the left one. Some studies also argue for asymmetric phenomenon in the amygdala but this question is still under study [78].

Finally, there is a high number of studies showing the activation of several other brain areas during emotional processes. An example is the work of Aftanas et al. [82] that showed significant differentiation of arousal based on EEG data collected from participants watching high, intermediate and low arousal images. This differentiation could be found in various areas such as parietal, parieto-temporal and occipital lobes. In [83] interactions between cortical regions during the presentation of emotional film clips were analyzed. It turned out that brain areas were differently synchronized depending on the type of stimulus. For instance, a higher synchronization between left and right frontal areas was observed for sad clips when compared to happy clips. Damasio also observed differences in brain activity during the feeling of self-generated emotions [84].

2.2.2 Peripheral nervous system (PNS)

The PNS is constituted of neurons that can be of two types: the sensory neurons that convey information from the sensory receptors to the CNS and the motor neurons that arouse or inhibit muscles and glands activity. The PNS can be divided in two subsystems: the somatic nervous system (SNS) that is connected to skeletal muscles and the autonomic nervous system (ANS) connected to autonomous muscles (the heart for instance) and glands. The SNS is associated with voluntary control of muscles while the ANS does not require conscious control to activate or inhibit connected organs. The SNS and the ANS are both involved in emotional reactions. For instance, the SNS is concerned with the voluntary production of facial expressions and the ANS

with involuntary facial expressions and physiological supportive functions (see section 2.1.3.b). Finally the ANS is also divided into two categories: the sympathetic nervous system and the parasympathetic nervous system. While the first prepares the organism for the uses of resources generally related to "flight and fight" responses, the second has the role of preserving and augmenting resources. Both of those systems are thus of interest in the study of emotions.

Since it is difficult and invasive to measure PNS electrical potentials by inserting electrodes directly into the nerves, PNS activity is generally indirectly assessed by measuring the activity of peripheral organs such as muscles. Below are described different measures of the PNS that will be used in this study.

a. Electrodermal activity

The ElectroDermal Activity (EDA) is related to the changes of electrical potentials and resistance of the skin. It was first measured by Féré [85]. While phenomenon inducing electrodermal responses are still not completely understood, it has been shown that they are strongly associated with sweat gland activity [86]. Sweat glands are situated deeply in the skin at the junction of the hypodermis and the dermis. A duct links a sweat gland to a pore situated at the surface of the skin (the epidermis) that can be closed or opened to let sweat spread out. Since sweat gland activity is known to be controlled by the sympathetic nervous system, EDA has become a common source of information to measure the ANS.

There are two methods to measure EDA:

- the **exosomatic** method consist of placing active electrodes (electrodes that inject a small current) on two sites of the skin to measure its resistance (or its conductance);
- the **endosomatic** method consists of placing two electrodes on the skin to simply measure the differences in skin potentials.

In both cases, since the sweat is a salty liquid that plays the role of a conductor, as soon as the ducts are filled with it, the electrical and impedance properties of the skin will change and be recorded by one the above settings. The palms of hands and foot, especially the fingers and the toes, are recommended places to position the electrodes because at these areas the skin contains many sweat glands.

The measured EDA can be decomposed in two different components [87]: the tonic level and the phasic response as highlighted in Figure 2.10. The tonic level represents the general resistance of the skin and is influenced by the hydration of the skin and the precedent sweat emissions. The phasic response is due to the accumulation of sweat in ducts and opening of the pores. Notice that it is not necessary that the sweat reaches the surface of the skin for the response to occur. This

response occurs from 1 to 4 seconds after a stimulus [88] and is often characterized by its latency (the time for the response to start after the stimuli), amplitude, rise time and half-recovery time (Figure 2.10) [87]. Sometimes these characteristics are hard to determine when there are several responses that overlap.

Depending on the method used to measure EDA as well as on the component of interest, several names can be found in the literature that refer to the same measurements: ElectroDermal Response (EDR), Galvanic Skin Response (GSR), Skin Resistance Level / Response (SRL and SRR), Skin Conductance Level / Response (SCL and SCR), Skin Potential Level / Response (SPL and SPR).

An increase of activity in sweat glands usually occurs when one is experiencing emotions such as stress or surprise. By asking participants to look at more or less arousing pictures of the IAPS, Lang et al. [7] discovered that the SCR amplitude is correlated with the level of arousal. Similar results where obtained in [89] showing that emotional words and neutral words can be differentiated based an EDA measures. By analyzing reactions to olfactory stimuli, Delplanque et al. [90] found a higher amplitude of the EDR for unpleasant than for pleasant odors. Moreover, EDA is known to be influenced by brain structures such as the hypothalamus, the limbic system and frontal cortical areas [89]. EDA can thus also be regarded as a window on emotional brain activity.

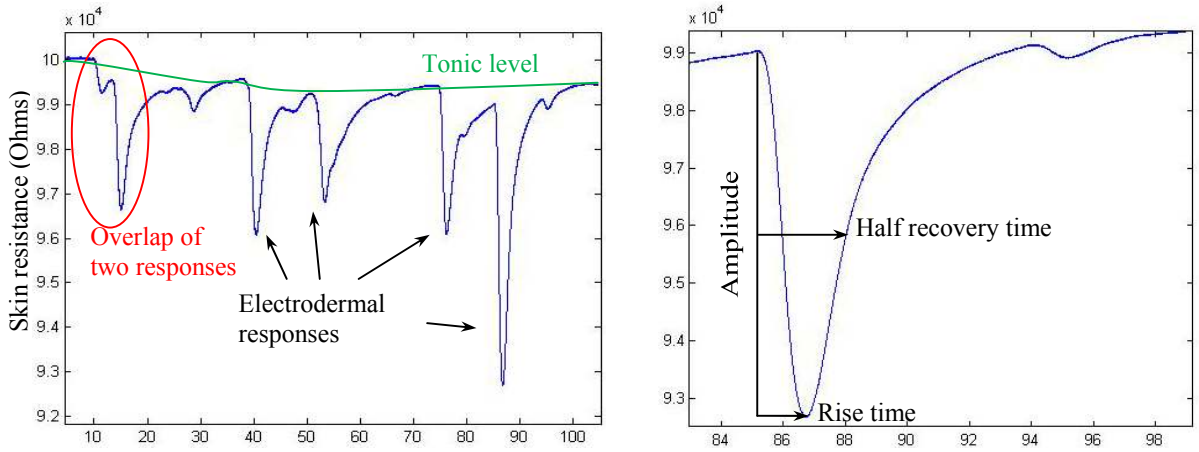


Figure 2.10. (left) Example of a signal representing the changes of resistance of the skin, (right) the characterization of an electrodermal response.

b. Blood pressure

The blood pressure (BP) is defined as the pressure that the blood exerts on the surfaces of the cardiovascular system (heart and vessels). Close to the heart the BP is the highest and decreases as the blood flows in the cardiovascular system. Changes in BP are influenced by cardiac output

(contractions of the heart), vasoconstriction and vasodilatation (reduction and augmentation of the vessels diameter by contraction and dilatation), and also by mechanisms occurring in other organs [91].

Generally two types of BP are distinguished: the systolic BP (SPB) is measured when the heart is contracted while the diastolic BP (DBP) is measured between two beats when the heart is relaxed. In turn, the pulse pressure is defined as the difference between SBP and DBP. Another value of interest is the mean arterial pressure (MAP) defined as the average BP. This value can be computed by averaging continuous measures of BP or by using an approximation formula based on the discrete measures of SPB and DPB [91]. Finally, it is also possible to compute pulse transit time (proportional to what is also called pulse wave velocity) defined as the time for the pressure wave to propagate from the heart to the area where the pressure is measured [91].

In order to avoid invasive methods, BP is generally measured using indirect methods. They can be divided in two types: intermittent methods that only allow for BP measures separated by several seconds and continuous methods that more or less continuously record the pressure signal. Manual or automatic cuffs tied around the arm to measure brachial artery pressure (sphygmomanometers) are examples of intermittent methods that measure SBP and DBP by identifying the so-called Korotkoff sounds. This method allows for 1 to 4 measures per minute [91]. By tying a photoplethysmograph around a finger it is possible to continuously record changes in blood pressure. The photoplethysmograph is a device that emits light and measures its reflection through the skin. Since light reflection varies according to the quantity of blood present in the vessels this sensor is able to measure Blood Volume Pulse (BVP). The obtained signal correlates with BP but can only provide relative changes of BP. Usually this sensor is placed at areas where there are many vessels near to the surface of the skin such as fingers and ear lobes.

Blood pressure has significant correlation with defensive reactions since these reactions are associated with vasoconstriction responses [87]. Several emotions are also known to be related to an increase or decrease of cutaneous blood flow giving rise to flushes or paleness [74]. Sinha et al. [92] refers to the increase of blood pressure during fear and anger as one of the most consistent findings in emotion research from autonomic activity. Moreover, they reported an increase in SBP during the visualization of emotional images compared to neutral images and an increase in DBP for anger compared to sadness. Finally, Lisetti and Nasoz [93] listed in their review two studies [94, 95] where SBP and DBP were found to be reliable indicators of task difficulty.

c. Heart rate

The Heart Rate (HR) is the number of heart contractions occurring over a given amount of time. It is generally expressed in Beats Per Minute (BPM) and is computed from InterBeat Intervals

(IBI), the time elapsed between two beats. An increase of HR can be due to an increase of the sympathetic activity or a decrease of the parasympathetic activity. Moreover, the cardiac response to sympathetic stimulations is slower than for parasympathetic stimulations giving rise to complex fluctuations of HR [96].

To better analyze those fluctuations Heart Rate Variability (HRV) is a common feature extracted from HR and IBI. Several methods can be used to compute variables related to HRV. One of the simplest is to consider the time series of IBI as a stochastic process and compute its statistical properties, such as standard deviation, to characterize the distribution of heart periods. It is also possible to compute the cardiac acceleration (i.e. the derivative of the HR). Another method consists of switching from the temporal to the frequency representation of the HR signal. In this case, three frequency bands are generally considered:

- the High Frequency (HF) band including frequencies between 0.15Hz and 1Hz;
- the Low Frequency (LF) band ranging from 0.05Hz to 0.15Hz;
- the Very Low Frequency (VLF) band from 0.0033Hz to 0.05Hz.

The exact boundaries of the different frequency bands are still under discussion in the literature and the given values are taken from [96]. The energy in the HF band is known to be principally mediated by the parasympathetic activity while energy in the LF band is influenced by the two autonomic systems. For those reasons many studies have proposed to use the ratio of the LF energy over the HF energy to reflect cardiac autonomic balance [96]. Other methods are also available to quantify HRV such as determining HR acceleration by computing the derivative of the HR signal.

In order to compute HRV features it is first necessary to identify heart contractions to construct the IBI and HR time series. This can be done non-invasively by analyzing the signals issued from several devices. One solution is to record heart electrical activity by placing electrodes at appropriate positions [97]. The recording is called an electrocardiogram (ECG or EKG). From an ECG it is possible to identify the peaks of the main waves, known as the R waves, which correspond to the main contractions of the heart. In order to measure HR, a microphone can be used to monitor heart sounds related to blood flow and valves activity of the heart. Finally, since heart pulses are visible in continuous BP and BVP signals, it is also possible to compute HR from this type of signals [97].

Increase or decrease of HR properties can be associated with different emotions [74]. For instance, Rainville et al. [6] observed an increase of mean HR for anger, fear, happiness and sadness compared to a neutral state. Moreover this increase was significantly different for fear

and anger, fear and happiness as well as for fear and sadness. In the same study, cardiac acceleration was also a relevant variable for the discrimination of those emotions while in [7] it was shown to be significantly correlated with pleasantness. Ekman [5] found that HR was able to separate happy, disgusted and surprised emotional states from angry, fearful and sad states. Concerning HRV, an energy shift from HF to LF frequencies is known to be associated with parasympathetic withdrawal reactions [96]. Finally, a decrease of energy in the HF bands was observed in [6] for fear and happiness compared to a neutral state while this energy was significantly different for fear and anger.

d. Respiration

Respiration involves organs such as lungs, airways and respiratory muscles. It is quantified by variables like lung volume, tidal volume (the quantity of air moved during inhalation and exhalation), pressure, air flow and breathing rate which can be influenced by resistive properties and contraction of the above-mentioned organs [98]. Breathing is mostly achieved by the contraction / relaxation of the diaphragm and the intercostals muscles that increase / decrease the thoracic volume. Contrarily to other measures described above, the respiration is operated by both the ANS and the SNS since breathing is generally involuntary but it is possible to control it for short periods of time. Concerning the ANS divisions both the sympathetic and parasympathetic divisions are implied in respiration.

Precisely measuring respiration characteristics requires the use of obtrusive equipment like spirometers for displaced air volume assessment and air-tight chambers for alveolar pressure computation (the last one being called body plethysmography). For a good review of such apparatus the reader is referred to [98]. A less obtrusive device that detects respiratory patterns is the respiration belt. It is attached around the abdomen and / or the chest in order to measure their expansion (Figure 2.11). Since this expansion approximates the quantity of inspired and expired air it is possible to measure the tidal volume with the proviso that calibration is performed beforehand and that two belts are used for measuring both chest and abdomen extension [98, 99]. Several types of belts exist, differing in the methods used to measure expansion (inductive sensors, piezo-electric sensor, etc.). One disadvantage of the belts is that they are relatively sensitive to movements. Temperature sensors placed under the nostrils and before the mouth can also be used to approximate respiration flow. However this method is not really reliable, especially during fast inspiration and expirations (Figure 2.11), prohibiting its usage for tidal volume computation.

The main function of respiration is to regulate the quantities of oxygen and carbon dioxide present in the blood to meet the needs of organs. However respiration is also influenced by emotional processes [98, 99]. For instance, a low breathing rate is linked to relaxation while

irregular rhythm, quick variations, and cessation of respiration correspond to more aroused emotions like anger or fear [6, 100]. Laughing is also known to affect respiration patterns, for instance by introducing high-frequency fluctuations in the signal measured by a respiration belt.

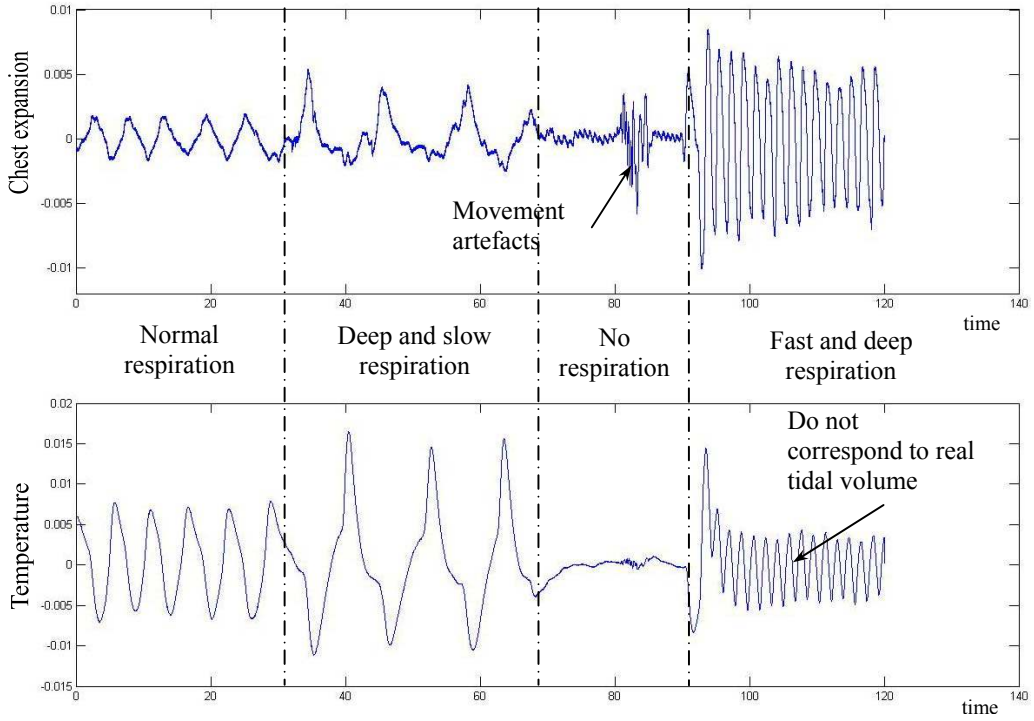


Figure 2.11. Examples of signals obtained from a respiration belt tied across the chest and a temperature sensor placed bellow the nostrils during different type of respirations.

e. Temperature

The internal temperature of the body is relatively stable while the temperature at the surface of the skin mainly depends on the surrounding temperature. The organism controls the internal temperature by balancing heat production and heat loss. Heat production is achieved by muscle contraction (such as shivering), by increasing chemical and metabolic activity and by vasoconstriction of skin blood vessels. Heat loss is obtained through reduction of metabolic activity, vasodilatation and sweating [101]. Since most of the heat is lost through the skin toward the surrounding environment, all those mechanisms also influence skin temperature.

Skin temperature can be recorded by standard temperature sensors like thermometers. Analog thermometers which allow for digital recording of temperature are generally based on the electrical resistance measurement of a piece of metal placed on the skin. To assess temperature, those devices make use of the predictable changes of the material resistance according to the heat. Less common devices include infrared thermal cameras. Since the amount of infrared light emitted by a surface is proportional to its temperature it is possible to use such a device to

monitor thermal changes. The main advantage of this device is that it is not necessary to attach a sensor on the skin, allowing for less obtrusive recordings.

Since skin temperature is influenced by vasoconstriction and sweat, both being related to emotions as discussed in the preceding sections, it is not surprising that researchers found it to be associated with emotional processes. Ekman [5] found a significant increase of skin temperature for anger compared to his five other basic emotions (sadness, happiness, fear, surprise and disgust). McFarland [102] found that stimulating persons with emotional music led to an increase of temperature for calm positive music and a decrease for excited negative pieces. In [103, 104] thermal images where investigated for the purpose of stress and anxiety detection. Finally, since skin temperature is also related to muscular activations, a thermal camera could be used to distinguish between different facial expressions [105].

2.2.3 Variability of physiological responses

As described in the preceding sections, physiological signals are influenced by several functions and features of the organism. They are thus very sensitive to modifications of the internal state but also to changes of the environment. For instance, a change of the surrounding temperature will impact skin temperature directly, BVP because of vasoconstriction and GSR because of possible change in sudation. A consequence is that the variability of physiological signals due to the process under consideration, in our case the emotion, is often much lower than the variability due to other phenomena. This last variability can then be considered as an important disturbance for an emotion assessment system.

The variability is generally decomposed into two components: **inter-participant** variability and **intra-participant** variability. Inter-participant variability accounts for the differences in physiological responses from one person to another. Those differences are due to several factors such as physiological characteristics, personal traits and behaviors. For instance body mass index, age and sex influence physiological responses while smokers and non-smokers may have different respiration patterns. Intra-participant variability accounts for the differences in physiological responses that can be observed for a given person. Those differences can be observed from day to day and can also be due to spontaneous events. Examples are a changes in mood from one day to another that influences EEG waves [81], and coffee absorption that modifies many physiological responses [91, 96, 97].

It is thus important to find a way to reduce the impact of those noise sources. Intra-participant variability can be reduced by having a high degree of control over the experiment, for instance by forbidding the ingestion of coffee or any drug before physiological recordings and keeping track of the mood of participants. It is worth to say that those methods can rarely be applied in the case of an emotion recognition system working in a real environment, generally far away from any

experimental control. Picard [106] emphasized the importance of including baseline monitoring to account for day to day variability in physiological signals. This method is equally important to reduce inter-participant variability. It consists in recording physiological signals during a "neutral" moment, before any emotional stimulation. The signals or the features extracted from the signals during emotion assessment are then expressed relatively to this baseline. Moreover, in emotion assessment from physiological signals, normalization of the signals for each participant is often performed to remove inter-participant variability [93, 106, 107]. Even if this method is quite effective, it is first required to have the range of all possible responses to normalize the signals which is generally not the case in real applications.

As shown in the preceding sections, different patterns of peripheral activation were found for a wide range of emotions. However, those findings are not always consistent since patterns which are ascribed to one emotion tend to vary between studies [108]. One explanation for this phenomenon is that the physiological activation that accompanies emotions is rather due to the tendency to action related to the emotion than to the emotion itself. In other words, the body is preparing for action (running in case of fear, aggressive action in case of anger, etc.) and this is the activity that can be measured on peripheral outputs. In [108] the authors proposed to unify those two views by developing the **context-deviation specificity** of emotions in which it is assumed that emotions have specific patterns of peripheral activation in the case where the conditions of emotional activation are similar.

These considerations are very important because they imply that any system dedicated to the recognition of affect should take into account the context specificity in its model of emotions. Though the goal of this thesis is not to design such a model, we believe that this concept should be kept in mind when one tries to perform affect recognition. From a practical point of view, taking into account all possible social, behavioral or psychological contexts of emotion elicitation seems a very hard task. Fortunately, to greatly reduce the number of contextual situations it generally suffices to concentrate on one particular application. From this point of view, it seems that a good protocol for emotion elicitation is not the one that will elicit emotion as close as possible to the real life emotions [107], but rather the one that will elicit emotions under controlled or monitored context close to the target application.

2.3 Emotion assessment from physiological signals

Over the last years, emotion recognition from physiological signals has received much interest. The goal is to find a computational model that is able to associate a given physiological pattern to an emotional state. This can be done by several methods discussed in this chapter.

Table 2.4 provides a (non exhaustive) list of relevant studies concerning emotion assessment from physiological signals. Unfortunately, it is difficult to make comparisons between these

studies because they differ on several criteria. Six criteria are introduced below to help the reader gain insight into the current state of the art as well as to discuss important aspects of emotion assessment from physiological signals. The studies are listed in Table 2.4 according to those six criteria.

Number of participants: in a study that includes a high number of participants the results can be regarded as more significant. Another point of importance is to know if the model obtained for emotion identification is user-specific or not. A user specific model avoids the problems related to inter-participant variability but a new model will have to be generated for each new user.

Emotion elicitation: how to elicit emotions from a participant is also a question of significant interest, especially considering the importance of the context as detailed before. Picard [106] proposed five factors to describe the context in which the emotion are elicited. One of those five factors divides emotion elicitation approaches into two categories: subject-elicited and event-elicited. In the first category, emotions can be generated by asking the participant to act as if he/she was feeling a particular emotion (for instance by mimicking the facial expression of anger) or to remember a past emotional event of his / her life. This method has often been used in facial expression recognition and it has been shown in [5] that it is effective to induce specific peripheral activity. In the second category, it is possible to use images, sounds, video clips or any emotionally evocative task. Several databases of stimuli have been designed for the purpose of emotion elicitation like the International Affective Picture System or International Digitized Sound system (IAPS, IADS)[54]. These databases are generally accompanied by affective evaluations from experts or average judgments of several people. However, since past experience plays a key role in emotion elicitation, it can also be important to ask the user to report and self-assess his / her feelings. Emotion elicitation is also influenced by the number and the complexity of the targeted emotions.

Time: temporal aspects are also relevant for emotion recognition but have only received little attention. According to [20], it is possible to define some time categories that range from the “full blown emotions”, lasting for some seconds or minutes, to the traits, lasting for years if not all the lifetime. In between are categories such as moods or emotional disorders. In human computer interaction, most of the applications under consideration deal with what Cowie defines as “full blown emotions” thus managing phenomena that last from seconds to minutes. In an ideal affective interface, the emotion of the user should be detected as soon as possible (let’s say in few seconds) in order to take the proper decision that directly matches the user expectation and not one that was expected minutes before. Synchronization of the different modalities is also an issue since the activation of physiological outputs can occur at different time scales. For instance, a change in temperature of the skin is much slower than a change in brain activity occurring a few milliseconds after emotion elicitation.

Ref.	Number of participants	Elicitation methods	Temporal aspects	Signals / sensors	Emotion classes	Methods (classifiers, feature selection, etc.)	Best results
Kim J. [100]	3 user specific	Music, AuDB (Augsburger database of bio-signals)	Time of a trial not specified 25 recordings over 25 days	Skin conductance, EMG, Respiration, ECG	Joy, anger, relaxation, sadness Positive / negative High / low arousal	SFFS feature selection, LDA with MSE SFFS feature selection, LDA with MSE SFFS feature selection, LDA with MSE	84% 84% 94%
Lisetti et al. [93]	29 not user specific	Film clips	70-231 s	GSR, heart rate, temperature	Sadness, amusement, fear, anger, frustration, surprise	Neural network with Marquardt backpropagation	84%
Rainville et al. [6]	43 not user specific[6]	Self Induction	90 s	ECG, respiration, skin conductance, EMG (zygomatic, masseter, corrugator)	Anger (15 part.), fear (15 part.), happiness (15 part.), sadness (17 part.)[6]	Step wise discriminant analysis	49%
Picard et al. [106]	1	Self induction	100 to 250 s 20 different days of recording	EMG, GSR, respiration, BVP, ECG	Neutral (no-emotion), anger, hate, grief, platonic love, romantic love, joy, reverence High / low arousal Positive / negative	SFFS-Fisher projection	81% 84% 87%
Katsis et al. [109]	10 not user specific	Driving simulation	Features computed on 10 s windows	EMG, ECG, respiration, GSR	High stress, low stress, disappointment and euphoria	SVM Adaptive neuro-fuzzy inference system	79% 77%

[6] The physiological activity of each participant was recorded for only one or two of the emotional conditions. The final number of participant for each condition is given in the "Emotion classes" column.

Ref.	Number of participants	Elicitation methods	Temporal aspects	Signals / sensors	Emotion classes	Methods (classifiers, feature selection, etc.)	Best results
Kim K.H. et al. [110]	50 children not user specific	Combination of story telling, visualisation and audio stimulus	50 s	Skin temperature, GSR, heart rate	Sadness, anger, stress Sadness, anger, stress, surprise	Subjects for training and others for testing SVM	78% 62%
Wagner et al. [111]	1	Music chosen by the participant	2 min 25 recordings over 25 days	EMG, ECG, GSR, respiration	Anger, sadness, joy, pleasure Positive / negative High / low arousal	LDA, KNN, MLP SFFS, Fisher, ANOVA	92% 86% 96%
Haag et al. [107]	1	images from IAPS	2 s several days	EMG, GSR, Skin temperature, BVP, ECG, respiration	Arousal Valence	Neural network for regression Accury is computed as the number of samples that fall in a 20% bandwidth of the correct value	97% 90%
Sibha et al. [112]	27 not user specific	Self induction	60 s 2 recording sessions on different days	ECG, GSR, finger temperature, blood pressure, EOG, EMG (zygomatic, corrugator, masseter, depressor muscles)	Fear, anger, neutral	LDA (first session as training set, second as test set)	67%
Takahashi et al. [113]	12 not user specific	Film clips	Time of a trial not specified	EEG, BVP, GSR	Joy, anger, sadness, fear, relaxation	Linear SVM one vs. all	42%
Leon et al. [114]	9 not user specific	images from the IAPS	6 s	Heart rate, GSR, BVP	Neutral, negative, positive	Autoassociative neural networks 1 participant for testing others for training	71%

Ref.	Number of participants	Elicitation methods	Temporal aspects	Signals / sensors	Emotion classes	Methods (classifiers, feature selection, etc.)	Best results
Sakata et al. [115]	16	Pictures	3 s	EEG Hear rate (results not presented)	6 emotions	LDA	29%
Rani et al. [116]	15 user specific	solving anagrams, playing pong	3-4 min 6 sessions on different days	ECG, GSR, bio-impedance, EMG (corrugator, zygomatic, trapezius), temp., BVP, heart sound	3 levels of intensity for: engagement, anxiety, boredom, frustration, anger. A classifier is trained independently for each emotion. Results are the average accuracy across participants and affective states	KNN Regression tree Bayes network SVM	79% 83% 78% 86%

Table 2.4. List of publications on emotion assessment from physiological signals. Signals acronyms are: Electromyography (EMG), Electrocardiogram (ECG), Galvanic Skin Response (GSR), Electroencephalography (EEG), Blood Volume Pulse (BVP). Classification acronyms are : Sequential Floating Forward Search (SFFS), Linear discriminant analysis (LDA), Support Vector Machine (SVM), Mean Square Error (MSE), Multi Layer Perceptron (MLP), K-Nearest Neighbors (KNN), ANalysis Of Variance (ANOVA).

Sensors / modalities: as described in chapter 2.2 various sensors can be used to measure physiological activity related to emotional processes. However, sensors used for emotion assessment should be chosen carefully so that they do not perturb the user. First, sensors should not be uncomfortable for the user in order to avoid parasite emotions like pain. Secondly, they should not prevent the use of classical modalities for instance by monopolizing the hands of the user. When physiological sensors are used for affective computing they switch from the standard status of sensors to the concept of modalities that communicate information about the emotional status of the user. It is then necessary to merge these modalities to perform emotions assessment in a reliable way, taking into account redundant and complementary information such as the relation that exists between HRV and respiration [117, 118].

Emotion models / classes: as discussed in chapter 2.1, several representations of emotions are available. Section 2.1.4 discusses their usability and extensions for the purpose of emotion assessment. In order to define a computational model that links physiological reactions to affective states, it is necessary to record physiological activity together with a ground-truth (i.e. the true elicited emotional state). However, constructing an emotional ground-truth is a difficult task because emotional states are influenced by several components of the organism as explained in Section 2.1.3.b. Several methods can be used to construct a ground-truth, for instance by defining the emotional state a-priori (generally based on precedent studies or evaluations of the stimuli) or by asking the participants to self-report their feelings. Depending on the methods used, the ground-truth can be quite different. The advantages and disadvantages of the ground-truth construction methods are discussed in Section 4.1.1.

Methods: a wide range of methods has been used to infer affective states. Most of them are part of the machine learning and pattern recognition techniques. Classifiers like k-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), neural networks, Support Vector Machines (SVM's) and others [119, 120] are useful to detect emotional classes of interest. Regression techniques [120] can also be used to obtain continuous estimation of emotions, for instance in the valence-arousal space. Prior to inferring emotional states it is important to define some physiological features of interest. It is very challenging to find with certainty some features in physiological signals that always correlate with the affective status of users. Those variables frequently differ from one user to another and they are also very sensitive to day to day variations as well as to the context of the emotion induction. To perform this selection researchers generally apply feature selection or projection algorithms like Sequential Floating Forward Search (SFFS) or Fisher projection.

As can be seen from Table 2.4 there are large differences in classification accuracy even for studies that employ the same number of classes. Although this can be partly explained by the factors detailed above, we believe that such variations mostly result from: the differences in

emotion elicitation strategies, the type of physiological signals (modalities) used, and the chosen model of emotions (or emotional classes).

For emotions elicited by an event, the best results obtained on more than three classes are those presented in [93] and [111]. One interesting point is that even though these studies use different stimuli (film clip versus music) they both use a method to control for the validity of the elicited emotions. In [93], the film clips presented were chosen according to the results obtained in a pilot study. In this pilot study, several participants were asked to evaluate clips by stating the emotion they felt as well as its intensity. In [111], the authors worked with the Augsburger Database of Biosignals (AuDB), the participants where asked to freely choose music that matched the desired emotions. Both of these methods insure that emotions felt during the experiment are intense and correspond to the expected ones. The good results obtained in [106] using a self induction protocol also tend to confirm the importance of reliable elicitation. In [106], the participant was the experimenter herself so that the emotions to be induced were perfectly clear to her. Moreover, the participant remembered past emotional events for emotion elicitation; this implies a strong intensity since remembered events are generally those that have induced intense feelings.

Diverse types of physiological activity measurements from both the peripheral and the central nervous system have been used. Up to now, most of the studies using brain activity for emotion assessment have used EEG signals with unconvincing results [36, 115]. One could conclude that EEG signals in the present state of the art are not effective for emotion assessment; the present thesis however will argue against this. Describing the state of the art for emotion recognition based on peripheral signals is a real challenge because most studies performed classification on feature sets that include features from many types of signals, thus preventing analysis for single modalities. However, there are signals that are employed in nearly all studies, like GSR and HR (extracted from ECG or plethysmography). These two signals are known to correlate well with affective states [5, 7, 92]. In many results it can also be seen that EMG signals are relevant for emotion assessment [6, 111, 112, 116, 121]. Generally, EMG electrodes are positioned to measure facial muscle activity. Muscles that are often recorded include: the venter frontalis (raising of eyebrows), the zygomatic (smiling) and the corrugator supercili (frowning of eyebrows). Since emotional states are often associated with facial expressions, measuring facial activity is strongly relevant to assess emotions. However having EMG, electrodes on the face is quite uncomfortable which could hamper their usage in a concrete application.

In order to assess emotions using physiological signals it is necessary to extract features from those signals that are relevant to the problem. In the current state of the art, most of studies [93, 106, 107, 122] consider physiological signals as independent Gaussian processes and thus extract statistical features such as mean, standard deviation and sometimes moments of higher order (skewness and kurtosis). However, it is likely that this independency assumption does not hold

since physiological signals are influenced by identical systems such as the ANS for peripheral signals. Moreover, correlations exist between the different signals because they are modulated by a single reaction to emotions. For instance both finger blood pressure and skin temperature are influenced by vasoconstriction. To our knowledge, there are no studies that try to consider those interactions at the feature extraction level. Most of the studies perform feature selection after feature extraction with the goal of removing redundant features and keeping relevant and synergetic ones. Other features frequently extracted in the temporal domain are statistical features (mean, standard deviation, etc.) of the first and second derivative of the signal [87, 93, 106, 113, 123]. Apart from temporal features, features from the frequency domain are also frequently extracted especially for HR, respiration and EEG [6, 109, 110, 115, 122, 124].

One of the most obvious observations that can be made from Table 2.4 is that different models of emotions lead to different classification accuracies. This is especially clear when comparing the basic-emotions models, which generally include more than three categories, to emotions in the valence-arousal space model, including two or three categories. Thanks to the works of Wagner [111] and Picard [106], it is possible to compare results on valence-arousal classes to those obtained on basic emotions classes in an intra-study framework. The conclusion emerging from this comparison is that the accuracies reported with the valence-arousal representation are lower than the accuracies reported with basic emotions, considering that the number of classes is different. A possible explanation for this is that acquisition protocols were designed for basic emotion elicitation while valence-arousal classes were defined by grouping basic emotions into two sets of classes (high vs. low arousal and high vs. low valence). This coarse definition can lead to confusing classes. For instance, grouping bliss and joy in the same positive class is prone to errors since peripheral activity for bliss, which is a rather calm feeling, is certainly different to the one of excited joy. However, results from other studies indicate that classification in the valence-arousal space is often associated with lower accuracies than when using basic emotion labels. Moreover, identification of valence classes is generally harder than identification of arousal classes (Table 2.4), which supports the idea that peripheral activity has a higher correlation with arousal than valence [7]. No clear differences in accuracy can be observed between the studies using different labels and number of classes to recognize basic emotions.

Only some of the studies listed in Table 2.4 performed emotion recognition online [25, 114, 125]. The others performed acquisition of the signals in a well controlled experimental environment and, in a second step, applied emotion recognition algorithms offline making use of cross-validation methods to determine the effectiveness of their system on unseen data. In [114] the authors first trained a model to associate physiological patterns to emotions and then applied the same model to detect emotions during the presentation of images from the IAPS. As can be seen from Table 2.4 the obtained accuracy on three classes (neutral, negative and positive) is quite

high (71%) demonstrating the feasibility of online emotion recognition. However, only one participant was tested with online detection preventing further conclusions concerning the reproducibility of this approach. Moreover, recordings and online classification where also performed in experimental conditions that are far from ecological and real application conditions. Another example of online emotion assessment is [125] where the authors designed a fuzzy emotion detection system for the purpose of studying emotions during game play. However, no online accuracy is reported for their system. Instead, the authors directly use the assessed emotions to compare different game play conditions. One of the most relevant studies concerning online emotion recognition is certainly [25] where pleasure (low vs. high) is assessed from the physiological signals of autistic children in order to adapt the behavior of a robot arm playing with them. The objective is to construct a robot able to change its gaming strategies as a psychologist would do. In this case the authors reported an average online accuracy of 81.1% when the two assessed emotional classes where compared to a ground-truth based on psychologist judgments. In this study, emotions were assessed while the children were interacting with the robot thus showing the possibility of emotion assessment for applications in a real HMI environment.

Chapter 3 Physiological signals recording and processing

3.1 Material

The material used for the acquisition of physiological signals is the Biosemi Active 2 System [126]. This system is composed of several components as detailed in Figure 3.1. The Biosemi system allows for simultaneous acquisition of EEG and peripheral signals. The analog signals are transmitted to the A/D converter for amplification and conversion to digital data. The data is then transmitted to the receiver through an optical fiber and finally to the PC for visualization and storage, both operated by the Actiview software (v. 5.21). Since the A/D converter is supplied with a low power battery and galvanic insulation from the main power supply is ensured by the optical fiber, there is no risk of electrocution. The system also provides a way to plug in external triggers that will be recorded with the other data sources. This is useful for synchronization with another acquisition system and for recording of manual event triggers (for instance by pushing buttons).

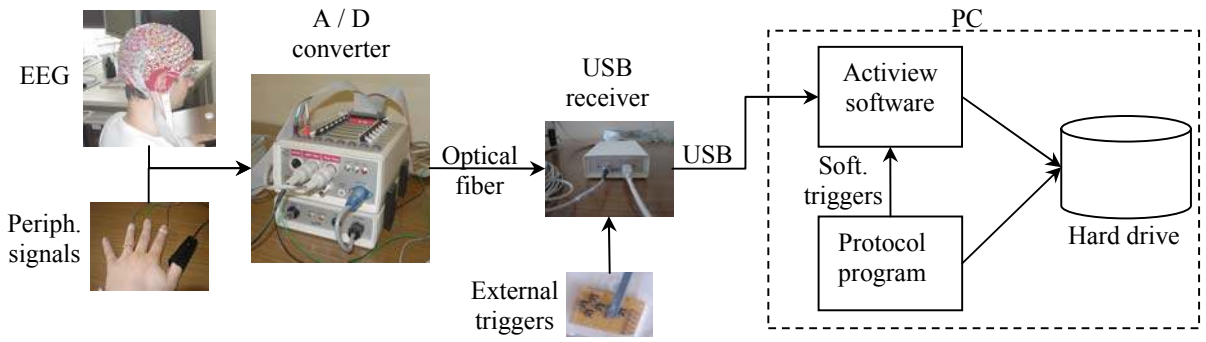


Figure 3.1. Hardware and software for signal acquisition.

During and after the recording of physiological data it is necessary to have a program that presents stimuli and records participant answers to electronic questionnaires. In Figure 3.1 this is named the "Protocol program". The protocol describes the scheduling of the different tasks that have to be performed by the participants during the experiment. It is important to synchronize the stimuli presented by the protocol program with the recorded signals to be able to retrieve the physiological signals that correspond to the events of interest. Since the Actiview software is implemented in Labview, the same platform was used to develop the protocol programs. In this way, we were able to send software triggers for the different events that were then recorded by the Actiview software together with physiological data. However, since the operating system was not real time there was nothing that could guarantee the precision of the software triggers. Experimentation showed that the lag between the stimulus presentation and the recorded trigger did not exceed 100 ms.

All the data were recorded in a room relatively immune to electromagnetic noise or in the Eckel C14 audiometric research chamber with electromagnetic insulation[7] (2.16m x 1.80m x 2.37m). In both cases aeration ensured that the ambient temperature did not increase due to the presence of a participant. The Faraday cage also provided acoustic isolation to ensure that participants were not disturbed by audio noise.

3.1.1 EEG

In order to avoid invasive recordings, surface EEG electrodes were used to monitor brain activity. The Biosemi system uses Ag-AgCl active electrodes to record electro potentials at the surface of the scalp which implies that a small current is applied on the skin. This design allows for noise and impedance reduction, low offset voltages and stable DC performance [126]. Two electrodes, the CMS and DRL electrodes, replace the usual ground electrode common to other EEG acquisition systems. As can be seen from Figure 3.2 which gives an overview of the names and positions of the EEG electrodes, the CMS and DRL electrodes were placed respectively on the left and right side of POz electrode position for all the recordings. The raw signals are referenced to the CMS electrode but it is necessary to re-reference them to obtain a better signal to noise ratio [126] (see Section 3.3.2).

In the different experiments performed, two electrode configurations were employed for EEG recordings. In the first configuration 64 electrodes were used to record an EEG with a relatively high spatial resolution (Figure 3.2). Those electrodes were positioned according the 10-10 system [76] which proposes positions for up to 74 electrodes. The second configuration consists of 19 electrodes placed according to the 10-20 system which allows for up to 21 positions [76]. The chosen electrodes are colored in green in Figure 3.2. The second configuration permits to reduce the time required to plug electrodes into the headcap. It is then possible to acquire data from more participants in a given amount of time, the main drawback being the loss of spatial resolution.

As can be seen from Figure 3.2 electrodes are plugged in electrode holders, themselves fixed on a cap attached on the head of participants. A cap is first chosen according to the head size of each participant and fitted on the head so that:

- Cz, Fpz and Oz electrodes are on the nasion-inion line (Figure 3.2),
- Fpz is at approximately 10% of the nasion-inion distance from the nasion,
- Oz is at approximately 10% of the nasion-inion distance from the inion.

[7] http://www.eckel.ca/ (retreived on 29 April 2009)

During the experiment, the cap was attached using Velcro straps to prevent displacement of electrodes. Before plugging an electrode, it is mandatory to fill in the corresponding electrode holder with a salty gel ensuring contact and conductivity between electrodes and the skin. The gel must be inserted using a syringe because of the small size of the electrode holders. Attention should be paid not to insert too much gel into the holders so that two different electrodes would be connected by a gel bridge and record the same broad activity. This is particularly important for the CMS / DRL electrodes since they are close to each other and a direct connection of those electrodes would reduce the signal to noise ratio for all electrodes. Two types of indicators were used to control that the electrodes were correctly plugged: impedance of the electrodes was kept below 5 KΩ and signals were visually verified. Visual verification consisted in ensuring that there is no sudden drift in the signals, controlling the appearance of alpha waves when the participant relaxes with eyes closed and observing the appropriate shape of the signals during eye blinks. Setting up the cap with 64 electrodes takes approximately 30 minutes for a trained experimenter.

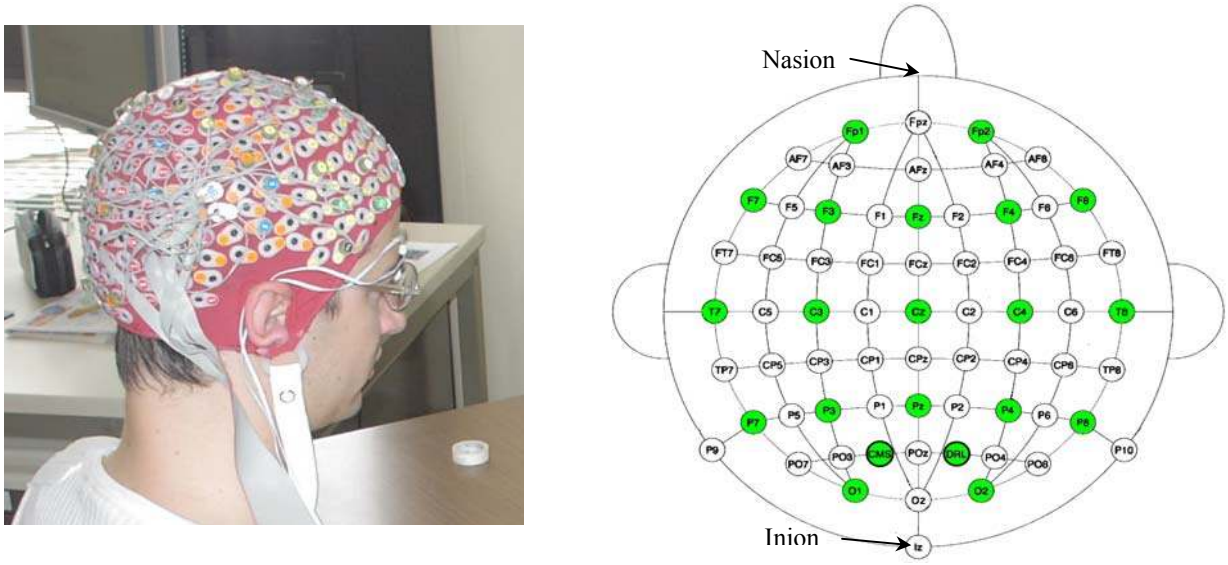


Figure 3.2. (left) A participant wearing the EEG cap with 64 electrodes plugged. (right) Top head view with the positions and names of the 64 electrodes used for EEG recording. For a 19 electrodes configuration only the green electrodes were used.

3.1.2 Peripheral sensors

The sensors used to monitor peripheral activity during emotional stimulations are presented in Figure 3.3. Apart from the respiration belt all the other sensors were attached on one of the hands. More precisely, they were placed on the left (resp. right) hand for right-handed (resp. left-handed) people to leave their preferred hand free to interact with the computer and answer questionnaires.

A **GSR** (Galvanic Skin Response) sensor, which measures the resistance of the skin in an exosomatic setting (see Section 2.2.2.a), was used to record EDA. This sensor is composed of two Ag-AgCl electrodes with a diameter of 8 mm that were placed on the distal phalanges of the index and middle fingers (Figure 3.3) as suggested by the guidelines of the society for psychophysiological research [127]. No electrolyte paste was applied and the GSR electrodes were directly fixed on the skin using medical tape. Proper contact with the skin was controlled using the indicators implemented in the Actiview software and by verifying the existence of an EDR when the participant is stimulated by a surprising event (clap of the hands hidden to the participant). Since the exosomatic setting implies that a small current is applied on the skin, the CMS / DRL electrodes have to be connected. If EEG was recorded simultaneously with GSR, the standard position was adopted (see section 3.2.1); otherwise the CMS / DRL electrodes were positioned close to each other, on the hypothenar eminences of the same hand as the GSR electrodes.

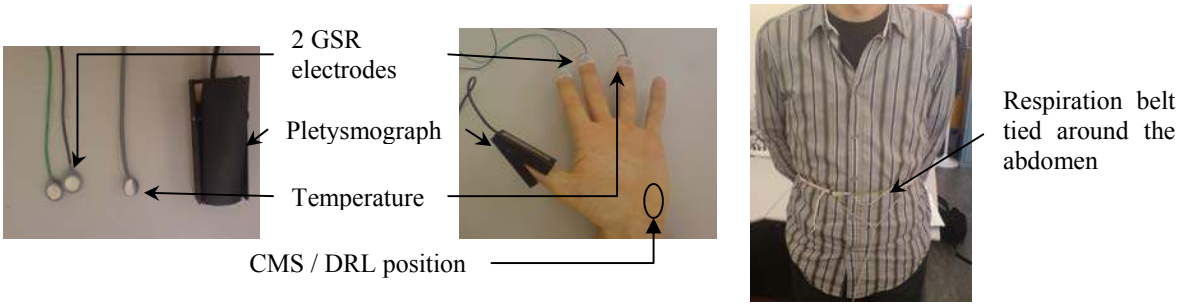


Figure 3.3. Pictures and positions of the sensors used to monitor peripheral activity. The CMS / DRL position was used only in the case where EEG activity was not monitored simultaneously with peripheral activity.

BVP (Blood Volume Pulse) was recorded by using a **photoplethysmograph** (later on called **plethysmograph**) that emits an infrared light and record the amount of light reflected by the skin (see Section 2.2.2.b). The plethysmograph was often clipped on the participant's thumb except if it did not fit or if bad signal quality was obtained at that place. In this case the little finger was used as the alternate position. The quality of the signal was verified by controlling the existence of a pulse including the usual peaks observed in BVP signals (see Section 3.3.3).

Respiration was monitored using a **respiration belt** tied around the abdomen of participants. When allowed by the participant, the respiration belt was placed directly in contact with the skin to avoid artifacts due to clothes movements. The respiration belt is useful to determine the respiration rate as well as deep breath depth. Since the chest expansion was not monitored together with the abdomen expansion it was not possible to determine the quantity of expired and inspired air (see Section 2.2.2.d). Between chest and abdomen positions, the abdomen was chosen because less noise was observed in the respiration signals at that position. Before the

beginning of each experiment, the quality of the respiration signals was controlled by observing their shape in different respiration conditions (slow / normal / fast respiration, deep breath).

Finally a **temperature** sensor was attached on the distal phalange of the ring finger with medical tape. Temperature measurements were assumed to be correct if they were around 33 °C.

3.2 Signals acquisition and preprocessing

3.2.1 Signal acquisition

All the signals were acquired at a sampling rate of 1024 Hz. This high sampling rate was chosen to have high time resolution EEG signals and it was applied to the other signal sources as well because they were plugged on the same acquisition system. At that sampling rate the bandwidth was of 268Hz due to an antialiasing filter that is applied to the signals by the Actiview system before storage on the hard drive. The complete bandwith was not used in the present studies; nevertheless it allows performing high frequency EEG analysis for future studies.

The physiological activity of each participant was recorded for the complete duration of the protocols. The obtained signals were then segmented in several trials (or epochs) each one being associated to a given stimulus. The starting time of each stimulus could be retrieved thanks to the triggers while the duration of a trial was defined by the protocols. Before extracting physiological features that are related to emotional activity, three preprocessing steps were applied to the signals: EEG and peripheral signals were denoised, EEG signals were re-referenced and the heart rate (HR) signal was computed from the non-filtered BVP signal.

3.2.2 Denoising

As a first step, the EEG signals were filtered by a 2-47 Hz Equiripple band pass filter (Table 3.1). This filter was applied to remove the DC offset of each electrode, drifts due to the difference of electrode impedance over time and power lines 50 Hz noise. This band pass filter also allows preserving frequency bands of interest for the study of emotional processes.

The peripheral signals were filtered by a moving average filter to remove noise. For this purpose we used filters of length 512 samples for GSR and temperature, 128 for BVP, and 256 for respiration (Table 3.1). Those different lengths were chosen to remove high frequencies without corrupting oscillations of interest in the different signals.

All the signals were filtered using the filtfilt function from the signal processing Matlab toolbox (v. 6.2.1) which processes the input signal in both the forward and reverse directions. This function allows performing a zero-phase filtering.

	High pass (-3dB)	Low pass (-3dB)
EEG	2 Hz	47 Hz
GSR	-	0.9 Hz
BVP	-	3.5 Hz
Respiration	-	1.7 Hz
Temperature	-	0.9 Hz

Table 3.1. Low and high pass cutoff frequencies at -3dB for the different filters.

3.2.3 EEG re-referencing

Since with the Biosemi system the original reference (signals are originally referenced to the CMS electrode) provides a poor signal to noise ratio, it is necessary to re-reference them afterward. To obtain a Laplacian reference the following Laplacian operator was applied to each electrode *i*:

$$x_i(n) = \tilde{x}_i(n) - \frac{1}{N_i} \sum_{j \in Neig(i)} \tilde{x}_j(n) \tag{3.1}$$

where $\tilde{x}_i$ is the CMS referenced signal of electrode *i*, x_i the Laplacian referenced signal, *n* the sample number, *Neig(i)* the neighbors electrodes of electrode *i* and N_i the size of this neighborhood. The neighborhood of an electrode was defined according to the Appendix B.

3.2.4 HR computation

As stated in Section 2.2.2.c an HR signal can be inferred from the BVP signal recorded by a pletysmograph. However, computing HR from a BVP signal is less reliable than from an ECG since the vaso-construction can influence the shape and timing of the heart pulses. As can be seen from Figure 3.4, the BVP signal is periodic because the blood pressure changes with heart contractions. Two points of interest can be identified in this signal: the foot of the systolic upstroke and the systolic peak both due to one of the heart contraction.

A method to determine HR from a BVP signal is proposed in [128]. This method is based on a complex analysis that identifies the systolic peaks as heart beats and requires recordings of long duration. For this study, this method was not used because:

- as can be seen from Figure 3.4 (right) it is sometimes difficult to identify which peak is which in a pulse, especially when the blood pressure is strongly increasing and decreasing or if the signal is of bad quality (noise due to movements, sensor badly placed or moved, etc.);
- in our experiments the trials generally lasted less than 10 seconds, the method was thus not adequate for these signals.

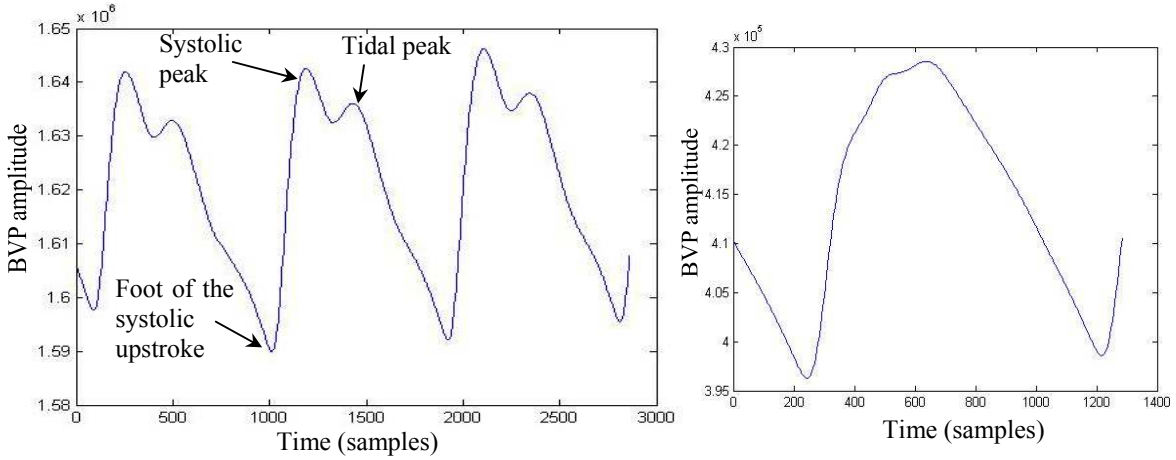


Figure 3.4. The heart waves in a BVP signal. (Left) Three pulses of the BVP signal with the different peaks, (rigth) example of a pulse where it is difficult to identify the different peaks.

In order to compute HR from signals of short durations without using the systolic peak, a method based on the detection of the foot of the systolic upstroke was implemented. The use of this point for identification of a heart beat is motivated by its frequent use for pulse wave velocity computation [96] (i.e. the time elapsed between the heart beat and the corresponding wave in a blood pressure signal). The developed method is composed of the following steps:

1. the linear trend of the BVP signal was removed from each trial to attenuate the effects of strong increase and decrease of blood pressure;

2. heart beats were assumed to be the local minima of the signal which were obtained by finding samples were the derivative is zero and the amplitude is switching from a decrease to an increase;

3. in the case where two such beats fall in the same interval of 0.5 second then only the beat that corresponds to the highest increasing BVP derivative is kept. The 0.5 second interval was chosen based on the assumption that the HR will not exceed 120 beats per minutes (BPM) which is somehow reasonable since in all the protocols participants were sitting in front of a computer screen without performing any significant physical activity;

4. the interbeat intervals (IBI) were computed as the time elapsed between two consecutive beats which could be converted to *B-1* HR values corresponding to the *B* detected heart beats. The time stamp of each HR value was placed in the middle of the corresponding IBI interval.

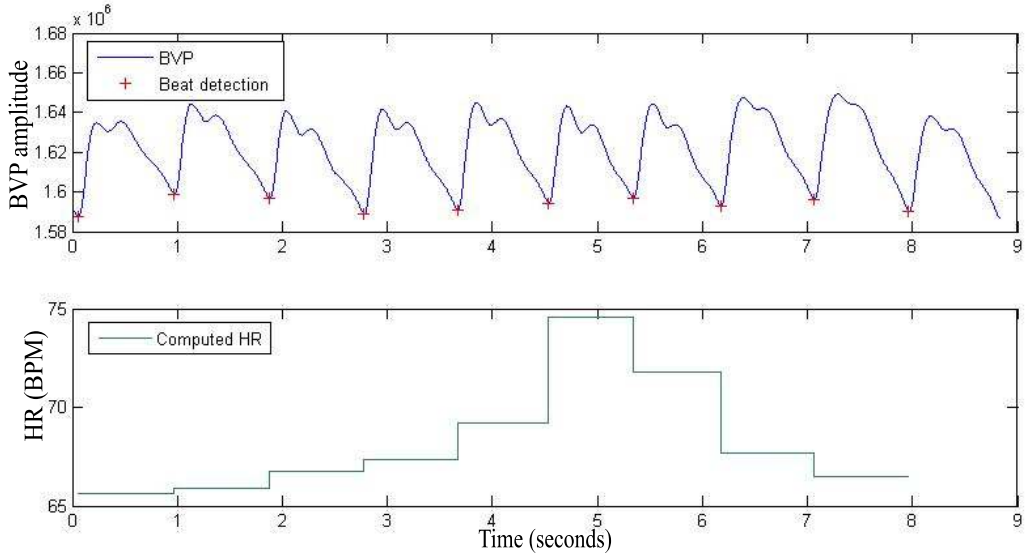


Figure 3.5. Example of the beat detection and HR computation algorithm on a 9 seconds signal. The HR signal is represented as a staircase function with the length of a step corresponding to the duration of an IBI.

As can be seen from Figure 3.5 this method performed fairly well on short duration trials (less than 10 seconds) but it was found to be less reliable on signals from one of the protocols where the length of a trial was longer (5 minutes) and signals were noisier. To improve the reliability of the peak detection, an algorithm was designed to detect and correct the falsely detected heart beats a posteriori. It is composed of two main steps described below.

1. Detection and correction of false positive peaks (a beat is detected but this is not a true one):
    a. for the *i*th IBI the median m_i of the 5 precedent IBI's was computed (except for the 5 first IBI);
    b. if $m_i - IBI(i) > \eta$ then the *i*th IBI is considered as corrupted, η being a parameter of the algorithm;
    c. for each corrupted IBI check if removing one of the two corresponding peak will solve the problem, in this case this peak is removed.
2. Detection and correction of false negative peaks (a beat has not been identified):
    a. similarly to step 1.b. if $m_i - IBI(i) < -\eta$ then the *i*th IBI is considered as corrupted;
    b. for each corrupted IBI *i* the number P_i of peaks to add was determined by $P_i = round(IBI(i)/m_i) - 1$; if adding those peaks does not violate the constraints

1.a and 2.a then the peaks were added in this IBI to construct P_i+1 new IBI's with equal duration.

The parameter η was empirically set to 0.2 seconds since it corresponds to reasonable changes in HR and it detected nearly all the false positive and false negative peaks.

3.3 Characterization of physiological activity

For each trial the physiological activity was characterized by computing several features from the signals. It is then possible to concatenate the different features to construct a feature vector associated to a trial. The current section explains for each feature why it has been chosen as well as the time constraints necessary for their accurate computation.

As explained in Section 2.2.3 it is important to use a baseline to account for inter and intra participant variability. However, most of the classifiers used in this thesis are participant specific which means that a classifier was trained for each participant. There is thus no need to account for inter-participant variability. Moreover, the physiological activity of each participant was recorded in a single session within a controlled context (stable environment temperature, no possibility to eat, etc.) alleviating the need to account for intra-participant variability. While baselines were computed for all protocols, subtracting baselines to the signals did not increase the accuracy of participant specific classification. For this reason the baselines were not used except for the video-game protocol presented in Chapter 7 because in this protocol a classifier was designed independently of the participant. The baselines computed for this protocol are presented in Chapter 7.

3.3.1 Standard features

This section described the physiological features that can be considered as standard because of their frequent use for emotion assessment [87, 93, 106, 113, 123]. Only the features that were used in this study are described.

By assuming that each measured signal is generated by a Gaussian process with independent and identically distributed samples, the two physiological features that can be used to characterize a physiological signal are its mean and its standard deviation:

$$\mu_x = \frac{1}{N_s}\sum_{n=1}^{N_s} x(n) \tag{3.2}$$

$$\sigma_x = \sqrt{\frac{1}{N_s}\sum_{n=1}^{N_s} (x(n)-\mu_x)^2} \tag{3.3}$$

where *x(n)* is the value of the signal *x* (an EEG electrode signal, GSR signal, etc.) at sample *n* and N_s is the number of samples in a trial.

In order to evaluate the trend of a signal *x* over a trial, the average of the signal derivative can be computed as:

$$\delta_x = \frac{1}{N_s - 1} \sum_{n=1}^{N_s - 1} (x(n+1) - x(n)) = \frac{x(N_s) - x(1)}{N_s - 1} \tag{3.4}$$

Finally the maximum and minimum of a signal can also provide information concerning the range of the signal amplitude:

$$Min_x = \min_n x(n) \tag{3.5}$$

$$Max_x = \max_n x(n) \tag{3.6}$$

Those features are quite general and can be applied to a wide range of physiological signals (EEG, GSR, EMG, etc.). Their usefulness (or weakness) will be discussed in the Section 3.4.2 for the different types of signal.

3.3.2 Advanced features

This section proposes specific features for each signal type based on the variables that are often analyzed in psycho-physiology.

a. EEG

The cognitive theory of emotions provides a strong motivation to go toward emotion assessment using signals from the CNS. As described in Section 2.2.1.c, several researchers have shown the involvement of brain structures in emotional processes. When using EEG to record emotional activity, most of the results were obtained by comparing the energy in different frequency bands. For instance, Davidson [81] demonstrated the lateralization of the frontal cortex by studying the energy of the EEG alpha waves. Aftanas et al. [82] reported energy differences between more or less arousing visual stimuli (images from the IAPS). Those last differences were observed from the energy in theta bands for the parietal and occipital areas, alpha bands for the frontal areas and gamma bands for the complete scalp. However, energy is not the only feature that can be used for the purpose of emotion assessment since studies also demonstrated that there are specific patterns of synchronization between brain areas during emotional processes [61, 129]. For those reasons different types of features were defined to characterize EEG signals. They were regrouped in five feature sets according to the type of features extracted.

Energy

This set of features, named *EEG_FFT*, was defined to represent the energy of EEG signals in frequency bands known to be related to emotional processes [81, 82]. For each electrode *i*, the energy in the different frequency bands displayed in Table 3.2 was computed over all samples belonging to a specific trial, using the Fast Fourier Transform (FFT) algorithm. This was done under the assumption that the EEG signals are stationary over the whole duration of a trial.

Feature for electrodes i	Frequency band
θ_i	4-8 Hz
α_i	8-12 Hz
β_i	12-30 Hz

Table 3.2. The energy features computed for each electrode and the associated frequency bands.

Moreover, the following *EEG_W* feature was added because it is known to be related to cognitive processes like workload, engagement, attention and fatigue [130-133]. To compute this feature the energy in each frequency was summed over the N_e electrodes:

$$EEG\_W = \log\left(\frac{\sum_{i=1}^{N_e}\beta_i}{\sum_{i=1}^{N_e}\theta_i + \alpha_i}\right) \tag{3.7}$$

The feature vector $\mathbf{f}^{\text{EEG\_FFT}}$, which corresponds to one trial, is thus composed of $3.N_e$ +1 features and the *EEG_FFT* feature set contains all the $\mathbf{f}^{\text{EEG\_FFT}}$ vectors.

Lateralization of EEG alpha waves

The results obtained in [19] are at the origin of the creation of this feature set called *EEG_Lateral*. This study demonstrated the correlation between a EEG asymmetry score in the frontal region and reported measures of general tendency to approach or withdraw. Similar results have also been reported, with less reproducibility, for stimuli of negative and positive valence [58, 78]. The asymmetry score was computed according to the following formula:

$$AS = \log(P_R) - \log(P_L) \tag{3.8}$$

where P_R and P_L are the power of the EEG signal in the alpha frequency band (defined as the 8-13 Hz band) respectively for a given right electrode and the left symmetrical electrode (for instance electrodes F4 and F3).

In this study the asymmetry score was computed for several pairs of electrodes, including non-frontal electrodes.

Asymmetry score	Right electrode	Left electrode
AS_1	Fp2	Fp1
AS_2	AF4	AF3
AS_3	F4	F3
AS_4	FC4	FC3
AS_5	C4	C3
AS_6	CP4	CP3
AS_7	P4	P3
AS_8	PO4	PO3
AS_9	O2	O1

Table 3.3. The pairs of electrodes used to compute 9 asymmetry scores.

For each trial, the alpha power of a given electrode was computed from the FFT applied on the complete duration of the trial; the asymmetry scores were then computed and a feature vector $\mathbf{f}^{EEG\_Lateral}$ was constructed by concatenation of the asymmetry scores:

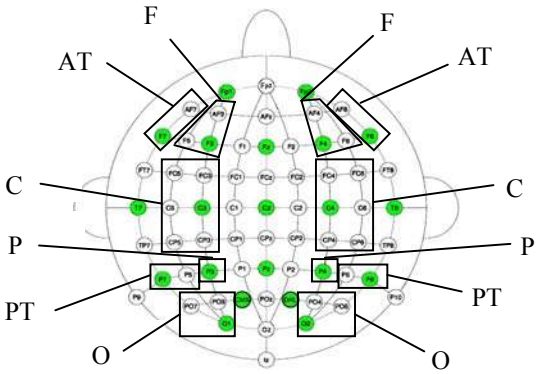
$$\mathbf{f}^{EEG\_Lateral} = [AS\_1, \ldots, AS\_9] \tag{3.9}$$

Finally, the *EEG_Lateral* feature set is composed of the ensemble of the $\mathbf{f}^{EEG\_Lateral}$ feature vectors obtained for each trial.

Average energy over brain areas

This feature set was computed to assess emotions elicited by the visualization of images with emotional content, as described in Chapter 5. The choice of those features was based on the study of Aftanas et al. [82] who showed a correlation between arousal elicited by images and responses in particular frequency bands and brain areas. As can be seen in Figure 3.6, areas were defined as groups of several electrodes (e.g. there are 6 electrodes in area PT, P, O). This figure also indicate the name of the rhythms of interest (θ1, θ2, γ, etc.) together with the associated frequency bands.

Power values of the 6 frequency bands listed in Figure 3.6 were computed for each electrode and for the whole duration of a given trial using the FFT algorithm. As several electrodes are located in the same area, the power over all these electrodes were averaged yielding a total of 6 features for this EEG feature set (e.g. feature one is the average power in band $\theta 1$ over all electrodes in areas PT, P, O). Most of the features concern the Occipital (O) lobe, which is not surprising since this lobe corresponds to the visual cortex and subjects were stimulated with pictures [82].



EEG feature	Location area	Frequency band
EEG_Area_1	[PT;P;O]	$\theta 1$ (4-6Hz)
EEG_Area_2	[PT;P;O]	$\theta 2$ (6-8Hz)
EEG_Area_3	[PT;P;O]	γ (30-45Hz)
EEG_Area_4	[AT;F]	$\alpha 2$ (10-12Hz)
EEG_Area_5	[AT;F;C]	$\beta 1$ (12-18Hz)
EEG_Area_6	[C;PT;P;O]	$\beta 3$ (22-30Hz)

Figure 3.6. Top head view with EEG electrode locations and corresponding frequency bands (from [82]).

For each trial a feature vector $\mathbf{f}^{EEG\_Area}$ was constructed by concatenating the 6 EEG features:

$$\mathbf{f}^{EEG\_Area} = [EEG\_Area\_1,\ldots,EEG\_Area\_6] \tag{3.10}$$

and the *EEG_Area* feature set was defined as the set of all vectors computed from the different trials of a given protocol.

Spectrogram

In the two feature sets, energy features were computed from EEG signals by applying the FFT algorithm on the whole duration of a trial. This could be done reliably under the assumption that the EEG signals are stationary for the duration of each trial. However this is rarely true since EEG signals can only be considered as stationary on short periods of time. This comment also applies for the computation of mean and variance features. To solve this issue the following feature set was defined under the assumption that EEG signals are stationary on short time windows of 0.5 seconds.

This set of EEG features was extracted by computing the Short-Time Fourier Transform (STFT) for each electrode with a sliding window of 512 samples and 50% overlap between consecutive windows. This window size was chosen in order to have a frequency resolution of Δf = 2Hz which allows to separate the different rhythms observed in EEG signals while maintaining a time resolution of 0.5 second. For each of the spectrograms (one per electrode), we selected 9 frequency bands of 2Hz ranging from 4Hz to 22Hz. This range was chosen because it includes most of the frequency bands used for the computation of the *EEG_FFT* and *EEG_Area* features set.

The $\mathbf{f}^{EEG\_STFT}$ feature vector for a given trial is then constructed by concatenating all the power values of the 9 frequency bands at the different time frames and for each electrode. The length of $\mathbf{f}^{EEG\_STFT}$ is thus:

$$length(\mathbf{f}^{EEG\_STFT}) = 9 \times N_e \times \lfloor 4T - 1 \rfloor$$

where 9 is the number of frequency bands, N_e the number of electrodes and $\lfloor 4T - 1 \rfloor$ gives the number of time frames for a trial of duration T seconds. The feature set composed of all the $\mathbf{f}^{EEG\_STFT}$ feature vectors (one per trial) is named *EEG_STFT*.

Mutual Information features

In this feature set, mutual information (MI) between pairs of electrodes is proposed as a measure of statistical dependencies between different areas of the brain. This set of features was motivated by studies that demonstrated synchronization of brains areas in emotional processes [61, 129]. With the assumption that the signal x_i of electrode i for a given trial is a stochastic process with probability mass function $P(X_i)$, the mutual information between electrodes i and j for this trial is expressed as:

$$I(X_i; X_j) = H(X_i) - H(X_i \mid X_j)$$

$$H(X_i) = -\sum_{x_i} P(X_i = x_i) \log(P(X_i = x_i)) \tag{3.11}$$

$$H(X_i / X_j) = -\sum_{x_i, x_j} P(X_i = x_i, X_j = x_j) \log(P(X_i = x_i / X_j = x_j))$$

where $H(X_i)$ and $H(X_i \mid X_j)$ are respectively the entropy and conditional entropy of random variables X_i and X_j. Mutual information was computed using Moddemeijer's Matlab toolbox [134][8] that estimates the different distributions based on histograms and automatically determines an appropriate bin size.

The MI feature vector $\mathbf{f}^{EEG\_MI}$ of this trial is then constructed by concatenation of mutual information between each pairs of electrodes:

$$\mathbf{f}^{EEG\_MI} = [I(X_1, X_2) \ldots I(X_1, X_{N_e}), I(X_i, X_{i+1}) \ldots I(X_i, X_{N_e}), I(X_{N_e-1}, X_{N_e})] \tag{3.12}$$

The total number of features of a trial for N_e electrodes is then: $\sum_{i=1}^{N_e-1} i = \frac{(N_e - 1)N_e}{2}$. The feature set containing all feature vectors was named *EEG_MI*.

[8] available at http://www.cs.rug.nl/~rudy/matlab/ (retrieved on 27 April 2009)

b. Galvanic Skin Response (GSR)

The mean skin resistance over a trial has been shown to be correlated with the arousal of a stimulus [7] and has been widely used for emotion assessment from peripheral physiological signals. An aroused emotion should also induce a decrease in the GSR signal. For those reasons mean skin resistance and the trend of the GSR signal were frequently added to the peripheral feature sets. However, as described in Section 2.2.2.a, the speed of the fall of a GSR signal is also of importance to characterize the EDA. To approximate this value the average decrease rate during decay time was defined as follows:

$$f_{GSR}^{DecRate} = \frac{1}{N_n} \sum_{n/GSR'(n)<0} GSR'(n) \tag{3.13}$$

with N_n being the number of samples where the GSR derivative GSR' is negative. The proportion of negative samples in the derivative was also computed to characterize the duration of GSR falls:

$$f_{GSR}^{DecTime} = \frac{N_n}{N_s - 1} \tag{3.14}$$

While these features can correctly characterize GSR signals when it is expected that only one EDR occurs, it is important to count the number of EDR's if several responses occur in a single trial. For this purpose an automatic EDR detection algorithm called *nbPeaks* was designed. In a first step this algorithm identifies consecutive high and low peaks of a GSR signal by finding the sign changes of the signal derivative. High peaks were identified as potentially being the beginning of an EDR whereas the low peaks were identified as the potential apex of a response. Finally a response was counted if the following criteria were met:

- the difference in amplitude between the beginning and the apex of a response is higher than 200 Ohms;
- the time elapsed between the beginning and the apex of a response is between 1 and 5 seconds.

Thus the last feature computed for the GSR signal is:

$$f_{GSR}^{NbPeaks} = nbPeaks(GSR) \tag{3.15}$$

Concerning the time aspects, the EDR is known to occur from 1 to 4 seconds after a stimulus [88]. It is thus important to record a GSR signal for more than 4 seconds after the presentation of an emotional stimulus to be sure that at least part of the reaction is recorded by the sensor.

c. Blood volume pulse (BVP)

The BVP signal given by the pletysmograph is a relative measure of the blood pressure (BP): it is correlated with the BP but does not measure its exact value. Since variations of BP and blood flow are known to be associated to several emotions (Sinha, Healey, Section 2.2.2b) the features computed from the BVP signal were the mean, the standard deviation and minimum / maximum values as described in Section 3.4.1.

d. Heart rate (HR)

The increase and decrease of HR is associated to many emotions (Section 2.2.2.c) thus the average HR computed over the whole duration of an emotional stimulus is a feature that can be used for emotion assessment. Another variable of interest for the study of emotions is the Heart Rate Variability (HRV). As detailed in Section 2.2.2.c the HRV can be determined from the HR power spectrum and by computing HR standard deviation.

Three components in the HR spectrum (frequency bands) of interest are generally considered to characterize HRV: a High Frequency band (HF, 0.15Hz-1Hz), a Low Frequency band (LF, 0.05Hz-0.15Hz) and a Very Low Frequency band (VLF, 0.0033Hz-0.05Hz). As a consequence, it is necessary to have signals of significant duration to correctly record HRV components. For instance the committee report of the Society for Psychophysiological Research [96] recommends using epochs of at least 1 minute to reliably compute the HF component and 2 minutes for the LF component. Those durations correspond to approximately 6-10 times the length of the wave with the lowest frequency (0.15Hz for HF and 0.05Hz for LF). Actually, it is possible to determine (less reliably) the energy in the HF band on epochs of 6.6 seconds and in the LF band on epochs of 20 seconds. If HRV is estimated using the standard deviation of heart rate, then it is computable on any period of time including at least 2 HR values (3 heart beats, which generally corresponds to less than 3 seconds); the result will however only provide information about the energy in frequency bands limited by the duration of the epoch.

The VLF component was not investigated in this study since it needs too long trial duration for its computation. Before computing energy in the HF and LF frequency bands the HR signal was interpolated with a cubic interpolation to obtain a new signal sampled at 1024Hz. This first step allows to:

- transform the original signal that has a low sampling rate and which is not sampled at regular intervals into a regularly sampled signal with a high sampling rate;
- obtain a signal with the same sampling rate than the others which facilitates the comparison of the signals over time (correlation measures, etc.).

The FFT algorithm was then applied to the interpolated HR signal to compute the power in the two frequency bands of interest. Finally, the ratio between the power in the HF and LF frequency bands was computed to reflect the cardiac balance between the sympathetic and parasympathetic activity (see Section 2.2.2.c). Table 3.2 summarizes the three features computed from the HR power spectrum.

Feature name	Description / formula
f_{HR}^{LF}	Power of HR in [0.05-0.15Hz]
f_{HR}^{HF}	Power of HR in [0.15-1Hz]
$f_{HR}^{LF/HF}$	$\frac{f_{HR}^{LF}}{f_{HR}^{HF}}$

Table 3.4. The three features computed from the HR

e. Respiration

The respiration rate ranges from 0.1 to 0.35 breaths / second at rest, while it can reach 0.7 breaths / second during exercise. Moreover, if respiration is measured with a belt, irregular respiration lead to energy increases in higher frequency bands. Laughing also affects the respiration pattern by introducing high-frequency fluctuations on the recorded signal. To measure the main respiration frequency (i.e. respiration rate) at rest it is thus necessary to have a signal of at least 10 seconds (based on the lower bound of the respiration rate). Nevertheless, shorter periods of time could still contain interesting information since respiration rate at rest can increase up to 0.35 breaths / second which can be measured on an epoch of around 3 seconds. To capture those fluctuations, features from both the frequency and time domain are therefore used.

Features of the frequency domain are obtained by computing the FFT of the original signal. Then six features were computed to measure the power of the respiration signal in consecutive frequency bands. Table 3.5 summarizes the different frequency bands of interest. Those frequency bands were concatenated in a feature vector called $\mathbf{f}_{Resp}^{Pow}$.

Power feature	Frequency band	Power feature	Frequency band
f_{Resp}^{Pow1}	[0-0.25Hz]	f_{Resp}^{Pow4}	[0.75-1Hz]
f_{Resp}^{Pow2}	[0.25-0.5Hz]	f_{Resp}^{Pow5}	[0.75-1.25Hz]
f_{Resp}^{Pow3}	[0.5-0.75Hz]	f_{Resp}^{Pow6}	[1.25-1.5Hz]

Table 3.5. The power features computed from the respiration signals and being part of the $\mathbf{f}_{Resp}^{Pow}$ feature vector.

The main frequency of the respiration signal was considered to represent the respiration rate and is computed as:

$$f_{Resp}^{Rate} = \arg \max_{0.1<f<0.7} P_{Resp}(f) \tag{3.16}$$

where $P_{Resp}(f)$ is the power of the respiration signal at frequency f. The precision of this feature depends on the frequency resolution obtained from the FFT which depends on the duration of the signal used for the FFT computation.

Finally, to identify the deepest inhalation or exhalation in a respiration signal *Resp*, the dynamic range feature was computed as:

$$f_{Resp}^{DR} = \max_{n} Resp(n) - \min_{n} Resp(n) \tag{3.17}$$

f. Temperature

Generally, changes in skin temperature are quite slow and difficult to detect on short time periods. However, since skin temperature is influenced by vasoconstriction and sweat, it is possible to observe short thermal reactions of a few seconds. For instance, Ekman [5] showed that the average value of skin temperature measured on a hand over an epoch of 10 seconds could be used to distinguish anger from fear and sadness. For this reason mean temperature was used as a feature in this study. The average derivative of the temperature signal was also used to indicate the trend of the signal.

3.4 Ethical and privacy aspects

Ethics is related to moral, law, honesty and privacy issues [135]. It is not often that researchers in computer science have to deal with those issues. However, there is an ever expanding trend in recording and using personal data for various domains of application such as sport, gaming and health. Due to its very private nature, this type of personal information requires that all necessary steps be taken to ensure privacy. In order to perform experiments on physiological emotion assessment it is necessary to acquire physiological data from persons while they are experiencing emotions. This raises several ethical issues since the human being is directly involved in those experiments and sensitive data is acquired.

This section describes all the steps that were taken to ensure ethical aspects. The binding regulations are European (The European Charter of Fundamental Rights, Art. 3, 8, 13), Swiss (Swiss Federal law from June 19, 1992 on data protection (LPD)), as well as at the University level (University of Geneva regulations concerning integrity in research). Rules and key points to address were also derived from the European commission document concerning ethics for researchers [135].

The recordings were performed with a limited number of **volunteers** (ranging from 4 to 11 persons depending on the experiment)**.** All of them were adults, researchers from our department or students interested in our work. It was controlled that participants were able to understand and question the experiment as well as give their participation approval on their own. The participants also needed to accept spending some time to do the experiment but no other selection criterion was applied.

All volunteers were carefully **informed** about the experiment and had to sign a **consent form** (Appendix A) to ensure that they were aware of:

- the purpose and scientific goals of the research (in human-computer interaction);
- how the experiment is performed (sensors used, type of data recorded, description of the protocol, duration of the experiment, possible income for the participant, etc.);
- the possible risks incurred (see Appendix A and next paragraph);
- the privacy issue, data being confidential and anonymous;
- the persons to contact in case of questions or problems (project leader and experimenters);
- the possibility to stop the experiment at any time, and to have data deleted on request, both without any loss of benefits.

Neither questions considered as too personal nor medical questions were asked. The only incentives to participate in the studies were self-motivation and curiosity. In only one study a (small) financial reward was given to some participants according to their performance (to increase their motivation see Chapter 7). Two copies of the (paper only) consent form were signed: one for the person participating to the experiment, one for the project leader. Moreover the content of this consent form was inspired from several sources like the binding regulations referred above and other consent forms used in medical and biological research.

Regarding **safety**, all recordings of physiological signals were **non-invasive**. The system used in the studies (see Section 3.2) records physiological signals using active electrodes, which mean that a small quantity of current is applied on the surface of the skin. To our knowledge this current is not harmful to the human. No counter-indications are given by the manufacturer and supplier of the acquisition system. The electrodes and the A/D converter are galvanicaly isolated (using an optical fiber) from the rest of the acquisition system (PC), avoiding any risk of electrocution from the mains supply. These systems are routinely and safely used in many laboratories worldwide, for a number of applications including human-computer interaction. Other threats could be headache due to a too long period wearing the fairly tight EEG caps and

epileptic reactions due to the manipulation of human computer interfaces (especially games). It is important to state that none of the experimenters have medical knowledge that could permit the incidental finding of diseases or abnormalities in the data; this was explicitly stated in the consent form.

Regarding **privacy**, all recordings were anonymized by assigning a numerical code to each user, and stored accordingly (e.g. Participant 1, Participant 2, etc.). The names of the participants were not stored in electronic form. Only the project leader and the experimenter had access to the participant's identity. The relevant recorded data, according to the purpose of the experiment, are kept for processing. Only authorized researchers have access to the recorded anonymous data (for electronically stored data: password and IP address restriction). Data once recorded will not be modified. The data is processed only for the research purposes as is explained in the consent form and will not be used for any other research. All data can be erased on request from the participant. Further, it is not possible to determine the identity of a participant from the physiological recordings we employ. The data is not shared with outside laboratories, and not used for any commercial purpose.

Finally, this section does not aim at answering ethical questions concerning the potential applications of this thesis. However, we believe that for each specific application the benefit / burden balance should be considered carefully.

Chapter 4 Methods for emotion assessment

4.1 Classification

Section 3.3 describes the different features extracted from the physiological signals to reflect emotional activity. The next step to go toward emotion assessment consists in finding a computational model that is able to associate a given instance of those features to an emotional state.

To obtain such a model it is possible to learn it from previously acquired data. In this thesis several protocols (described in the next chapters) were designed to acquire emotional data elicited by different types of stimuli (images, recall of past emotional episodes, etc.). As explained in Chapter 3, the signals acquired from those protocols were segmented in N trials, each one corresponding to an emotional state y_i, and features were extracted for each trial. The result is a database for each protocol including:

- a feature set **F** composed of N feature vectors $\mathbf{f}_i$, one for each trial i, containing some of the features described in Section 3.3 (a feature vector $\mathbf{f}_i$ can then be regarded as an instance or sample of the feature set **F)**;
- a vector **y** with each value y_i being the emotional state associated to trial (or sample) i.

On such a database, supervised learning algorithms [119, 120] can be applied to obtain a computational model that transduces the values of feature vector $\mathbf{f}_i$ into an estimated emotional state $\hat{y}_i$. In supervised learning such a model is learned from a database that contains both a feature set **F** and the associated labels vector **y**. However, in order to construct the **y** vector it is first necessary to define what an emotional state is and how to determine the real emotion associated to a trial. This is known as the ground-truth construction and is detailed in Section 4.1.1. Section 4.1.2 explains the methods employed to validate the models learned from the data and Section 4.1.3 presents the different supervised algorithms used to learn the model.

4.1.1 Ground-truth definition

Defining a ground-truth for the purpose of emotion assessment strongly depends on the protocol used to record emotional reactions. There are actually two ways to elicit emotions (Section 2.3):

- by asking the participant to self-generate a given emotion;
- by stimulating him / her with material containing emotional content (images, sounds, video clips, etc.).

In the first case the annotation is straightforward since a trial could be annotated with the emotion that the participant was supposed to express. However, assumes that the requested emotions were successfully elicited, and thus requires appropriate control of the elicitation procedure.

In the second case, the annotation of each trial (corresponding to a stimulus) can be done by either one of the following three methods:

- by a-priori defining the emotional label of each stimulus;
- by determining the elicited emotion from the observation of the participant's emotional expressions (for instance facial expressions);
- by asking the participant to self-assess his / her feelings after the stimulation.

The a-priori annotation can be done arbitrarily for instance according to the judgment of the experimenter. However this method is not recommended since it does not take into account the variability of judgments that can be observed in a population. Another possibility is to ask a large population of persons to extensively evaluate the stimuli. Each stimulus can then be associated to the most frequent label or to the average of judgments if emotions were evaluated in a continuous space such as the valence-arousal space. For instance, each of the IAPS images [54] is provided with the mean and the standard deviation of the valence-arousal judgments of 800 persons. However this method is still not optimal since the participants can have various emotional reactions under the same stimulus, due for instance to differences in past experience. As an example, an image of someone skiing can elicit pleasure but it can also elicit negative emotions for a participant that had a bad experience on skis. To alleviate this problem, it is possible to ask to each participant to select and evaluate images before the experiment. By using this method the elicited emotions are known in advance. The problem is then that the stimuli will be known by the participants and thus elicit different emotional response during the experiment (for instance less intense responses).

Analyzing the emotional behavior and expressions of the participant is a possible method to determine the elicited emotion. However this method requires that at least one expert (a psychologist for instance) is present during the experiment. In order to avoid bias in the estimated emotions, it is even better to group the judgments of many experts, raising the question of the combination of different judgments. The participant should be free to fully express his / her emotions, which is not always the case. For instance, when physiological signals are recorded movements are generally limited to avoid noise in the signals. Finally, if a model is trained from this type of ground-truth then it will be able to detect the expression of the emotion and not the other factors involved in emotions such as the subjective feelings (see Section 2.1.3.b).

Presenting the stimuli to the participants and then asking them to self-assess their feeling is also an alternative to the two precedent annotation methods. This can be done by asking the participants to fill in questionnaires, to give emotional labels and to evaluate their emotions in the valence-arousal space. While this method alleviates the problems related to the subjectivity of emotions it relies on the assumption that the participants are reliable experts in evaluating their own feelings. This latter statement can be discussed because of the following issues. Firstly, the participant will determine the elicited emotions mostly based on his / her subjective feeling which is only one of the factors involved in the emotion (see Section 2.1.3.b). Secondly, misunderstanding of the material used for self-assessment could lead to wrong annotations. As stated in Section 2.1.4, emotional words can have different meanings across persons and cultures. Moreover, representing an emotion in the valence-arousal space is not straightforward and it is mandatory to provide explanations before someone can use this space. Thirdly, a participant can hide his / her true felt emotional state because of social rules. For instance, a man can hesitate to report high arousal while watching a picture of a nude man.

The best way to annotate data is certainly to combine the different annotation methods described above. An example can be to combine annotations from experts that evaluate an emotional state based on the analysis of several emotional cues (facial expressions, speech, behavior, physiological signals etc.) together with self-reported measures of emotions. This would enable to determine a ground-truth based on several components involved in emotional processes. However it is not clear how the fusion of the different annotation should be performed.

In this study both self-generated and stimuli-based emotions were elicited in different protocols. For the protocol where emotions were self-generated, the trials were directly annotated with the corresponding emotional state as explained above. For the protocols where emotions were elicited using stimuli two of the annotation methods described above were employed: a-priori annotation and self-assessment. The effectiveness of those methods was thus estimated based on the accuracy of the emotion assessment.

The above paragraphs detail how to collect a ground-truth without any assumption on the type of annotations that are collected. Those annotations could be continuous (for instance by using the valence-arousal space) or discrete (using emotional labels such as fear and anger). Essentially because of its generality but also for other reasons detailed in Section 2.1.4 the valence-arousal space was chosen as a representation for emotional states. Accurately determining a point in this space based only on physiological features is a difficult task. For this reason we preferred to take a first step by defining valence-arousal classes of interest. Thus each value y_i of the $\mathbf{y}$ vector can take values in a set of emotional labels $Y=\{\omega_1,\ldots,\omega_c\}$, where C is the number of classes.

Depending on the protocol, the valence-arousal space was thus divided in different regions and each region associated to a target label. For instance, three regions of interest (C=3) can be defined by segmenting the valence-arousal space in calm, excited-positive and excited negative areas. Another possibility is to segment the space in two areas (C=2) such as calm vs. excited or positive vs. negative areas. The segmentation can be done a-priori, the participants thus annotate the emotions accordingly to the classes defined; or a-posteriori from continuous annotations gathered during the protocol. All those different possibilities give rise to different **y** ground-truth vectors and thus corresponds to different formulations of the emotion assessment problem that are called classification schemes. Several classification schemes were studied for the data gathered from each protocol and they will be detailed in the appropriate chapters.

4.1.2 Validation strategies

From the ground-truth acquired according to the methods presented in Section 4.1.1, the emotion assessment task is defined as supervised classification. It is supervised because a ground-truth is available to learn a model (the y_i values) and it is classification because the goal is to retrieve emotional classes of interest $\hat{y}_i$. In this case, the methods usable for emotion assessment originate from the pattern recognition and machine learning fields [119, 120].

When training classifiers overfitting can occur when the obtained model perfectly fits the data from which it is learned but performs poorly on new unseen data [119]. In order to control for the generalization capability of the model, it is thus important to test the performance of a learned model (the classifier) on a different dataset than the one used for learning. Validation strategies consist in segmenting the data in two sets: a training set from which the model is learned and a test set on which the performance of the model is tested (Figure 4.1).

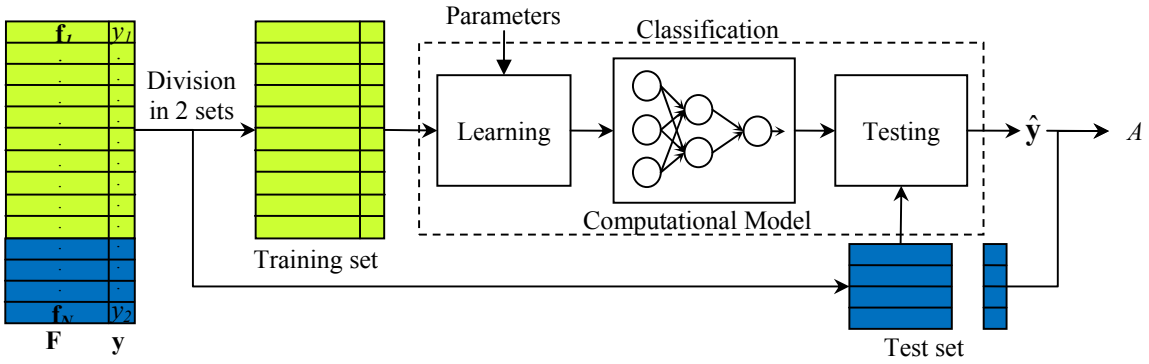


Figure 4.1. Validation scheme for classification, where $\hat{\mathbf{y}}$ is the vector of the classes estimated by model for the test set, A is the accuracy.

In this study, the performance of a model was tested by using the following measure of accuracy:

$$A = \frac{N_c}{N_t} \tag{4.1}$$

where N_t is the number of samples in the test set and N_c is the number of test samples correctly classified (test samples where $\hat{y}_i = y_i$). A confusion matrix (Table 4.1) will also be used to determine how the samples are classified in the different classes. A confusion matrix gives the percentage of samples belonging to class ω_i and classified as class ω_j.

		Estimated labels $\hat{y}$				
		ω_1	.	ω_j	.	ω_c
Ground truth labels y	ω_1	$P_{1,1}$				$P_{1,c}$
	.					
	ω_i			$P_{i,j}$		
	.					
	ω_c	$P_{c,1}$				$P_{c,c}$

Table 4.1. A confusion matrix, $P_{i,j}$ is the percentage of samples belonging to class ω_i and classified as class ω_j.

The accuracy A can be retrieved from the confusion matrix by summing its diagonal elements $P_{i,i}$ weighted by the prior probability $p(\omega_i)$ of occurrence of the class ω_i:

$$A = \sum_{i=1}^{c} p(\omega_i) P_{i,i} \tag{4.2}$$

For the model to correctly represent the data, it is important that the training set contains enough samples (or instances). On the other hand it also important that the test set contains enough samples to avoid a noisy estimate of the model performance. This can be problematic because it often occurs that the amount of collected data is limited in practice. This is particularly true in our case since the number of emotional stimulations is limited by the duration of the protocols which should not be too long to avoid participant fatigue as well as elicitation of undesired emotions. Cross-validation methods help to solve this problem by splitting the data in different training / test sets so that each sample will be used at least once for training and once for testing.

The two well known cross-validation methods are the k-fold and the leave-one-out [136]. In the k-fold cross-validation, the data is split in k folds containing the same amount of samples. Generally the folds are determined so that the prior probability $p(\omega_i)$ of observing each class i is the same for each fold. Each fold is then used in turn as the test set and a model is learned from the remaining k-1 folds. By using this method k accuracies are obtained from the k test sets so that it is possible to compute the average accuracy and its standard deviation. The leave-one-out cross-validation is similar to the k-fold cross-validation except that the size of the test set is always 1. Thus N models are tested in turn on each sample and learned from the N-1 remaining samples of the database. The advantage of this cross-validation method is that it provides the

maximum possible size for the training set which generally helps to find a better model, especially in the case where few samples are available in the database. On the other hand only the average accuracy can be computed reliably since the test set contains only one sample (the accuracy is thus either 0 or 1).

When designing a general computational model for emotion assessment (i.e. a model that is not person dependent but can be used to assess emotions of anyone) based on physiological features it is important to take into account the high variability that can be observed in physiological reactions. To control the performance of such a model it should be tested on physiological data of persons whose features were not used for the learning of the model. For this reason the participant cross-validation method was proposed. The database was segmented in folds where each fold contains the samples computed from the physiological signals of a single participant. Then the classification performance was computed similarly to the k-fold cross-validation, by using each fold in its turn as the test set. This method allows testing the classification performance as in a "real-case" where the emotions of a user would be assessed by using a model defined from the physiological activity of other persons.

4.1.3 Classifiers

Section 4.1.2 detailed how to determine the performance of a classifier. This section will describe the different classifiers used in this study, all being part of the pattern recognition and the machine learning fields [119, 120]. For most classification algorithms, it is important that the features be normalized (i.e. belongs to the same range of value). This normalization was applied at each cross-validation step by whithening each feature using mean and standard deviation computed from the training set.

a. Naïve Bayes

Several classifiers rely on the Bayes' rule to find the most probable class in which a sample represented by its feature vector **f** should be classified. This is done by attributing the class ω_i that maximize the posterior probability $p(\omega_i \mid \mathbf{f})$ to the estimated label $\hat{y}$. According to the Bayes' rule:

$$p(\omega_i \mid \mathbf{f}) = \frac{p(\mathbf{f} \mid \omega_i)\, p(\omega_i)}{\sum_{i=1}^{C} p(\mathbf{f} \mid \omega_i)\, p(\omega_i)} \tag{4.3}$$

One of the advantages of this type of classifier is that it is able to output the posterior probability $p(\omega_i \mid \mathbf{f})$ that a sample belong to a given class ω_i. Notice that it is enough to find the maximum value of the numerator to maximize $p(\omega_i \mid \mathbf{f})$ since the denominator has the same value for any

class ω_i. The main differences between Bayesian classifiers is the way by which the conditional probabilities $p(\mathbf{f} \mid \omega_i)$ are estimated.

For the Naïve-Bayes classifier the assumption of conditional independence of the features given ω_i is made:

$$p(\mathbf{f} \mid \omega_i) = \prod_{j=1}^{F} p(f_j \mid \omega_i) \tag{4.4}$$

where F is the number of features in the feature vector $\mathbf{f}$. In this study, the conditional probability $p(f_j \mid \omega_i)$ of a feature j was estimated by quantizing the features in 10 bins of equal sizes and computing the associated conditional probability mass function from the training set. The prior probability $p(\omega_i)$ could also be computed from the training set; it would however be biased by the stimuli presented for each class in the protocol. For instance, if the aim of the classifier is to distinguish between calm and excited emotional states and more excited stimulus were presented to the participants then the prior probability would be higher for the excited class. While this is coherent in this particular protocol it does not have any meaning in real applications since there is nothing that guarantees the higher occurrence of excited states in this case. For this reason the prior probability $p(\omega_i)$ was set to $1/C$ under the assumption of equiprobability of the classes.

b. Discriminant analysis

Two discriminant analysis methods, namely the linear discriminant analysis (LDA) and the Quadratic discriminant analysis (QDA) are used in this study. Both are based on the Bayes rule to find the class with the highest posterior probability $p(\omega_i \mid \mathbf{f})$ [119]. For this purpose the following g_i discriminant functions are defined:

$$g_i(\mathbf{f}) = \ln(p(\mathbf{f} \mid \omega_i)p(\omega_i)) \tag{4.5}$$

Finding the class ω_i with the highest g_i value is then similar to finding the class that maximizes the numerator of equation 4.3 Under the assumption that the conditional distributions $p(\mathbf{f} \mid \omega_i)$ are Gaussians with different means $\boldsymbol{\mu}_i$ and covariance matrices $\boldsymbol{\Sigma}_i$ this rule automatically defines a (hyper-)quadratic decision boundary (hence the name QDA for the associated classifier):

$$g_i(\mathbf{f}) = -\frac{1}{2}(\mathbf{f}-\boldsymbol{\mu}_i)\boldsymbol{\Sigma}_i^{-1}(\mathbf{f}-\boldsymbol{\mu}_i)^T - \frac{F}{2}\ln 2\pi - \frac{1}{2}\ln\left|\boldsymbol{\Sigma}_i\right| + \ln p(\omega_i) \tag{4.6}$$

where T and $|.|$ respectively stands for the transpose and determinant operators. Vectors $\boldsymbol{\mu}_i$ and matrices $\boldsymbol{\Sigma}_i$ are computed from the training set. In the case where $\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}_j, \forall i \neq j$ the boundary becomes linear, yielding an LDA classifier. With the LDA it is sufficient to compute a single

covariance matrix $\mathbf{\Sigma}$ from the complete training set without distinction between classes. Similarly to the Naïve Bayes classifier, the prior probability $p(\omega_i)$ was defined as $1/C$.

In the case where the size of the feature space F is large and the number of samples available for learning is small, the discriminant analysis can fall in the singularity problem where the $\mathbf{\Sigma}_i^{-1}$ matrix is not invertible. In this case we used the diagonalized version where covariance matrices are assumed to be diagonal, containing the variances of the features. Notice that in this case the discriminant analysis is a Naïve Bayes classifier with a conditional independent Gaussian assumption for the $p(f_j \mid \omega_i)$ distributions. The Matlab statistics toolbox (v. 5.0.1) implementation of those algorithms was used in this study.

c. Support Vector Machines (SVM's)

A SVM [120, 137] is a two class classifier (C=2) using a linear model of the form:

$$h(\mathbf{f}) = \mathbf{w}\phi(\mathbf{f})^T + b \tag{4.7}$$

where a feature vector $\mathbf{f}$ is estimated as being from class ω_1 if $h(\mathbf{f})<0$ and ω_2 if $h(\mathbf{f})>0$. The ϕ function projects a feature vector in another feature space, generally of higher dimensionality, thus allowing for non linear separation of the data in the original feature space. In order to find the model weights $\mathbf{w}$ and b, an SVM tries to maximize the distance between the decision surface created by the h function and a margin to this surface as well as to minimize the error on the training set. The trade-off between margin maximization and the training error minimization is controlled by a parameter C_{SVM} that was empirically set to 1 in this study. The advantage of SVM's is that they minimize an upper bound on the expected risk rather than only the error on the training data, thus enabling good generalization even for undersampled datasets, as well as interesting performances in high dimensional feature spaces [138]. Moreover, they provide sparse solutions where not all of the data points are used for classification.

The SVM optimization problem can be expressed in a dual form [120, 137], where a kernel function $k(\mathbf{f},\mathbf{f}') = \phi(\mathbf{f})\phi(\mathbf{f}')^T$ is introduced between two samples $\mathbf{f}$ and $\mathbf{f}'$. In this new formulation, the decision boundary becomes a function of only some of the data points called the support vectors. In this study, both linear and radial basis function (RBF) kernels were used:

$$k^{linear}(\mathbf{f},\mathbf{f}') = \mathbf{f}.\mathbf{f}'^T \tag{4.8}$$

$$k^{RBF}(\mathbf{f},\mathbf{f}') = e^{-\gamma\|\mathbf{f}-\mathbf{f}'\|^2} \tag{4.9}$$

where ||.|| is the norm operator. In the case of RBF kernels, the size of the kernel γ was chosen by applying a 5-fold cross-validation procedure on the training set and finding the γ yielding the best accuracy. The tested γ values belonged to the 5.10^{-3} to 5.10^{-1} range with a step of 5.10^{-3}.

There are two drawbacks to the use of SVM's as classifiers: they are intrinsically only two-class classifiers and their output is uncalibrated so that it is not directly usable as a confidence value in the case one wants to combine outputs of different classifiers or modalities. In this study the first point was addressed by using the one-versus-one approach where *C(C-1)/2* classifiers are trained on each possible pair of classes. The class associated to a test sample is the one that receives the highest number of votes from the *C(C-1)/2* classifiers.

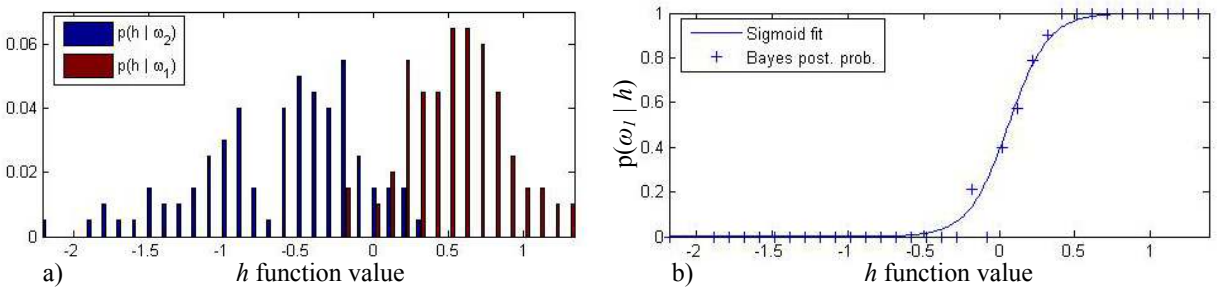


Figure 4.2. Obtaining posterior probabilities p(ω_i | *h*) from SVM outputs. a) Histograms representing the distributions of the SVM output for two classes. b) Posterior probabilities estimates from the Bayes rules applied on the histogram of a) and from the sigmoid fit proposed by Platt [139].

For the second point, Platt (2000) proposed to model the probability p(ω_1 | *h*) of being in the first of the two classes knowing the output value *h* of the SVM. As can be seen in Figure 4.2 this could be done by applying the Bayes rule. The discrete posterior probability plotted in Figure 4.2 approximately follows a sigmoid curve; this is why Platt proposed to model those probabilities by using:

$$p(\omega_1 \mid h) = \frac{1}{1+\exp(\alpha h + \beta)} \tag{4.10}$$

where the α and β values are found by the algorithm proposed in [139] and improved in [140]. Figure 4.2 shows the result of the sigmoid curve fitting. Concretely, the *h* values were obtained from a 5-fold cross-validation on the training set and the parameters α and β were determined from those *h* values. The posterior probabilities of the test samples were then computed using equation 4.10. Finally, to compute the posterior probabilities p(ω_i | *h*) when there are more than two classes to separate, the solution proposed in [141] was employed. The libSVM [142] Matlab toolbox was used as an implementation of the SVM and probabilistic SVM algorithms.

d. Relevance Vector Machines (RVM's)

RVM's [143] are algorithms that have the same functional form as SVM's but embedded in a Bayesian learning framework. They have been shown to provide results similar to SVM's with generally sparser solutions. They have the advantage that they directly give an estimation of the posterior probability of having class ω_i.

RVM's try to maximize the likelihood function of the training set using a linear model including kernels. The main difference with classical probabilistic discriminative models is that a different prior is applied on each weight thus leading to sparse solutions that should generalize well. For all the following studies, the multiclass RVM version presented in [144] was used.

4.2 Feature selection

The features extracted from the physiological signals, especially EEG signals, were sometimes of high dimensionality. For instance the $\mathbf{f}^{EEG\_STFT}$ feature vector presented in section 3.4.2.a would contain 16704 features for an EEG recorded over a period of 7.5 seconds with 64 electrodes. Although increasing the number of features in a database should theoretically decrease classification error, this is not the case in practice because models of classifiers, or parameters of distributions, are estimated from a training set of limited size. Having fewer samples than features for classification, as could be the case when the feature space is of high dimensionality, is known as the undersampled or singularity problem [138, 145]. To alleviate this problem, one can resort to feature space reduction techniques. Reducing the size of a feature space can be done either by finding a function that projects the data points in a space of lower dimensionality (generally preserving or improving the discriminability between classes) [119, 146] or by discarding the features that are not of interest for classification [147-149]. In this study algorithms from the second approach, called feature selection algorithms, were implemented.

Two different approaches are generally considered to deal with feature selection [147]: the filter approach where the quality of each feature is estimated independently of the classification scheme, generally using statistical measures; and the wrapper approach that relies on the classification algorithm to evaluate feature subsets as well as on heuristic methods to find the best subset. Although the second approach can yield better results, it suffers from an important computational cost, especially for very high dimensional space. There are also classification algorithms that perform embedded feature selection such as the Adaboost [150]. The SVM also contains a regularization term in its error function that keeps the weights $\mathbf{w}$ low and thus provides a form of feature selection.

In this study, feature selection algorithms were always applied on the training set to determine the selected subset of features. The classifiers were then trained on this selected subset and the

resulting model was applied on the test features subset. This section describes the filter and wrapper algorithms used for feature selection.

4.2.1 Filter methods

a. ANOVA

The ANOVA (ANalysis Of VAriance) is a statistical test that estimates if a difference in mean observed between several groups for a given variable is significant. Two values can be computed from an ANOVA test: the F value that gives the magnitude of the mean difference and the p value that gives the probability of making an error by assuming a difference in mean. If the p value is low (near to 0) than this reflects the strong belief that at least one group mean is different from the others.

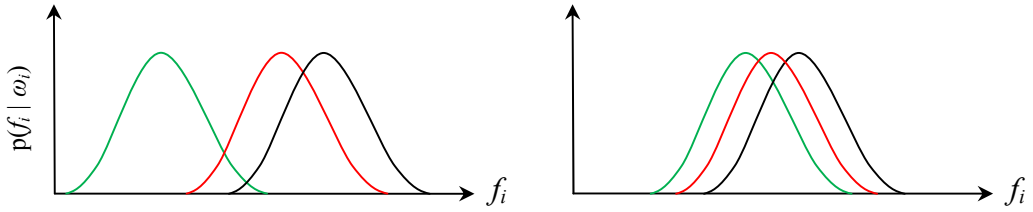


Figure 4.3. Different possible distributions of a feature value for a 3 classes scenario (green, red and black classes). (left) The feature is relevant since it is usefull to distinguish the green class from the others (low p value). (right) A non relevant feature (high p value).

Under the assumption that the conditional distributions $p(f_i \mid \omega_i)$ are Gaussians with similar variances than a feature is relevant for classification if the means of those distributions are different for at least one of the classes (Figure 4.3). Thus an ANOVA test was applied on each feature with the groups defined by the classes and the resulting p values were used to determine if a feature was relevant. A feature was considered to be relevant if and only if $p < \delta_{\text{ANOVA}}$, with δ_{ANOVA} set to 0.1 which corresponds to an error probability of 10%. All non-relevant features were discarded. It was thus not possible to a-priori choose an exact number of features to be selected.

b. Fisher criterion

The Fisher projection [119, 146] is a well known algorithm that projects the samples of a given feature space into a subspace of lower dimensionality having the highest linear discrimination between the classes according to the Fisher criterion. However, the choice of the dimension F_s of the subspace is limited by the number of classes ($F_s < C$). In order to perform feature selection instead of projection and to be free to choose the size of the resulting feature space, the Fisher criterion used in the Fisher projection algorithm was applied independently to each feature. For

this purpose, vectors $\tilde{\mathbf{f}}_i$ containing all the samples for each feature f_i were constructed from the feature set **F (**a vector $\tilde{\mathbf{f}}_i$ is a column of **F)**. It is then possible to apply the following Fisher criterion on such a vector:

$$J(\tilde{\mathbf{f}}) = \frac{\sum_{j=1}^{C} N_j (m_j - m)^2}{\sum_{j=1}^{C} \sum_{\tilde{f} \in D_j} (\tilde{f} - m_j)^2} \tag{4.11}$$

where m and m_j are respectively the mean of the vector $\tilde{\mathbf{f}}$ and the mean of this vector for samples belonging to class j, N_j is the number of samples belonging to class j and D_j represents the ensemble of values from $\tilde{\mathbf{f}}$ that belongs to class j.

From the numerator of Equation 4.11 it can be seen that the larger the distance between the feature means of each classes, the higher the value of J. On the other hand, the denominator represents the average variability of the feature across the classes and the smaller the better. Thus this criterion will have high values for features that have different mean for each class and a low average intra-class variance which is very similar to what the ANOVA test does.

To filter out features, the Fisher criterion J was computed for each feature and the features were then ranked by decreasing order of J. Finally, the F_s features with the highest J value were kept and the others removed. This algorithm thus allows selecting the size F_s of the resulting feature space.

c. Fast Correlation Based Filter (FCBF)

The FCBF algorithm [149] was proposed as a filter method that selects features according to their relevance to the class concept and removes those that are redundant. By removing redundant features the algorithm reduces the size of the selected subset of features without changing its disciminant capacity. The absolute value of the linear correlation measure was used to evaluate both the relevance and redundancy of a feature:

$$Corr(\mathbf{x}, \mathbf{z}) = \left| \frac{\sum_{i=1}^{m} (x_i - \overline{x})(z_i - \overline{z})}{\sum_{i=1}^{m} \sqrt{(x_i - \overline{x})^2} \sum_{j=1}^{m} \sqrt{(z_i - \overline{z})^2}} \right| \tag{4.12}$$

where **x**, **z** are two vectors of size m respectively containing x_i and z_i values with means $\overline{x}$ and $\overline{z}$.

From the feature set **F**, vectors $\tilde{\mathbf{f}}_i$ were constructed. These vectors contains all the samples of each feature f_i. The algorithm then selected a feature subset in two steps:

- removal of each irrelevant feature f_i where $Corr(\tilde{\mathbf{f}}_i, \mathbf{y}) < \delta_{FCBF}$;
- removal of redundant features; a feature f_i being considered as redundant with respect to another feature f_j if $Corr(\tilde{\mathbf{f}}_i, \tilde{\mathbf{f}}_j) > Corr(\tilde{\mathbf{f}}_j, \mathbf{y})$ and $Corr(\tilde{\mathbf{f}}_i, \mathbf{y}) < Corr(\tilde{\mathbf{f}}_j, \mathbf{y})$.

The first step removes features that have low correlations with the classes. In the second step the pairs of features that are highly correlated are identified and for each feature pair the feature that has the lowest correlation with the class is removed.

As for the ANOVA feature selection, the number of selected features is determined by the δ_{FCBF} parameter but the exact number of features can not be chosen explicitly. The possible values for the δ_{FCBF} parameter range from 0 to 1. With $\delta_{FCBF} = 0$ all the features are selected in the first step but redundant features are removed in the second. When δ_{FCBF} is increased the number of selected features decreases with the minimum number of selected features being reached when $\delta_{FCBF} = 1$ (only the features that are fully correlated with the classes are kept).

4.2.2 The SFFS wrapper method

In most of the filter algorithms, the relevance of a feature is estimated independently from the other features. However it can happen that a feature is not relevant by its own but is highly relevant when combined with another feature; this is called feature interaction [151, 152]. Wrapper algorithms have the advantage that a feature is added or removed from a feature set based on the analysis of the newly generated feature set. As a consequence, those algorithms are able to remove redundant features but also to take into account the interaction between features. Moreover, the wrapper algorithms generally use the predictive accuracy of the classifier used for final classification to evaluate the performance of a feature subset. Thus the features are selected in accordance with the intrinsic properties of the classifier. However the main drawback of those algorithms is that they are computationally intensive since they require the evaluation of a number of feature subsets that is generally higher than the number of features.

The Sequential Floating Forward Selection (SFFS) [148] was implemented as a wrapper algorithm if the number of features was sufficiently low. This algorithm starts with an empty feature set. At each iteration of the algorithm a forward step is taken followed by a number of backward steps. A forward step consists in adding to the current feature set the feature that maximizes the performance of the newly created feature set. A backward step consists in removing a feature from the current feature set only if doing so improves the performance. This last step is taken as long as removing a feature improves the performance. This method searches

the feature space more deeply than by using only forward or backward algorithms [148]. The SFFS algorithm stops when the number of features in the subset is equal to the number of requested features F_{SFFS} and no backward step is taken. The best subset of selected features is then returned which can be of size 1 to F_{SFFS}. In this study the classification accuracy, computed on the samples of the training set, was used as the performance measure. The classifier used to evaluate a feature subset was the same as the one used to generate the final classification model. For instance, if the SVM classifier was used for emotion assessment then it was also used for feature selection.

4.3 Fusion

As stated in Sections 1.1.2 and 2.1.3.b emotion elicitation is a multimodal process that involves several components of the organism. As a consequence, emotions are expressed through several channels, giving rise to many emotional cues that can be recorded by different sensors. Since the information recorded by those sensors can represent the activity of the different components involved in emotional processes, combining the information obtained from those sensors can improve the reliability of emotion assessment. The combination of multi-sensor information is known as fusion. In this study, several sensors described in Section 3.1 were used to monitor the activity of both the peripheral and the central nervous system. Moreover, different types of features were also computed for each sensor. This section describes how the different information obtained from the sensors were fused for the purpose of emotion classification.

When the goal is to perform classification, information fusion can be done at different levels including the sensor data, the feature and the classifier levels [153, 154]. In this study the performance of fusion was evaluated at the feature and classifier level. In those cases, the problem could be formulated as having a dataset containing several feature sets $\mathbf{F}^j$ (instead of one as presented in Section 4.1) and the associated label vector $\mathbf{y}$. Each of those feature sets can contain:

- the features extracted from the signals produced by a given sensor (for instance all the features extracted from the GSR signals or all the features extracted from the EEG signals);
- the features extracted from the signals corresponding to one of the two parts of the nervous system (for instance the features extracted from the PNS or those extracted from the CNS);
- the different features extracted from the signals of a unique sensor (for instance a feature set containing the MI features extracted from EEG signals and another one containing the STFT features extracted from the same signals).

Under this formulation, fusion at the feature levels consists in combining the features sets $\mathbf{F}^j$ before classification, while fusion at the classifier level consists in classifying each feature set independently and combine the classification results in a second step.

4.3.1 Feature level

Fusion at the feature level can be done by concatenating the feature sets or by linearly or non-linearly combining the features [153]. Concatenation of features from the peripheral and central nervous system has already been done in [113] with no improvement of the performance. However the features computed from the EEG signals were only standard features (see section 3.3.1) not specially related to emotional processes which could explain their low performance for emotion assessment.

In this study concatenation of the features was done to combine the features extracted from the different peripheral sensors (GSR, Respiration belt, etc.) in a unique feature set. It was also used to fuse peripheral features with features computed from the EEG signals. Concatenating N_F feature sets $\mathbf{F}^1$, …, $\mathbf{F}^{N_F}$ in a new feature set $\mathbf{F}$ consists in the concatenation of feature vectors $\mathbf{f}_i^j$ for each sample i and all feature set j:

$$\mathbf{f}_i = [\mathbf{f}_i^1 \cdots \mathbf{f}_i^j \cdots \mathbf{f}_i^{N_F}] \tag{4.13}$$

4.3.2 Classifier level

According to Sanderson et al. [153] the fusion at the classifier level (or post-mapping fusion) is divided in two categories: decision fusion and opinion fusion. In decision fusion, the estimated classes given by several classifiers are combined to decide which class will be attributed to a sample. The most common method for decision fusion is certainly majority voting: the class that has been chosen by most of the classifiers is attributed to the sample. However, decision fusion strategies are generally applicable only in particular cases (for instance the number of classifier is constrained by the number of classes for majority voting). This is why the present study focuses on opinion fusion.

a. Opinion fusion: sum rule

In opinion fusion, a classifier outputs a score for each class. This score generally represents the classifier confidence that the test sample belongs to a given class. Product and sum rules can then be applied to fuse the scores given by multiple classifiers [153]. In this work, the probabilistic outputs of the classifiers were used as a measure of confidence. The following sum rule was applied to fuse those output probabilities for a class i:

$$k_i = \frac{\sum_{q \in Q} P_q(\omega_i \mid \mathbf{f})}{\sum_{j=1}^{C} \sum_{q \in Q} P_q(\omega_j \mid \mathbf{f})} = \sum_{q \in Q} \frac{1}{|Q|} P_q(\omega_i \mid \mathbf{f}) \tag{4.14}$$

where Q is the ensemble of the classifiers chosen for fusion, $|Q|$ the number of such classifiers and $P_q(\omega_i \mid \mathbf{f})$ is the posterior probability of having class ω_i according to classifier q. The final choice is done by selecting the class ω_i with the highest value k_i. It can be observed that k_i can also be viewed as a confidence measure on the class given by the fusion of classifiers.

b. Opinion fusion: Bayes belief integration

One drawback of the sum rule is that it does not take into account the errors made by each of the classifiers [154]. As an extreme example, it is possible that a (very bad) classifier q classifies most of the instances of a class ω_i into a class ω_j. Thus, when the class estimate $\hat{y}_q$ of this classifier is ω_i there is a high probability that the true class be in fact ω_j. For Bayes belief integration [154], the errors produced by the classifiers are expressed by the probabilities $P(\omega_i \mid \hat{y}_q)$ computed from the confusion matrices obtained from the training set. The fusion is then performed by assuming classifiers independency and choosing the class ω_i that maximizes the following probability:

$$P(\omega_i \mid \hat{y}_1 ... \hat{y}_{|Q|}) = \frac{\prod_{q \in Q} P(\omega_i \mid \hat{y}_q)}{P(\omega_i)^{|Q|-1}} \tag{4.15}$$

where Q is the ensemble of classifiers used for the fusion and $\hat{y}_q$ is the class estimate of classifier q. Notice that the probability $P(\omega_i \mid \hat{y}_1 ... \hat{y}_{|Q|})$ could be computed directly from the training set without any assumption on classifiers independence. However it is generally difficult to estimate it reliably because of the low number of training samples compared to the high number of combinations of the $|Q|$ classifiers estimates.

4.4 Rejection of samples

In classification, the samples that lie close to the decision boundary can be considered as less reliably classified than those that are far from the decision boundary. For this reason a classifier can assign a class to a sample only if it is sufficiently far away from the decision boundary (i.e. the confidence in the classification is sufficiently high).

As a final step, rejection of trials that have a confidence value k_i (determined using the method presented in Section 4.3.2.a) below a threshold δ_{reject} was performed to improve classification

accuracy. When a sample is rejected because the confidence value is not sufficiently high, the sample is not classified and the classification accuracy is computed only on the samples with high confidence. The percentage of rejected samples as well as the accuracy computed on the remaining samples thus become a function of the δ_{reject} threshold. A good value for δ_{reject} would be the one which provides a compromise between accuracy maximization and rejection rate minimization. This approach is taken is chapter 6.

Chapter 5 Assessment of emotions elicited by visual stimuli

5.1 Introduction

Among all the senses, the visual modality is the one that has the largest capacity (i.e. it can carry a large amount of information in a given time) [155]. It is thus not surprising that, in HMI, most of the information is communicated to the user through the visual modality (for instance via screens). Consequently, recognizing visually elicited emotions is important to enhance affective computing methods. In this chapter, images were chosen as visual stimuli to elicit emotions. The performance of several methods to assess the valence and arousal dimensions of these emotions from physiological signals was then analyzed. Since this work uses images as stimuli, its results can also have some impact on the automatic affective annotation of multimedia data (see Section 1.2.3.c).

Section 5.2 describes how an emotional database of physiological features was constructed. In Section 5.3 different classes were defined in the valence-arousal space and the classification methodology employed to recognize them are presented. Section 5.4 discusses the results obtained and stresses the interest of EEG features alone as well as fused with other peripheral features in emotion assessment.

5.2 Data collection

This section details the creation of a database of physiological features where emotions, defined as classes in the valence-arousal space, where elicited by using images. It explains how the image stimuli where chosen, describes the protocol designed to acquire the data and finally presents the extracted features.

5.2.1 Visual stimuli

In this study we used a subset of images from the 700 emotionally evocative pictures of the IAPS (International Affective Picture System [54]). Each of these images has been extensively evaluated by north-American participants, providing valence and arousal values on nine points scales (ranging from 1 to 9). Experimentation showed a 0.8 correlation with evaluations performed by Europeans [156]. Each image is associated with average arousal μ_A and valence μ_V computed from these evaluations as well as with their standard deviations σ_A and σ_V. However, as observed during our experiments, feelings induced by an image on a particular participant can be very different from the ones expected. This is likely due to difference in past experience. Self-assessment of valence/arousal was therefore performed in the present study by each participant and for each image. Two different subsets of the IAPS images were presented to the participants, one to study the valence dimension (valence subset) and the other to study the arousal dimension

(arousal subset) of emotions. The complete list of the pictures presented to the participants can be found in Appendix C.

To construct the **valence subset**, 50 negative and 50 positive images were selected according to the following constraints:

$$\text{negative images: } \mu_A > 5; \quad \mu_V < 3; \quad \sigma_V < 2;$$

$$\text{positive images: } \mu_A > 5; \quad \mu_V > 7; \quad \sigma_V < 2.$$

The constraint on the mean arousal allows keeping only images that have high arousal and thus intense positive or negative content. The constraint on the mean valence is useful to distinguish negative from positive images while a low standard deviation ensures stability in the judgments of the participants (i.e. there is a high probability that a participant rates an image with a valence close to the mean IAPS valence). Pictures were than randomly selected from the subset of the IAPS images satisfying the constraints.

For the **arousal subset**, 50 images of high arousal and 50 images of low arousal were selected The pictures were taken from the 3 subsets of the IAPS images defined by the following constraints:

$$\text{low arousal and neutral: } \mu_A < 3; \quad 4 < \mu_V < 6;$$

$$\text{high arousal and positive: } \mu_A > 5.5; \quad \mu_V > 5;$$

$$\text{high arousal and negative: } \mu_A > 5.5; \quad \mu_V < 5.$$

50 images were chosen randomly from the “low arousal and neutral” set to construct the final low arousal set. To construct the high arousal set 25 images were chosen from both the “high arousal and positive” and the “high arousal and negative” sets. This procedure was employed because there are few pictures with high arousal and neutral valence. It is important that the high arousal set contains the same number of negative and positive images to ensure that the arousal axis will be assessed at the classification stage (and not the difference between neutral and negative images as could be the case if only negative image were present in this set). No constraint was added on the variance with the disadvantages that participants’ evaluations can strongly differ from the IAPS evaluations. However, if a constraint on the standard deviation would be added, the number of images would have been too low to construct sets of 50 pictures.

By using the constraints above, some of the high arousal images of the arousal subset can also be part of the valence subset. This should be avoided because the arousal and valence sets were

presented consecutively to the participants (see Section 5.2.2) and the second time an image is presented, the intensity of the felt emotion can be weakened. For this reason the images were chosen in order to minimize the intersection of the arousal and valence subsets. This result in a total of 19 images present in both sets (the list of these images is given in Appendix C).

5.2.2 Acquisition protocol

We acquired data from 4 participants, 3 males, 1 female, aged from 28 to 49. One of the participants is left handed. Cortical activity was acquired by recording an EEG with 64 electrodes (see Section 3.1.1). The peripheral sensors used were the GSR sensor, the plethysmograph to measure BVP, the respiration belt to evaluate abdominal and thoracic movements, and the temperature sensor. All signals were sampled at 1024 Hz.

For each experimental recording, the participant equipped with the above sensors was sitting in front of a computer screen in a bare room relatively immune to electromagnetic noise. The valence and the arousal sets of images were presented consecutively to the participant. For both sets, each trial corresponding to a visual stimulus was scheduled as described in Figure 5.1. A dark screen was first displayed for 3 seconds to "rest and prepare" the participant for the next image. A white cross was then drawn on the screen center for a random period of 2 to 4 seconds, to attract user's attention and avoid accustoming. An IAPS image was subsequently displayed for 6 seconds, while at the same time a trigger was sent for synchronization. The task of the participant was to watch each image and self assess the valence and the arousal of his / her emotion using a simplified version of the SAM [73] (see Figure 5.1). The valence/arousal scales were composed of 5 possible numerical judgments for each dimension with scales ranging from -2 to 2 for valence and 0 to 4 for arousal. This self-assessment step was not limited in time to allow for a resting period between images.

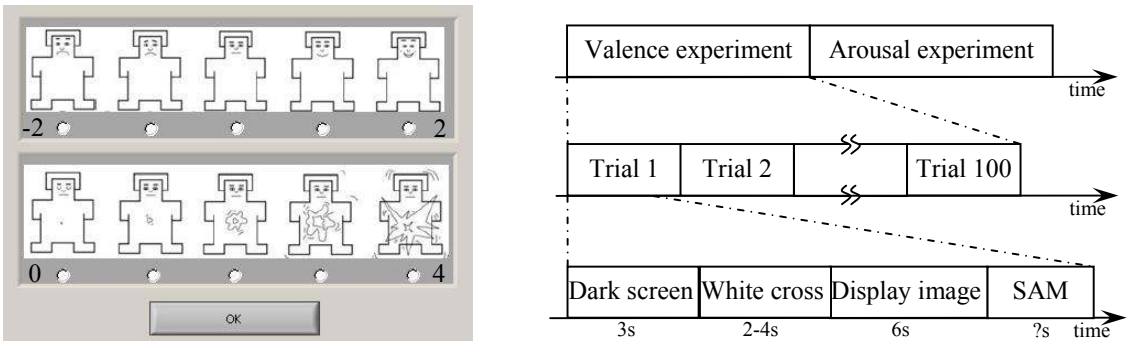


Figure 5.1. Description of the acquisition protocol. (left) the modified SAM used for self assessment. (right) the schedule of the protocol.

According to the 5 factors defined by Picard [106] this protocol corresponds to an *open-recording* condition since the participant knew that his / her physiological activity was recorded.

The emotions were *event-elicited* using the images. The emphasis was put on the *feeling* of emotions rather than the expression of the emotion since participants had to self-assess their feelings. The complete recording was performed in a *lab setting* (special room dedicated to the recording).

5.2.1 Features extracted

For the EEG signals two feature sets were used for classification (see Section 3.3.2.a):

- the *EEG_Lateral* feature set was employed for classification on the valence dimension;
- the *EEG_Area* feature set was used for the classification on the arousal dimension.

The choice of the *EEG_Lateral* feature set is motivated by the fact that studies have shown that the asymmetry index can be useful to discriminate positive from negative stimuli while Aftanas et al. [82] demonstrated that the *EEG_Area* features are significantly different for low and high arousal stimuli. Notice that most of the *EEG_Area* features concern the Occipital (O) lobe, which is interesting since this lobe corresponds to the visual cortex and subjects are stimulated with pictures.

Concerning peripheral signals, HR was estimated from the blood pressure signal by using the method presented in Section 3.2.4. The 5 peripheral signals to analyze are therefore: GSR, BP, HR, respiration and temperature. Table 5.1 presents the features extracted from each of these signals over the 6 seconds epoch. They correspond to some of the standard features presented in Section 3.3.1. A total of 18 features were thus obtained for the peripheral signals. Since no post-processing algorithm was applied to improve the peak detection on short time BVP signals (see Section 3.2.4) and since the minimum and maximum features are sensible to outliers, those features were not computed for the HR signal.

Signal	μ_x	σ_x	Min_x	Max_x
BVP	X	X	X	X
HR	X	X		
GSR	X	X	X	X
Respiration	X	X	X	X
Temperature	X	X	X	X

Table 5.1. The 18 features extracted from the peripheral signals.

For arousal classification, the EEG features and the peripheral features were concatenated as presented in section 4.3.1 to analyze the performance of fusion of peripheral and central information at the feature level. This step was not taken for valence classification as for this problem the classification accuracy obtained with peripheral features was at the random level (see Section 5.4.1).

5.3 Classification

5.3.1 Ground-truth definitions

The visual stimuli used in both the valence and arousal experiments were purposely chosen to belong to two distinct classes: negative vs. positive for the valence experiment and calm vs. excited for the arousal experiment. Since self-evaluations were also collected, the ground-truth can be defined either a-priori, based on the classes defined by the IAPS evaluations, or a-posteriori using the self-evaluations. Some of the advantages and disadvantages of these two methods are discussed in Section 4.1.1.

This section compares the IAPS evaluations to the self-assessment values and discusses the construction of the ground-truth for emotion assessment. For this purpose, the IAPS evaluations (ranging from 1 to 9) were linearly projected in the same range as the self-assessment values (ranging from -2 to 2 for valence and 0 to 4 for arousal) using the following formulas:

$$\begin{aligned} A &= \frac{(A_{IAPS} - 1)}{2} \\ V &= \frac{(V_{IAPS} - 5)}{2} \end{aligned} \tag{5.1}$$

V_{IAPS} and A_{IAPS} being the original IAPS valence / arousal values and V and A being the new valence / arousal values.

a. Valence experiment

As can be seen from Figure 5.2 the valence distribution of the self-evaluations is very close to the one obtained from the IAPS evaluations. This tends to validate that the visual stimuli elicited the expected emotions: either positive or negative emotions. However, 3 of the 4 participants judged some of the stimuli as being of neutral valence with 79% of those stimuli belonging to the positive class according to the IAPS evaluations. Moreover only one participant ranked three of the stimuli with a value of 2. Those results demonstrate that, according to the 4 participants self-assessments, positive emotions were more difficult to elicit than negative emotions.

Concerning arousal evaluations, the IAPS and self-assessment distributions were found to be quite different. For instance most of the stimuli were self-evaluated with an arousal value of 1 while, because of the constraints applied for the selection of images, none of the stimuli had an IAPS arousal value below 2. This result can be explained by two factors. First, the stimuli may have elicited lower arousal than expected. Secondly, the participants may have used the complete scale to report for the arousal difference between the stimuli (which is a measure relative to the set of stimuli) rather than to report absolute arousal value. When looking at the IAPS arousal

histogram with a higher number of bins than in Figure 5.2 (which is possible because the IAPS values are means computed from a 9 points scale), the histogram computed in the interval [2,4] is then very similar to the one obtained from the self-assessment. This encourages the argument of relative self-assessment and shows that the self-assessed arousal values were not so different from the IAPS evaluations.

Since the self-assessments were close to the IAPS evaluations, particularly for valence evaluations, the ground-truth was defined based on the IAPS evaluations only. This allows to construct two classes of interest: one class corresponding to the positive stimuli and one corresponding to the negative stimuli.

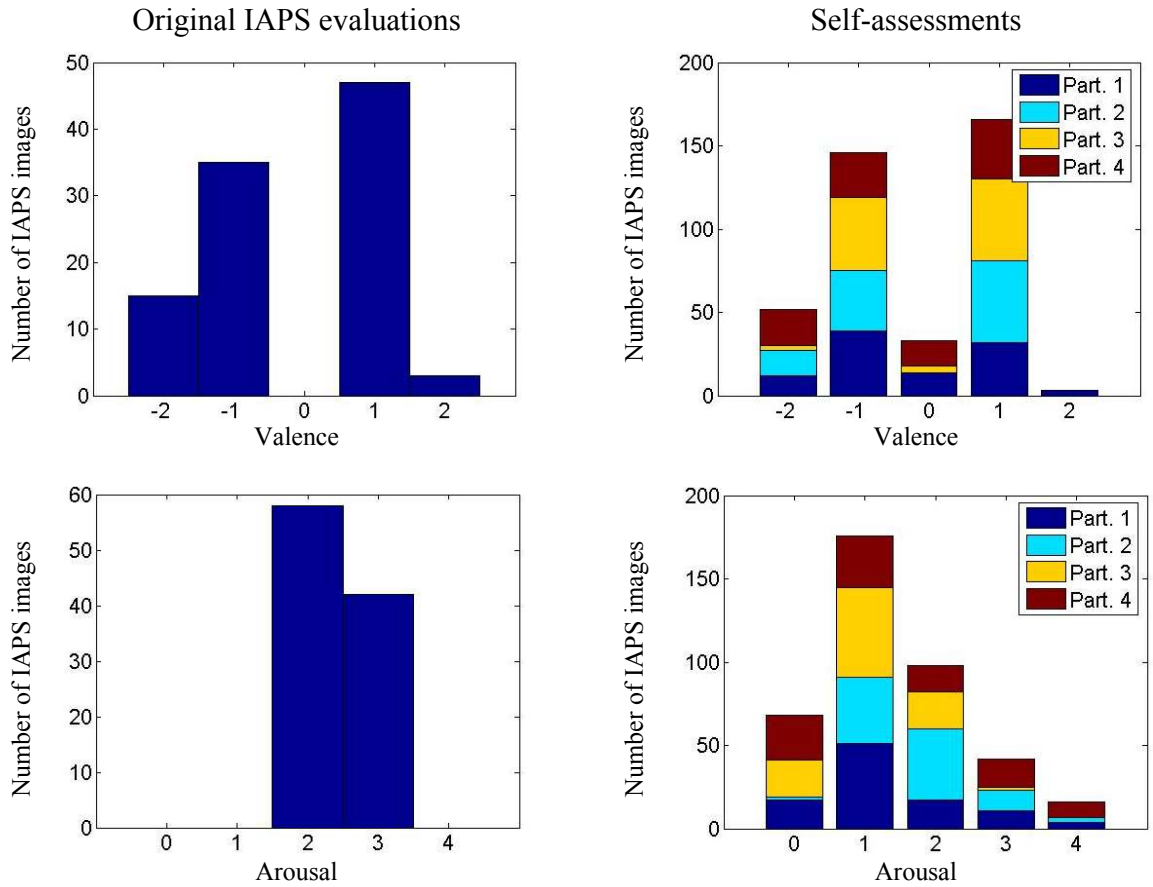


Figure 5.2. Histograms of the IAPS and self evaluations (valence and arousal) for the valence experiment. For easier comparison of IAPS evaluations and self evaluations the IAPS values have been normalized to the same range as the self evaluations.

b. Arousal experiment

As for the valence experiment, the distribution of valence obtained from the IAPS and self-assessment values are quite similar (see Figure 5.3). The higher number of images self-evaluated as having a valence of 1 compared to the IAPS evaluations is mostly due to participants

evaluating originally neutral images (valence value to 0) as slightly positive. This is particularly true for participant 3 as can be seen from Figure 5.3.

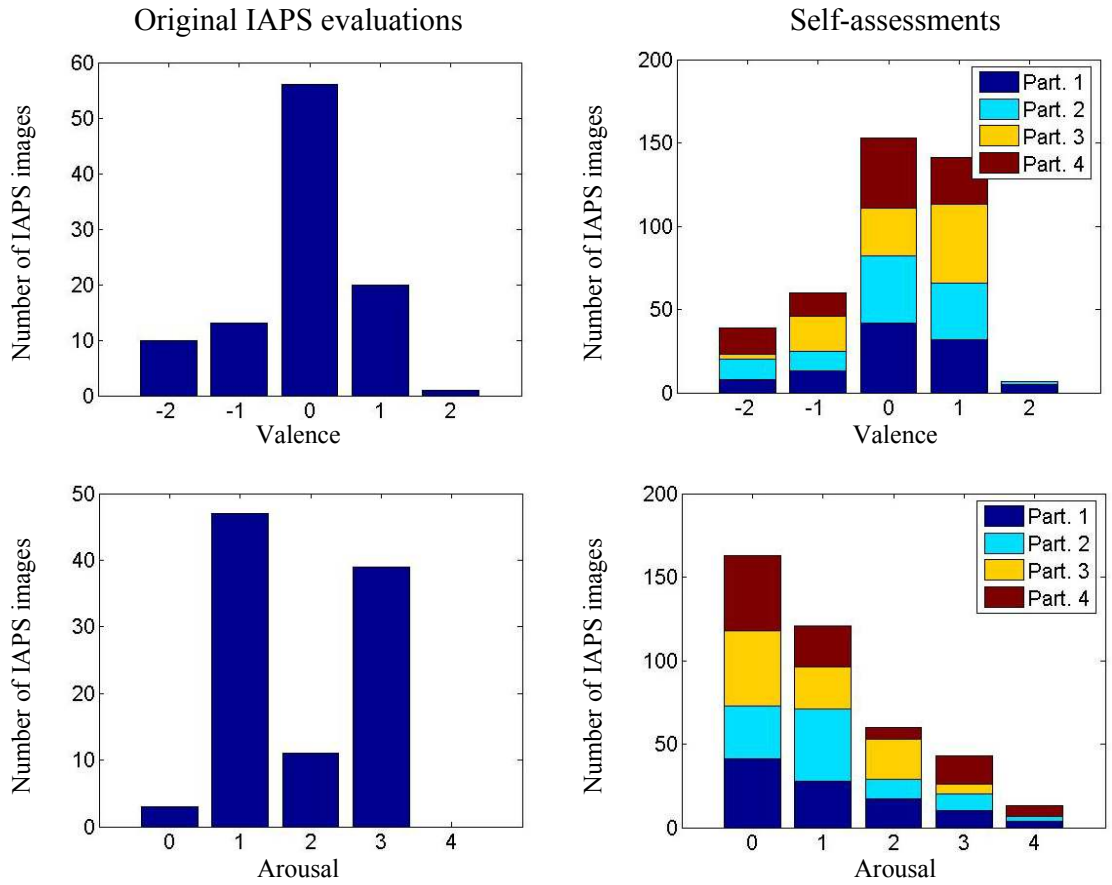


Figure 5.3. Histograms of the IAPS and self evaluations (valence and arousal) for the arousal experiment. For easier comparison of IAPS evaluations and self evaluations the IAPS values have been normalized in the same range as the self evaluations.

For arousal, the histograms from the IAPS and self-assessment values were again different. Due to the constraints applied to construct sets of low and high arousal stimuli, two peaks can be observed for the arousal IAPS histogram. However, those peaks are not clearly visible for the histogram obtained from self-assessments. Only participant 3 obtained similar peaks with a high number of images rated with arousal values of 0 and 2. Since no constraint was applied on the variance of arousal during the selection of the stimuli, the histogram difference is certainly due to a large variability of the arousal judgments across participants and demonstrates the difference in evaluation that can be observed for the same stimuli.

Since the distributions of self-evaluations and IAPS values were found to be different for arousal and the purpose of this experiment is to assess the arousal dimension of emotions, different sets of classes were constructed based on either the IAPS values or the self-assessments.

The images used for the arousal assessment were purposely chosen to be of either very low or very high IAPS arousal values, that is they essentially should have belonged to 2 classes. For this reason, when using the IAPS judgments as a basis to build ground-truth classes, it was natural to divide data into two sets, one for the calm emotions and the other for the exciting emotions. In this way, two well balanced ground-truth classes of 50 patterns each were obtained.

It is more difficult to determine classes from the self-assessment values. As shown by the histograms of arousal, the evaluations are not equally distributed across the 5 choices and in particular do not readily correspond to 2 classes. Taking this into account, two different classification experiments based on the self-assessment were done:

- with 2 ground-truth classes, were the calm class contained patterns judged in the calmest category and the exiting class the others,
- with 3 ground-truth classes (calm, neutral, exciting) were the calm class corresponded to the first of the 5 judgment values, the neutral class to the second and third, and the exciting class to the last two.

Both class definitions led to unbalanced classes, especially for the 3-classes problem: the exciting class contained very few samples (6 to 23 depending on the participant) compared to the calm class (32 to 45) and the neutral class (32 to 55).

5.3.2 Methods

An ANOVA test was applied on the features of the *EEG_Lateral* feature set to control that significant differences were observed in the asymmetry scores between the positively and negatively valenced emotional states. This was done to verify the precedent findings concerning alpha lateralization in the case where emotions are stimulated by pictures and also to check if the asymmetry scores can be used as features for the purpose of classification. Since classification was done in an intra-participant framework (a model was designed for each participant) the ANOVA test was run separately for each participant. For the *EEG_Area* feature set, no ANOVA test was applied because Aftanas et al. [82] already demonstrated the interest of those features for arousal discrimination in a protocol very similar to the one proposed in this chapter.

A Naïve Bayes classifier was first applied on the features of each participant. This classifier is known to be optimal under the assumption of conditionally independent features and in the case of complete knowledge of the underlying probability distributions of the problem. Modeling the underlying distributions is unfortunately difficult in our study, since very few samples are available to construct them; a performance decrease is thus unavoidable. For the sake of comparison, classification based on LDA was also performed. In this case the distributions are assumed to be multivariate Gaussians with no assumption of independence. Due to the rather

limited number of patterns, a leave one out cross validation was preferred to a k-fold strategy in order to maximize the size of the training set (see Section 4.1.2). Results presented in the next section are the percentage of well classified examples.

5.4 Results

5.4.1 Valence experiment

An ANOVA test was applied on the *EEG_Lateral* features to check for a difference in mean between the two valence conditions (Table 5.2). Contrary to what was expected from Davidson's theory [19, 81], no significant differences were found in the frontal area. This could be due to the a lower signal to noise ratio in this region because of facial muscular artifacts that were not removed. Another explanation could be that the lateralization of alpha waves was demonstrated for approach and withdrawal stimuli [19] which are different from positive and negative stimuli. However the lateralization was significant in areas located more at the rear of the brain such as parietal and occipital areas. Those results could partly be explained by the nature of the stimuli since visual processing takes place in occipital areas. Moreover, the absence of alpha hemispheric lateralization for the frontal areas and a significant effect for the parietal area (P3-P4 electrodes) was also found in [157] where the participants had to mentally review film sequences. Since some of the features were shown to have significantly different means between the two conditions, this feature set was used for the purpose of classification.

Electrodes pairs	p-values			
	Paticipant 1	Paticipant 2	Paticipant 3	Paticipant 4
Fp2-Fp1	0.78	0.62	0.57	0.30
AF4-AF3	0.22	0.71	0.59	0.70
F4-F3	0.72	0.43	0.27	0.79
FC4-FC3	0.83	0.17	0.80	0.06
C4-C3	0.52	0.20	0.53	0.45
CP4-CP3	0.75	0.08	0.68	0.04
P4-P3	0.39	0.10	0.09	0.13
PO4-PO3	0.01	0.08	0.01	0.12
O2-O1	0.11	< 0.01	0.59	0.01

Table 5.2. p-values of the ANOVA test applied on the lateralization features for the two groups defined by the IAPS classes (negative vs. positive visual stimuli) and for each participant. p-values < 0.1 are highlited in gray.

The Naïve Bayes classifier accuracies obtained on both the peripheral and EEG features were all around the random level of 50%. Since better results were obtained with an LDA (Figure 5.4) on the EEG feature set this suggest that the Naïve Bayes assumption of conditional independence is not valid in this case. By looking at the covariance matrices of the normalized EEG features for the two classes, some features were found to be correlated. This shows the EEG features conditional dependency and explains the poor results of the Naïve Bayes classifier.

The accuracies obtained with the LDA for the peripheral and EEG features are presented in Figure 5.4. The mean bars on the right of Figure 5.4 represent the average accuracy obtained for the 4 participants (P1 to P4). As can be seen from this figure, the accuracy is higher than the random level for the EEG features and around the random level for peripheral features. The average EEG accuracy across participants is of 58%. This result demonstrates the importance of EEG features for better assessment of the valence dimension of emotions and shows that EEG signals should not be neglected for the assessment of emotions from physiological signals. The lower accuracy obtained for peripheral features is not surprising since the peripheral activity, more specifically the autonomous nervous system activity, is known to better correlate with the arousal dimension than with the valence dimension of emotions [7, 87].

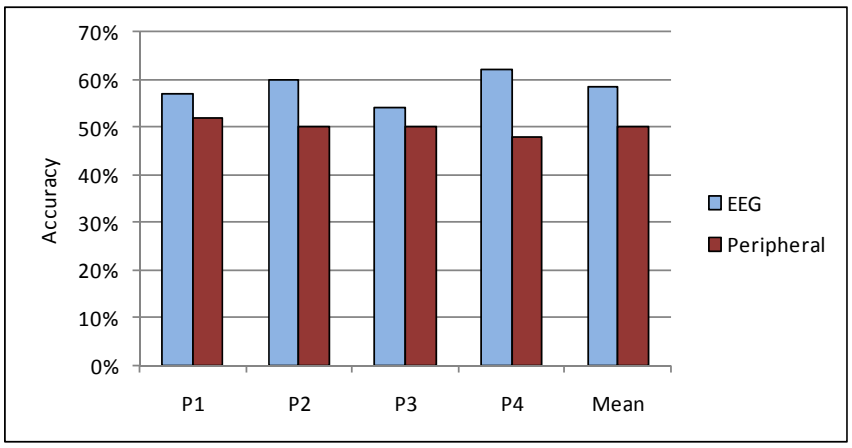


Figure 5.4. LDA accuracy for classification of negative and positive stimuli.

5.4.2 Arousal experiment

When using the two arousal ground-truth classes defined according to the IAPS judgment, the Naïve Bayes classifier average accuracy across participants exceeded the chance level only for EEG features (54% vs. 50%). The LDA classifier performed slightly better, with an average accuracy of 55%, 53% and 54% for EEG, physiological and fused features respectively. Those relatively low accuracies are likely due to large differences between the IAPS values and the actual emotion felt by the participant (as detailed in section 5.3.1.b). We concluded that in our experimental setting the IAPS arousal judgments could not be recovered from actual physiological measurements, and had to use self-assessments. However, using the LDA, the accuracy of peripheral features is higher than the random level. This confirms that the peripheral features computed in this study are more suitable for classification of the arousal dimension than the valence dimension of emotions.

Results with ground-truth classes obtained from self-evaluations are presented in Figure 5.5 and Figure 5.6. The percentage of well classified patterns for the four participants and the average across participants are shown. Compared to accuracies obtained with the ground-truth defined by

the IAPS judgments, accuracies obtained with the self-assessed ground-truth are higher, especially for participants 2 and 3 (Figure 5.5). This tends to confirm that physiological signals better correlate with personalized self assessment of emotion than with the generalized IAPS judgments. The best performance of 72% is obtained by using the EEG signals of participant 2 and a Naïve Bayes classifier. A similar result is obtained with the LDA (70%). For both classifiers, the average accuracy obtained with EEG signals is higher than the one obtained from peripheral signals. Those results stress again the added value of using EEG signals for emotional assessment.

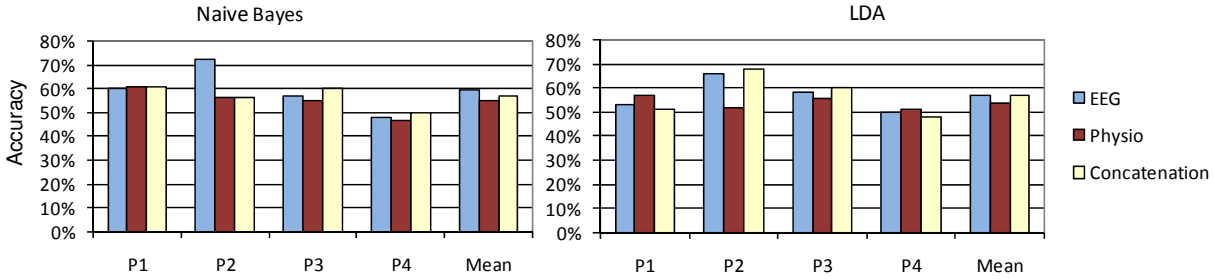


Figure 5.5. Classifiers accuracy with 2 classes constructed from self-assessment.

It is worth mentioning that there are two drawbacks to the problem of unbalanced classes in the case three arousal classes are assessed: (*i*) some of the classes are strongly undersampled and (*ii*) the accuracy measure is less reliable since there are more samples belonging to one of the class than to the others. The first drawback implies that the probability distributions of the undersampled classes cannot be correctly determined which results in a weak assessment of those classes. Concerning the second drawback, since the number of samples in each class was equal for the two feature sets and similar across participants, we believe that the comparison of the emotion assessment performances based on this measure of accuracy is still reliable.

Figure 5.6 shows results for the three class problem. Again, the features extracted from the EEG of participant 2 yield the best result of 58% of well classified patterns (compared to a chance level of 33%). Participant 4 still obtained the worst accuracy. This is likely due to the high number of eye-blinks that were found in the EEG signals of this participant (approximately one blink per second). Participant 1 obtained better results with a Bayes classifier than with a LDA. Extreme results for participants 2 can be explained by a better understanding of the self assessment procedure since he had a good knowledge about emotions, and was likely to accurately evaluate his feelings.

The results obtained for fusion by concatenation are different depending on the participant, the classifier and the number of defined classes. For the Naïve-Bayes classifier, the concatenation of peripheral and EEG features slightly increased the average accuracy for the 3 arousal classes and decreased it for 2 arousal classes. For the LDA, concatenation of features increased the average

accuracy by 5% for 3 arousal classes and did not affect it in the other case. Thus the LDA seems more appropriate for fusion at the feature level, which could be explained by the weak assumption of conditional independence of the Naïve-Bayes classifier. Also, fusion provides more robust results since some participants had better scores with peripheral signals than with EEG's and vice-versa.

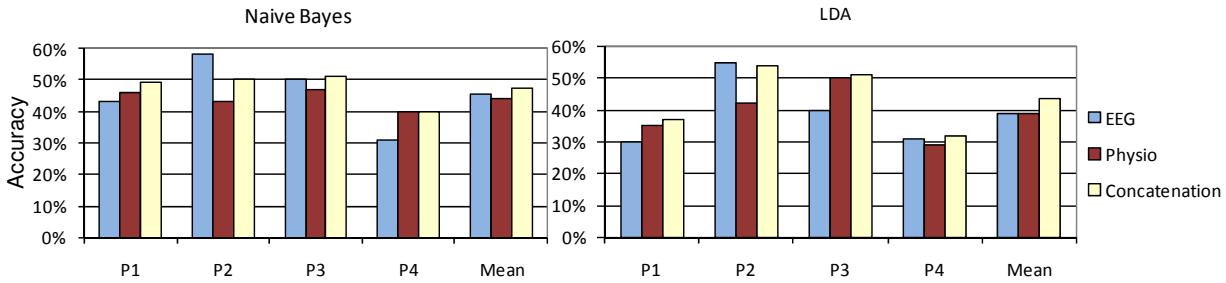


Figure 5.6. Classifiers accuracy with 3 classes constructed from self-assessment.

5.5 Conclusion

In this chapter two categories of physiological signals, from the central and from the peripheral nervous systems, have been evaluated on the problem of assessing the arousal and the valence dimension of emotions elicited by IAPS images. Those assessments were performed as classification problems, with ground-truth valence / arousal values provided either by the IAPS or by self-assessments of the emotion. Two classifiers were used, a Naïve-Bayes classifier and a LDA.

Results showed the usability of EEG's in both arousal and valence recognition and the interest of EEG features over peripheral features. Moreover, the fusion of EEG features with peripheral features improved the assessment performance. This improvement was better with a LDA than with the Naïve-Bayes classifier. Results also markedly improved when using classes generated from self-assessment of emotions. When trying to assess emotions, one should avoid using predefined labels but rather ask for the user's feeling. However, by using self-assessment the generated classes were unbalanced which gave rise to classification problems such as the undersampling of some classes. Moreover, using the self-assessments as a ground-truth implies that the user is an expert in evaluating his / her feelings, which is not always the case as discussed in Section 4.1.1.

Future work on arousal assessment will first aim at improving on the current results by using other non-linear classifiers, such as Support Vector Machines. Feature selection and more sophisticated fusion strategies will also be examined, jointly with the examination of other features such as temporal characteristics of signals that are known to be strongly implied in emotional processes.

Chapter 6 Assessment of self-induced emotions

6.1 Introduction

Fairly recent psychological studies regarding the relations between emotions and the brain have uncovered the strong involvement of cognitive processes in emotions [61, 79-81, 84, 158]. Among those studies, some used the recall of past emotional events to self-induce emotions [84, 158]. This method has the advantage of activating many brain areas because cognitive processes related to memory retrieval are located throughout the brain. Asking participants to self-induce emotions is also useful because the remembered emotions will correspond with the emotion that is required by the protocol. This allows for an optimal control over the number of emotions that are elicited per class. From a classification point of view, this is interesting because it avoids the problem of unbalanced classes encountered in Chapter 5. Other advantages of this elicitation method are presented in Section 4.1.1. To our knowledge, despite of all those advantages, this method has never been used for emotion assessment from EEG features (see Section 2.3).

Having in mind the previous considerations, the present Chapter aims at investigating the usefulness of EEG and peripheral signals in a self-induction paradigm. In order to be as much application independent as possible, we used the valence-arousal space as a prior model to define three emotional classes of interest that are calm-neutral, positive-excited and negative-excited. The protocol and the features computed from the recorded signals are presented in Section 6.2. Section 6.3 describes the complete framework used for emotion recognition. Finally, results are presented discussed in Section 6.4.

6.2 Data acquisition

6.2.1 Acquisition protocol

In [106] five factors that can influence recordings were defined: subject-elicited vs. event-elicited, laboratory setting vs. real world, focus on expression vs. feeling of the emotion, openly-recorded vs. hidden recording and emotion-purpose vs. other-purpose. The following protocol description addresses those five factors as well as those emphasized in Section 2.3.

In the present study a self-induced method, using recall of strong emotional episodes, is employed to elicit reliable and short time emotions. An episode is defined as a situation that lasted for a several minutes and potentially containing several events and actions with the same emotional orientation. An example is the funeral of a relative including events such as moments of the ceremony and the burial in itself. The elicited emotions are considered reliable because *(i)* thinking of the same episodes ought to produce similar reactions from one trial to another, *(ii)* emotional episodes are often stored in memory because the emotions felt were quite intense,*(iii)*

emotional recall is a cognitive task that induces EEG activity [84, 158] as well as modify peripheral activity [6, 92].

Compared to other studies [93, 106, 111], where emotions are elicited and assessed over several minutes, the duration of an emotion epoch in this study is merely 8 s. This epoch duration was chosen because it is the maximum duration that allows maintaining the total length of the protocol below one hour to avoid participant fatigue. Within the requirement of the one hour duration this epoch is maximized for three reasons. Firstly, some peripheral features need to be determined over a sufficiently long period of time in order to be reliably computed; this is for instance the case for statistical features extracted from HR. In general, an epoch of 8 s should suffice for this purpose if we do not consider the very low frequency features such as low frequency HRV [96]. Secondly, recalling past episodes and eliciting the corresponding emotions are difficult tasks and participants might need a few seconds to accomplish them. Thirdly, the reaction time of peripheral activity from the moment where the emotion is elicited is of several seconds, with the GSR being the slowest response with a lag around 1-4 seconds [88].

The 11 participants (7 males, 4 females) who took part in the study were aged from 26 to 40, one being left-handed. One week before the recording, participants were told to retrieve from their memory one excited-positive and one excited-negative episode that had occurred in their life and that they consider as being most powerful. On the day of the experiment, each participant was given a consent form where the context, the goal and a short explanation of the experiment were provided. Participants had to sign this consent form to continue further and could stop the experiment whenever they wanted. After signing the consent form, sensors were attached to the participant who was seated in front of a computer screen. A precise description of the protocol was provided with a support demonstration. This type of experiment corresponds to Picard's factors for an open-recording (participants knew they were recorded), emotion-purpose (participants new the objective of the study), and laboratory settings (participant are recorded in a controlled environment).

The complete recording session was divided into trials. During each trial participants had to accomplish a particular task according to the visual cue displayed on the monitor after a random duration display of a dark screen (Figure 6.1). This task could be to self-generate one of the two excited emotions by using the past emotional episodes of their life as a support, or to stay calm and relax in order to define a third emotional state called calm-neutral. A total of T = 300 trials (100 trials per emotional state) were performed in a random order. Since facial muscle artifacts can contaminate EEG signals, participants were encouraged not to express their feelings through facial expressions, not to blink, and not to close their eyes during the 8 s of recordings (despite of this instruction some involuntary facial expressions artifacts can still occur in the signals). Emphasis was thus put on the feeling of emotions rather than on the cognitive task of

remembering and on the motor expressions of emotions. A resting period of unlimited duration to relax and stretch muscles was proposed to participants after each block of 30 trials. As can be seen from Figure 6.1, the chosen emotional states do not cover all areas of the valence-arousal space, especially in the bottom half of the space. This choice was made because there are actually few emotions that are calm-negative or calm-positive [23, 54].

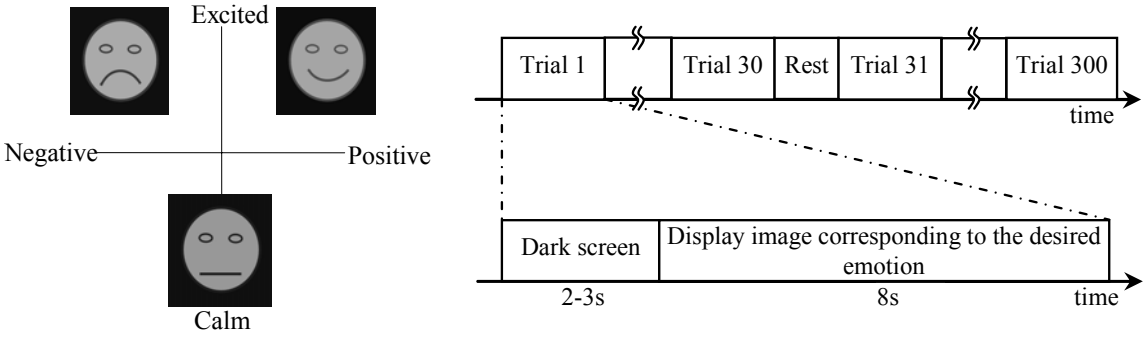


Figure 6.1. (left) The different emotional classes represented in the valence-arousal space and their associated image. (right) schedule of the protocol and detail of a trial.

Data were recorded using the Biosemi Active II system. EEG signals were recorded using 64 surface electrodes. Plugged on the same system to simplify synchronization, other sensors were used to record peripheral activity: a GSR (Galvanic Skin Response) sensor to evaluate sudation, a respiration belt to record abdominal expansion and a plethysmograph to measure blood pressure. Both EEG and peripheral signals were sampled at 1024 Hz.

After data acquisition, participants were asked to report verbally on their experiences in an informal interview. Participants were not asked to provide a detailed description of the chosen episodes because we believe that for personal and ethical reasons a participant may hesitate to refer to his / her strongest emotional experiences. For this reason the differences in the cognitive tasks between different trials could not be fully controlled. However, as argued in [84] the known effectiveness of mental imagery as an elicitator of powerful emotions can compensate this problem.

The present protocol for off-line acquisition of physiological signals is very close to those encountered in the BCI community so that the conclusions drawn from this study may also have some impact in this direction. An emotion elicitation task can then be regarded as a mental task that the user tries to perform in order to communicate his / her feelings. This can be useful for severely disabled people that cannot directly express their emotions. Current BCI paradigms [27, 29, 30] aim to detect brain activity that corresponds to complex tasks (mental calculus, imagination of finger taping, etc.) not related to the objective of the user (moving a mouse cursor, choosing a letter, etc.). Generally the user needs training before using such systems. In case the

objective of the user is to express an emotion, classical BCI tasks (e.g., imagination of finger tapping) seem to be really far from this objective and it is more appropriate to use tasks such as the remembering of a similar emotional episode.

6.2.2 Feature extraction

From the EEG signals, 2 feature sets were extracted to analyze their performance for emotion assessment. Those features were computed on the last 7.5 seconds of the signal. The first 0.5 seconds were removed for the two following reasons. Firstly, this time window was not expected to contain emotional information related to the requested emotional episode since it is rather unlikely that a participant start to remember the episode directly after the display of the associated image. Secondly, this part of the signal contains the P300 [159] wave and may contain emotional information related to the nature of the presented image (smiling or sad smiley) which is not the stimulus studied here.

The 2 feature sets computed from the EEG signals are (See Section 3.3.2.a):

- the *EEG_STFT* feature set, obtained from the spectrogram of the EEG signals, containing in that case $9 \times 64 \times 29 = 16704$ features (9 frequency bands, 64 electrodes and 29 time frames);
- the *EEG_MI* feature set, containing the mutual information between each pairs of the 64 electrodes with a total of $\frac{(64-1)64}{2} = 2016$ features.

As can be seen, both feature sets are of very high dimensionality.

Peripheral signal	Standard features			Advanced features	Reference
	μ_x	σ_x	δ_x		
GSR	X		X	$f_{GSR}^{DecRate}$, $f_{GSR}^{DecTime}$	Section 3.3.2.b
BVP	X				
Heart Rate (HR)	X	X	X		
Respiration	X	X	X	$\mathbf{f}_{Resp}^{Pow}$, f_{Resp}^{DR}	Section 3.3.2.e

Table 6.1. The features extracted from the peripheral signals

The peripheral features extracted from the corresponding signals are given in Table 6.1. A detailed explanation of the features can be found in Section 3.3.1 for the standard features. The "Reference" column indicates the section of Chapter 3 that corresponds to an explanation for the given advanced features. All the peripheral features were computed from the complete duration of the 8 seconds trial and concatenated in a feature vector. The HR signal was computed from the BVP signal of the plethysmograph as discussed in Section 3.2.4. No a-posteriori correction of the

wrongly detected beats was applied because of the short duration of a trial. For a given trial, all the peripheral features were concatenated in a unique feature vector containing a total of 18 features. The peripheral feature set constructed by that way was called *Peripheral*.

6.3 Classification

6.3.1 The different classification schemes

From the protocol detailed in Section 6.2.1, the ground-thruth is easily defined by attributing to each trial the corresponding label of calm, positive-excited and negative-excited. However, to analyze the classification performance in different classes' formulations, the five classification tasks described bellow were tested.

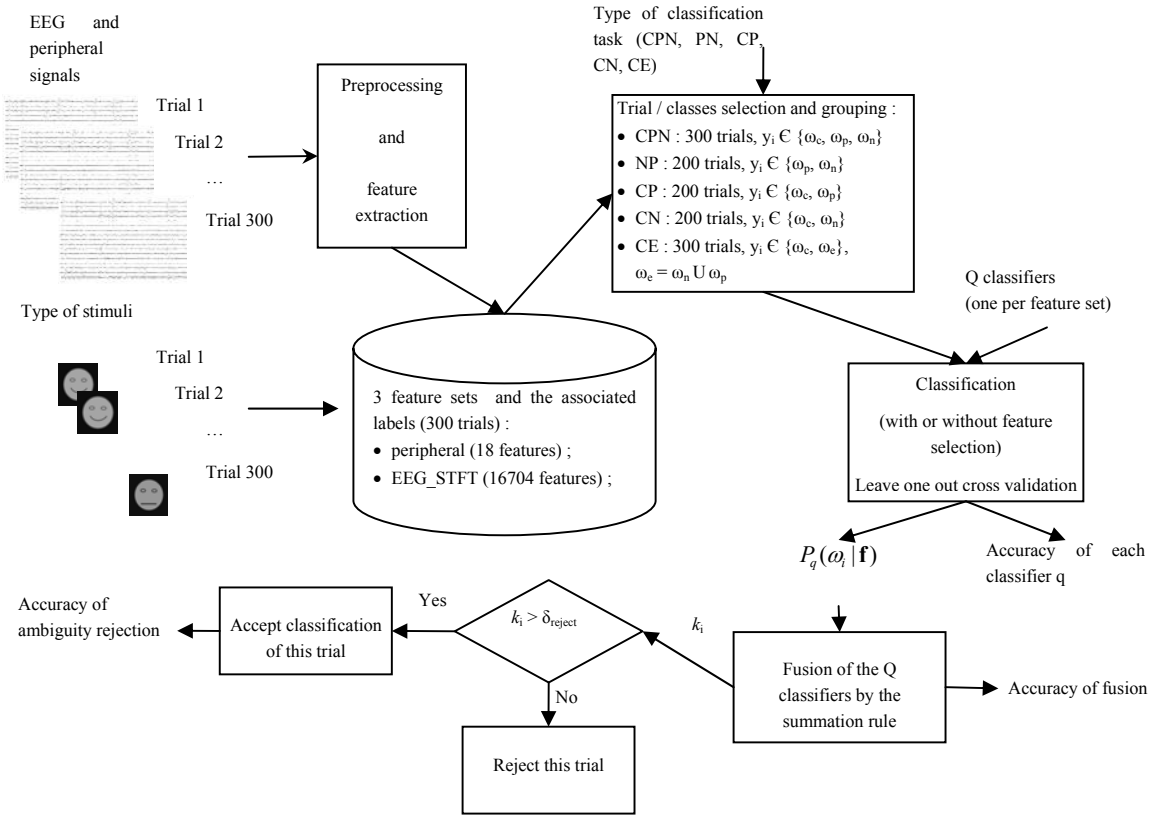


Figure 6.2. Complete process of trial acquisition, classification, fusion and rejection for a given participant. As defined in Chapter 4, k_i is the confidence measure of class ω_i after opinion fusion and δ_{reject} is the rejection threshold.

Since each recorded trial corresponds to a particular emotional state, it is easy to formulate a classification task (called CPN for "calm", "positive", "negative") where the three ground-truth classes ω_c, ω_p, ω_n correspond to calm-neutral, positive-excited and negative-excited patterns. A target class vector $\mathbf{y}^{CPN} = [y_1 \ldots, y_i, \ldots, y_{300}]^T$ is constructed, where $y_i \in \{ \omega_c, \omega_p, \omega_n \}$ represents the

class of the trial *i*. We also address other classification tasks by constructing different target vectors to distinguish between the following emotional states: negative excited vs. positive-excited (NP), calm-neutral vs. positive-excited (CP), calm-neutral vs. negative excited (CN), calm vs. excited (CE) by regrouping samples of the positive-excited and negative-excited states.

As summarized in Figure 6.2, there are three sets of features, *Peripheral*, *EEG_STFT* and *EEG_MI* that contain respectively peripheral features, STFT EEG features and MI EEG features for all trials. Those feature sets are associated with the class vectors $\mathbf{y}^{CPN}$, $\mathbf{y}^{NP}$, $\mathbf{y}^{CP}$, $\mathbf{y}^{CN}$ and $\mathbf{y}^{CE}$, depending on the classification task to address.

6.3.2 Classifiers

Since the EEG feature sets (*EEG_STFT* and *EEG_MI*) are of very high dimensionality (thousands of features) compared to the number of samples in the sets (200 or 300 depending on the classification scheme), there is always a linear boundary that can completely separate training samples of the different classes. For this reason only linear classifiers were applied on those feature sets. Another advantage of using linear classifiers is that they give better generalized solutions. The issue of feature space reduction was also investigated as detailed in Section 6.3.3.

The following linear classifiers were applied on the EEG feature sets:

- the LDA because it can provide probabilistic output which is useful for the purpose of fusion. Since the EEG feature spaces are of high dimensionally it sometimes occurred that the covariance matrix of the features was singular, in that case the diagonalized version of the LDA was employed;
- the linear SVM since they are known to have good performance in high dimensional spaces [138];
- the probabilistic linear SVM to obtain probabilistic output;
- the RVM since, as the SVM, it should perform well in high dimensional spaces and provides probabilistic outputs.

The above mentioned classifiers were also applied on the *Peripheral* feature set. Since this feature set is of a lower dimensionality than the EEG feature sets, the performances of the following non linear classifiers were also investigated:

- the QDA to obtain quadratic decision boundaries together with probabilistic outputs;
- the RBF SVM, where the RBF kernel size was chosen according to the procedure explained in Section 4.1.3.c;

- the probabilistic RBF SVM to obtain probabilistic output.

In the present study, different classifiers were trained on the three feature sets to recover the ground truth classes. A classifier was trained separately for each participant and the accuracy of each classifier was evaluated using the leave-one-out strategy. This involves using each feature vector in turn as the test set and the remaining ones as the learning set. At each step, a classifier is trained from the learning set and then applied to the test sample. This leave-one-out strategy was chosen since it provides the maximum possible size of the learning set. This is preferable in this problem because the number of samples (N=300) is very low compared to the size of the EEG feature spaces.

6.3.3 Reduction of the feature space

In order to study the effects of feature space reduction on the classification accuracy, several methods were tested: the Fast Correlation Based Filter (FCBF) presented in Section 4.2.1c and filtering based on the Fisher criterion presented in Section 4.2.1b. The classification accuracy of the LDA and the linear SVM were evaluated for different parameter values of these feature selection methods. Since both these methods are computationally expensive, they were tested on one participant only, namely participant 1 (who was the first recorded participant). The reduction of the *EEG_STFT* feature set was addressed since this is the one that contains the higher number of features.

The δ_{FCBF} parameter of the FCBF algorithm represents the minimum correlation value with the class labels $\mathbf{y}$ that a feature f should have in order to be selected. This parameter is thus related to the number of selected features and the higher the parameter the lower the number of selected features. In this study, the value of this parameter ranged from 0.05 to 0.3 with a step of 0.05. The higher bound of 0.3 was chosen because most of the features had a correlation value below this value (for instance, for the CPN classification scheme only one feature has a correlation value higher than 0.3). At the lower bound of 0.05 approximately half of the features are removed in the first step of the FCBF algorithm.

The Fisher criterion defined in Section 4.2.1b can be used to rank the features by relevance order. It is then possible to select only the most relevant features and filter out the others, which allows to compute the classification accuracy for a given number of selected features. In this study, the performance of this filter technique was studied for numbers of features taken in the interval [1 F_{EEG_STFT}] (where F_{EEG_STFT} is the number of features in the complete *EEG_STFT* feature set). This method was used only for the CPN classification scheme.

6.3.4 Fusion and rejection

To investigate the advantages of fusion of several physiological feature sets, classification accuracy was studied for both of the methods presented in Section 4.3: fusion at the feature level and fusion at the classifier level. Fusion was done independently for each participant and the combination of the following feature sets was studied:

- *EEG_STFT* and *EEG_MI* feature sets, to analyze the performance of fusion of different feature sets computed from the same EEG modality,
- *EEG_STFT*, *EEG_MI* and *Peripheral* to analyze the performance of fusion of information originating from the CNS with information originating from the PNS.

Fusion at the feature level was done by the concatenation of features as explained in Section 4.3.1. The fusion at the classifier level was done by opinion fusion, using the sum rule detailed in Section 4.3.2a. In this case, a classifier was chosen for each feature set to form the ensemble Q of classifiers (with $|Q| = 2$ or $|Q| = 3$ depending on the number of feature sets used for fusion). The ensemble Q was composed of the classifiers with probabilistic outputs that generally had the best accuracy on the associated feature set.

As a final step the method detailed in Section 4.4 to reject samples with a low confidence k_i value was applied to the CPN, NP and CE classification scheme. Those schemes were chosen because they are the most relevant for HCI applications. To analyze the performance of this rejection, the average accuracy across participants computed on the non-rejected samples and the percentage of rejected samples were plotted as a function of the δ_{reject} threshold (taken in the range [0 1]). Since the label of each trial is already determined after fusion, it is possible to compare the number of badly classified trials that are rejected to the correctly classified ones. The sums over participants of the correctly classified and badly classified samples that are rejected are also plotted as a function of the δ_{reject} threshold. From the analysis of those curves, a value for the δ_{reject} threshold that improves the accuracy while keeping the percentage of samples reasonably low will be suggested.

6.4 Results

6.4.1 Participants reports and protocol validation

Out of the 11 recorded participants 10 reported a successful elicitation of the emotions by recalling emotional episodes. As can be seen from Figure 6.5, which represents the average accuracies obtained from the 10 participants cited above, the peripheral activity is useful to distinguish between different classes of emotions. This implies that different patterns of physiological activity where induced for each emotional task and thus supports the idea that

emotions were successfully elicited. However, all participants reported that it was really difficult to stay concentrated throughout the entire recording. A recurring observation was also that switching from one emotion to another very quickly was sometimes confusing and hard to accomplish. The effects of such observations can be missing trials where the participants did not accomplish the requested task, the elicitation of the undesired boredom emotion which can interfere with positive and negative excited trials, and noisy EEG signals due to fatigue. The emotions used in the two excited categories were also different from a participant to another (for instance a participant elicited a negative-excited emotion by remembering a past episode where he felt fear while another participant used an episode where he felt anger).

In the protocol presented in Section 6.2.1 brain activity can be induced by two cognitive components: the actual events of the episode (for instance thinking of someone crying) and the emotion elicitation following the event. Since our aim is to detect emotions, it is important to control that the events used to induce emotions were not always the same to ensure that what is detected from brain signals is the emotion and not the cognitive task related to the event (for instance mental imagery of the act of crying). Since participants did not report about the episodes they used to induce emotions it is difficult to control for this, however the following remarks lead us to assume the protocol is valid:

- two participants reported that they thought of different episodes within the same category (i.e. positive-excited and negative-excited). The classification accuracy obtained from the signals of one of those two participants actually corresponds to the best results across the 10 participants while the other one obtained average accuracies;
- one participant reported that he thought to the episodes without concentrating on the feeling of emotions which resulted in a weak emotion elicitation. All the accuracies computed from the signals of this participant are at the random level;
- since an episode was defined as including several emotional events of the same category, it is unlikely that the participants always thought of the same event to elicit one of the emotions;
- the participants were explicitly told to focus on the feeling of emotions and emotions were successfully elicited as stated above.

Notice that the participant who did not concentrate on the feeling of emotions was removed for further analysis since he did not follow the protocol properly.

6.4.2 Results of single classifiers

Figure 6.3, Figure 6.4 and Figure 6.5 respectively present the mean accuracy across participants for the *EEG_STFT* features, the *EEG_MI* features and the *Peripheral* features. The accuracies of different modalities and classifiers are compared below to answer the following questions: what

is the effectiveness of EEG and peripheral features to assess emotions according to the different classification schemes and which classifiers should be used for latter fusion of feature sets.

The *EEG_STFT* features provided interesting results with a mean classification accuracy of 63% for three classes and a SVM classifier (the random level is at 33% accuracy). The best average accuracy for two classes is obtained from the CP classification task with nearly 80% of well classified trials (random level at 50%), followed by the CE and NP classification tasks with respectively 78% and 74% of accuracy. For all participants and all classification tasks, the results are higher than the random levels (33% for three classes and 50% for two classes). The *EEG_MI* features seem to be a bit less suitable for emotion classification than *EEG_STFT* features with an approximate decrease of well classified trials of 2% to 4%, except for the NP classification task where a slight performance increase was noted. It is hard to compare those results to the state of the art because there are only few studies using EEG. In the previous study reported in Chapter 5 the best accuracy on two and three arousal classes was respectively of 72% and 58%. In this study the highest accuracies for the CE and CPN classification tasks are respectively of 88% and 86.3%. The best result for a two class task is obtained on the NP task with 96% of accuracy. In [115] an accuracy of 29% was obtained for 6 different emotional classes while the accuracy was of 42% for identification of 5 emotional states in [113]. Our results are thus superior to the previous studies using the EEG modality for detecting emotions expressed in the valence-arousal space and in alignment with results obtained on emotional labels.

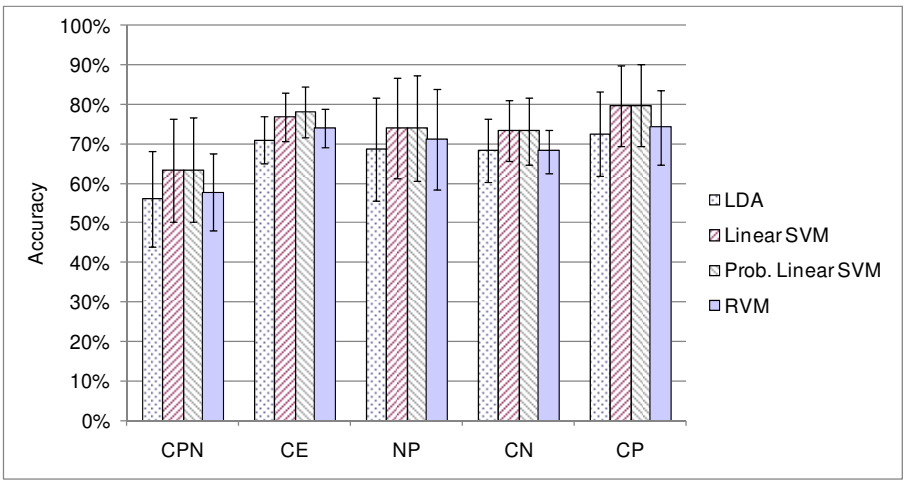


Figure 6.3. Mean classifier accuracy across participants for the *EEG_STFT* feature set and the different classification schemes. The bars on top of each column represents the standard deviation across participants.

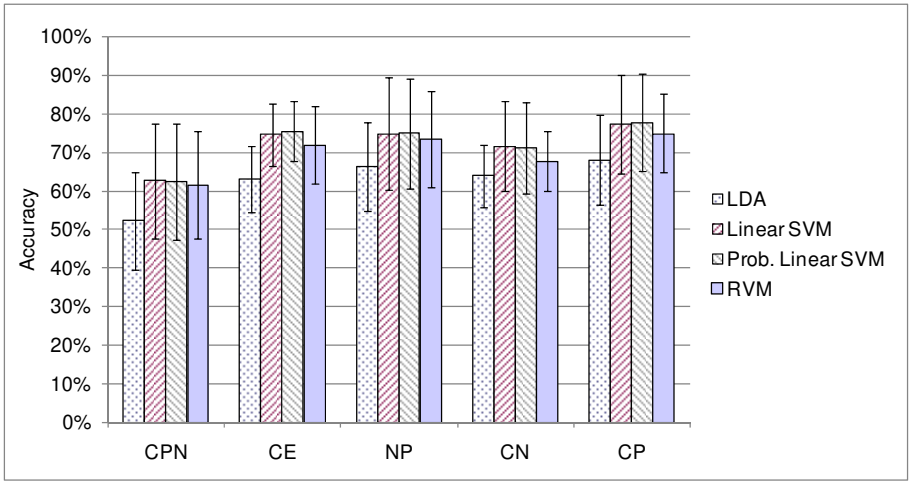


Figure 6.4. Mean classifier accuracy across participants for the *EEG_MI* feature set and the different classification schemes. The bars on top of each column represents the standard deviation across participants.

To check for the usability of this emotional protocol as a new BCI paradigm our results were compared to BCI accuracies. In [160] the authors showed that around 75% of the 99 untrained participants that took part in a two class BCI paradigm without feedback obtained accuracies between 60% and 79%. The distributions of the accuracies for our recall paradigm are similar; however more participants should be recorded to validate this statement. Our results are also far from those of more recent BCI studies where the accuracy can reach more than 90% for two classes for many untrained participants [30]. This can be due to the definition of mental task that are chosen to activate well separated areas of the brain, contrary to the task definition used in this study.

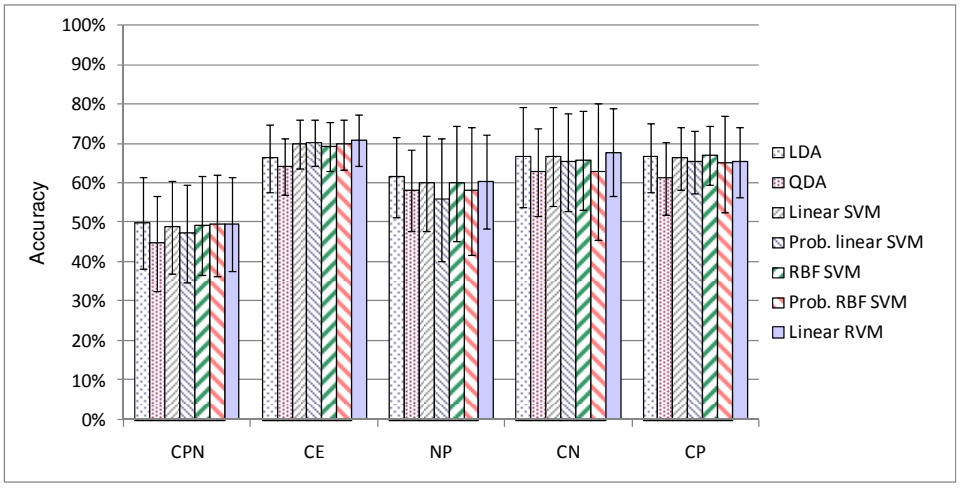


Figure 6.5. Mean classifier accuracy across participants for *peripheral* features and the different classification schemes. The bars on top of each column represents the standard deviation across participants.

If the three figures (Figure 6.3, Figure 6.4 and Figure 6.5) are compared, it is obvious that the EEG features lead to better accuracy than peripheral features for all classification schemes. For peripheral features the LDA classifier is the best with an average accuracy of 51% for three classes and around 66% for two classes (except for the NP classification task with accuracy around 61%). Results ranged from nearly the random level up to around 80% for two class formulations and from 37% up to 75% for three classes, showing the importance of this modality for at least one participant. However there is an exception, the CE classification task, where the LDA does not have the best accuracy. In this task the sparse kernel machines have better accuracies but they were sensitive to the unbalanced nature of this configuration with 200 samples belonging to the excited class and 100 samples belonging to the calm class. As can be seen from the confusion matrices of Table 6.2, sparse kernel machines tend to always assign the excited class to test samples. Those results where thus considered as irrelevant and the LDA classifier chosen as the most relevant classifier for fusion.

(a)

Truth \ Classified	Calm	Excited
Calm	65%	35%
Excited	33%	67%

(b)

Truth \ Classified	Calm	Excited
Calm	31%	69%
Excited	11%	89%

(c)

Truth \ Classified	Calm	Excited
Calm	13%	87%
Excited	3%	97%

(d)

Truth \ Classified	Calm	Excited
Calm	33%	67%
Excited	10%	90%

Table 6.2. Average confusion matrices across participants for peripheral features and different classifiers: (a) LDA, (b) Linear SVM, (c) RBF SVM and (d) Linear RVM.

Compared to the state of the art of emotion assessment from peripheral signals and time segments of similar duration our results are under those reported. In [107] 90% and 97% of accuracy was obtained using time windows of 2 s for valence and arousal assessment respectively. However, the accuracy they report represents the number of samples from which the output of a neural network regressor falls in a 20% interval of the target value. Thus this accuracy cannot be directly compared to classification tasks. In [114] the classification strategy discriminated three emotional states (neutral, positive and negative) with an accuracy of 71% from 6 s signals. The classification strategy used in [114] included a detection of signals corruption, which demonstrates the importance of such a procedure for correct emotion assessment. However, the accuracy was computed for only one participant after training the algorithm on 8 participants. To give an example of the variability of results that can be obtained from a participant to another, in our study results ranged from 39% for the worst participant to 78% for the best considering only the LDA classifier (the classifier having the lowest variance) and the CPN classification task.

The large differences in accuracy between the EEG and peripheral features can be explained by two factors. Firstly, the protocol is based on a cognitive elicitation of emotions where participants are asked to remember past emotional episodes which ensures strong brain activities. Moreover, the emphasis was put on the internal feeling of emotions rather than on the expression of emotion that can help to induce peripheral reactions [5]. Secondly, the 8 s length of trials may be too short for a complete activation of peripheral signals while it may be sufficient for EEG signals.

For both EEG and peripheral features there is always high variability of results across the participants. For instance the accuracies ranged from 45% to 86% to classify emotions in three classes using the *EEG_STFT* features and the Linear SVM classifier. This variability can be explained by the fact that the participants had more or less difficulty in accomplishing the requested tasks as reported during the interview. Another remark that holds for all feature sets is that the detection of arousal states is more accurate than the detection of valence sates. This is not surprising for peripheral activity since it is known to better correlate with the arousal axis than with the valence axis [7], and sheds some new light on the usability of EEG for the detection of arousal and valence. Notice that the standard deviation is lower for arousal identification than for all other combination of emotional states showing that arousal states are detected with more stability across participants. The RVM classifier also tends to classify emotion with more stability but unfortunately obtained a lower average accuracy than SVM for most classification schemes.

This study also allows comparing the performances of the different classifiers in the three feature spaces. For the *Peripheral* feature set (Figure 6.5), the classifiers have relatively similar accuracies except for the QDA which performs poorly compared to the others. Since this algorithm needs to compute a covariance matrix for each class, the low number of samples that are available for learning (around 100 per class) explains this result. The RBF SVM does not perform as well as the other classifiers for the two classes formulations, suggesting that those problems are linear by nature. For the high dimensional spaces of EEG features the LDA accuracy is always about 10% below the results obtained by SVM classification. This confirms the effectiveness of SVM's in high dimensional spaces [138]. One of the goals of the present work was also to determine which of the RVM and probabilistic SVM would have the best accuracies in order to use the best algorithm for the purpose of fusion. As can be seen from Figure 6.3 and Figure 6.4, the probabilistic SVM performs as well as the standard SVM demonstrating the interest of such a classifier to perform fusion on the basis of standardized scores. The RVM classifier outperforms the LDA, showing its adequacy for high dimensional spaces but does not outperform the SVM. An explanation could be that RVM's generally used less support vectors than SVM's which is not desirable in those undersampled classification tasks where good generalization is hard to obtain.

6.4.3 Results of feature selection

Feature selection was applied for participant 1 on the *EEG_STFT* feature set. Only the data recorded from the first participant was subject to this analysis because of the computational time needed to select features from this high dimensional feature set. The FCBF algorithm was applied on all classification schemes while the filter method based on the Fisher criterion was applied only the CPN classification scheme.

a. FCBF

For the FCBF feature selection, Figure 6.6 shows classification accuracies as a function of the threshold δ_{FCBF}. For LDA, accuracies for all two-class sets without feature selection were obtained using a subset of the original feature set, generally with δ_{FCBF}=0.2 and a subset size of 30 to 80 features. FCBF even succeeded in improving the results for about 6% for the CN classification task. The significant reduction of the number of features (less than 0.5% of original features are kept) can help improve computation time and storage capacity in a practical application.

The number of features that have a correlation with the class labels lower than 0.05 represents around half of the original size of the *EEG_STFT* feature set. However for all classification schemes the FCBF algorithm keeps only around 100 of those features. This is due to the second step of the FCBF algorithm that removes redundant features and shows that many features are correlated. This correlation is certainly due to the construction of the *EEG_STFT* feature set: two energy features extracted from the same electrode signal but in a neighboring time window are likely to be correlated, as are two features computed over the same time window and from the signals of two neighboring electrodes.

SVM accuracy is higher without feature selection, confirming its intrinsic capacity of good generalization in high dimensional spaces. For both LDA and SVM feature selection was not effective on the set of three classes (42% of accuracy for one selected feature and up to 47% for 6 selected features). This can be due to the nonlinear nature of this problem since the FCBF algorithm relies on linear correlation. One solution can be to substitute linear correlation by mutual information as a measure of relevance [149]. Another explanation for the accuracy decrease when using feature selection is that the FCBF algorithms eliminates features without taking into account the quality of the whole feature subset as is the case with wrapper algorithms. FCBF then fails to find features that are interacting and such that they are improving classification accuracy with the complete set of features.

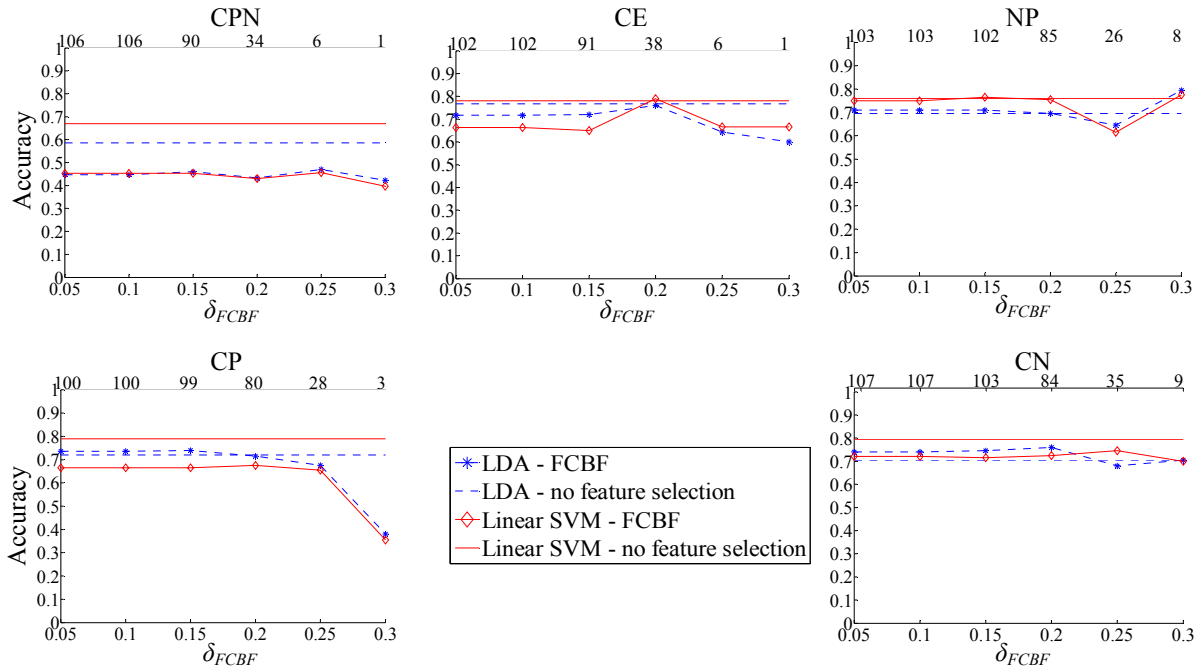


Figure 6.6. classification accuracy using participant 1 EEG_STFT features with LDA and with SVM on the five sets of classes, with or without FCBF feature selection. The bottom horizontal axis indicates the value of the threshold δ_{FCBF}, while the top horizontal axis corresponds to the number of selected features.

b. Fisher feature selection

Since the FCBF algorithm has a poor performance on the CPN classification scheme, the filter method based on the Fisher criterion was applied on this problem and the accuracies of LDA and linear SVM classifiers were analyzed. Figure 6.7 shows the accuracy of those classifiers as a function of the number of selected features, the last point of the curve being the accuracy for the complete set of 16704 features.

As can be seen from by comparing Figure 6.6 (CPN classes) and Figure 6.7 the accuracies obtained with the Fisher feature selection are better than those obtained with the FCBF algorithm. With Fisher selection, the accuracy of the LDA is 47.3% by selecting only one feature and 59% by selecting 100 features compared to the best accuracy of 47% obtained with the FCBF algorithm (the number of feature ranging from 1 to 106). The first peak in accuracy can be observed for 1000 selected features using a LDA and 1500 selected features using the SVM. This point corresponds to the best accuracy for the LDA and demonstrates the usefulness of feature selection for this classifier. By keeping 1000 features only 6% of original features are selected which is interesting for computational complexity and storage improvement. However, the number of features is still high at this point which shows that many features from the STFT feature set are required for accurate detection of emotions.

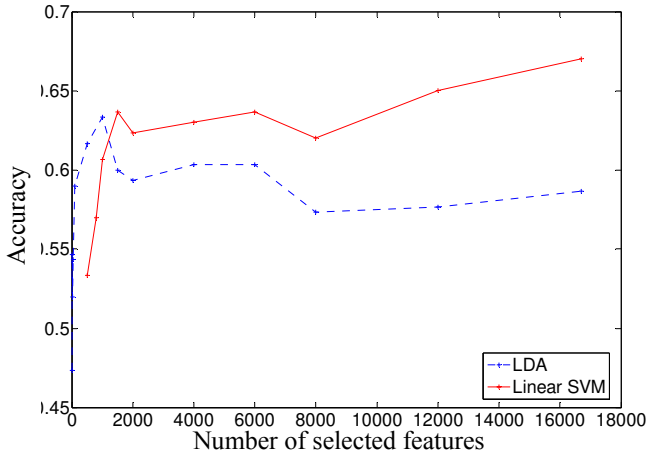


Figure 6.7. Accuracy of LDA and the Linear SVM classifiers for different numbers of selected features of the EEG_STFT feature set using the Fisher criterion (only for the CPN classification scheme). Only the number of features marked with a '+' have been computed while the other values are linearly interpolated.

Figure 6.7 shows that the SVM accuracy is lower than the LDA accuracy in low dimensional spaces (under approximately 1250 selected features) while it is better in high dimensional spaces. The SVM performance in high dimensional spaces can be explained by its intrinsic properties [138]. The good performance of the LDA can be explained by the fact that the diagonalized version of the LDA was employed. Since in this classifier only the variance and the mean of the features are taken into account to determine the linear boundary between the classes (contrarily to the SVM with a linear kernel); the diagonalized LDA is very close to the Fisher criterion employed for feature selection.

The decrease of accuracy observed after the first peak is likely due to the addition of noisy and redundant features. However, after the addition of more than 8000 features the accuracy increases with the number of selected features. This demonstrates that less relevant features (in the sense of the Fisher criterion) can interact with others to improve the accuracy.

As can be seen from Section 6.4.3.a and Section 6.4.3.b, feature selection can improve the accuracy for the LDA with a strong reduction of the number of features (from 0.5% to 6% of the features are kept), providing advantages for real applications where the issues of storage and computation time are important. However, the methods used for feature selection were not able to increase the best accuracy obtained by the SVM classifier. For this reason, feature selection algorithms were not employed for the next steps that are fusion of the modalities and rejection of samples.

6.4.4 Results of fusion

Concerning the fusion by concatenation of the feature vectors, no significant improvement of the accuracy was observed. Most of the time, the classification accuracy was the same as the

accuracy obtained on the EEG_STFT which have the highest number of features. This is likely due to the problem of high dimensionality of this space. While applying feature selection methods after or before fusion can help to solve this issue, this method was not analyzed in view of the results presented in Section 6.4.3. Instead the fusion at the classifier level was performed.

Fusion of classifier decisions is done according to the explanation given in Section 6.3.4. According to the obtained results, fusion was performed choosing probabilistic SVM as the classifiers for EEG features sets (q^{SFFT} and q^{MI}), and the LDA as the classifier for the peripheral feature set (q^{Periph}).

Results from the fusion of MI and STFT EEG features as well as fusion of all EEG and peripheral features are presented in Figure 6.8. As can be seen, combining EEG feature sets increased the best average accuracy by 2% to 4% while combining the three feature sets increased it by 3% to 7% depending on the classification scheme. For instance, the accuracy of the SVM classifier is 63.5% for the CPN classification scheme and reaches 70% after fusion of the three feature sets. In all the present cases combining feature sets leads to an increase in average accuracy, even when fusing modalities with low accuracies such as the peripheral signals. This demonstrates the importance of combining multiple sources of information from both the central and peripheral nervous system in emotion detection from physiological signals.

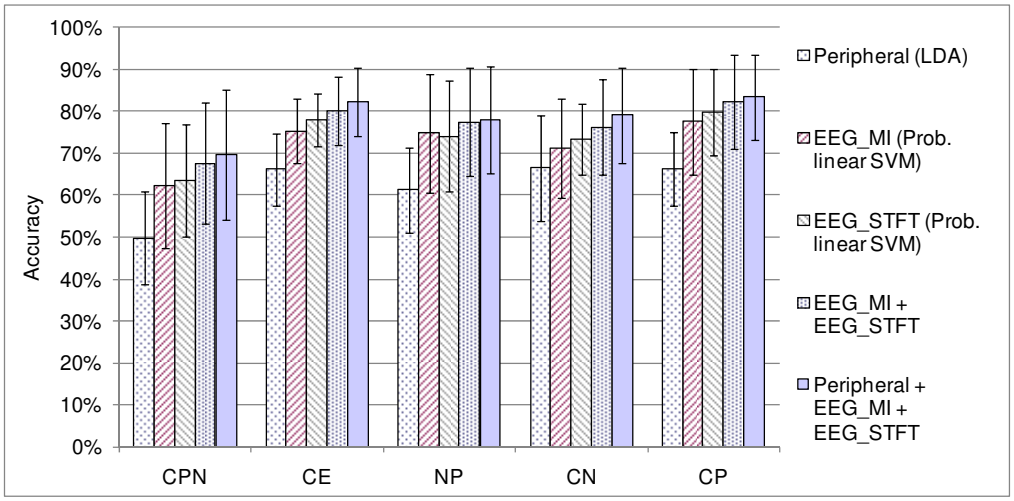


Figure 6.8. Average accuracy across participants for different modalities and their associated classifiers, as well as for fusion of the two EEG and the three physiological modalities.

To our knowledge there only is one study that tried to fuse peripheral and EEG information at the feature level [113]. In this study the authors found that the fusion did not improve accuracy compared to EEG classification. The same fusion was also analyzed in Chapter 5 where an increase was reported only for some classifiers and sets of classes. Since poor results for fusion at the feature level were also obtained in the current study, we believes that for the purpose of

emotion assessment the fusion between the peripheral and the EEG modalities should be operated at the classifier level. This is especially true when the fusion involves feature spaces that are of high dimensionality.

6.4.5 Results of rejection

Finally, samples with low confidence values are rejected using the method described in Section 4.4 and the corresponding increase in accuracy is analyzed in Figure 6.9. In this figure, only the results of the CPN configuration are presented for the trials of all 10 participants (3000 trials) and different values of the δ_{reject} threshold. As can be seen from Figure 6.9, no samples are rejected until δ_{reject} reaches the value of 33%, which is normal since $\max_i g_i$ cannot be inferior to 33% (there is the constraint $\sum_{i=1}^{K} k_i = 1$). The number of rejected samples that are badly classified is higher than the number of correctly classified samples until δ_{reject} becomes higher than 47%. We choose this value to stop rejecting samples since most of the badly classified samples are rejected at this point.

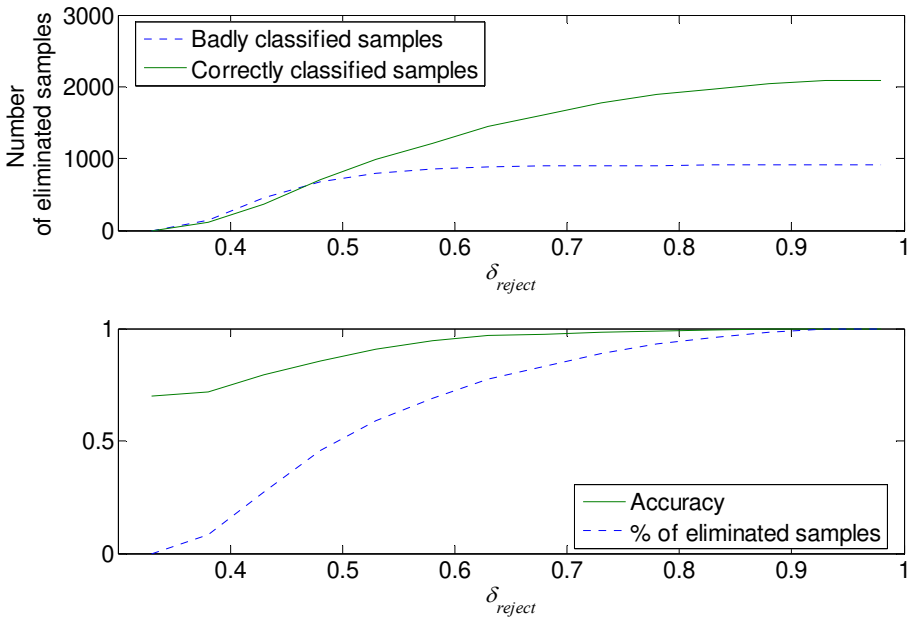


Figure 6.9. Relation between the δ threshold value, classification accuracy and the amount of eliminated samples for the CPN classification task.

This value corresponds to a mean accuracy across participants of 80%, thus increasing it by about 10%. This is to be compared with the 70% accuracy when performing fusion without rejection, but at the cost of rejecting 40% of the samples. Such high rejection rate could seem problematic for a real application, but is however compensated by the short recording period needed to

perform classification and give a decision. For instance if two consecutive trials are rejected, and the third one correctly classified the whole process would still be completed within 25 s.

Similar curves were plotted for the NP and CE classification schemes. The results were quite similar than those obtained from the CPN scheme. Using the same value of 40% for the percentage of rejected samples, the increase of accuracy was respectively of 11% and 10%, resulting in an accuracy of 89% and 92% for the NP and CE schemes. This shows the interest of rejecting samples to improve classification accuracy for other classification schemes than the CPN as well.

6.5 Conclusions

This chapter proposes an approach to classify emotions in the three main areas of the valence-arousal space by using physiological signals from both the PNS and the CNS. A protocol based on the recall of past emotional episodes was designed to acquire short-term emotional data from 11 participants. From the data of 10 participants we extracted three feature sets, one for peripheral signals and two high dimensional feature space for EEG signals. Using the different feature sets, the accuracy of several classifiers was compared on the discrimination of the different combinations of three emotional states. The reduction of the feature space dimensionally was studied for different feature selection algorithms. The fusion of the three feature sets at the classifier level, by combining the probabilistic outputs of classifiers, was analyzed. Finally, rejection of trials where the confidence of the resulting classification is low was performed. In the case the trials with low confidence are those that are misclassified such rejection should lead to an increase of accuracy.

Results showed the importance of EEG signals for emotion assessment by classification as they had better accuracy than peripheral signals on the 8 s of recorded signal. Classification of time-frequency features derived from the EEG signals provided an average accuracy of 63% for three emotional classes and between 73% and 80% for two classes. A new set of features containing the mutual information between each pairs of electrodes was proposed to represent the interaction between different brain areas during emotional processes. The accuracies obtained with this new feature set were only slightly lower than those obtained with the energy features, showing their potential interest for emotion assessment.

Despite of their low accuracy compared to EEG features, the peripheral features were shown to increase accuracy when fused with the EEG modality at the classifier level. Fusion of the two different EEG feature sets also increased the performance of the emotion assessment. This also demonstrates the interest of the new feature set based on mutual information. By fusing the three physiological feature sets the obtained accuracy is of 70% on three classes. Finally, the rejection

of 40% of samples having a low confidence value increased the accuracy to up to 80% on three classes.

The analysis of feature selection methods to reduce the dimensionality of the EEG feature spaces showed that those methods are effective to decrease computational and storage costs without loosing too much accuracy. However, the accuracy obtained after feature selection was always lower than the best accuracy obtained without feature selection.

Since following the stimulus onset emotional processes in brain and peripheral signals are expected to be observable at different times, the exploration of different time resolutions is needed to determine the time scales favorable to emotional assessment from EEG and peripheral activity. For this purpose a protocol where the exact time of the emotion elicitation is known should be designed.

The high number of electrodes used in this study is also an issue since it leads to a high dimensional space where classification is difficult and it forbids the use of this system for real applications. From the analysis of feature selection results, it was shown that a lot of features were correlated, potentially because they were extracted from neighboring electrodes. Grouping the features extracted from close electrodes in order to keep only relevant information and remove part of the noise could also be a solution to reduce the size of feature spaces. The study from Ansari-Asl et al. [129] that tries to select relevant electrodes, based on the data from the same protocol, is also a step toward the reduction of feature spaces sizes.

Analysis of EEG in other elicitation contexts should also be performed to confirm the efficiency of EEG features for emotional assessment in less cognitive tasks, as well as when interacting with computer interfaces. For HCI, the described work can also be used as a guideline to decide which classification strategy to use. Finally, while the rejection of non-reliable trials has been shown to improve accuracy, the percentage of rejected samples is high and further analysis should be conducted to confirm that this rejection can improve the information transfer rate.

Chapter 7 Assessment of emotions for computer games

7.1 Introduction: the flow theory for games

The experiments presented in Chapters 5 and 6 demonstrated the usefulness of physiological signals for emotion assessment. In this chapter similar experiments were performed in a context closer to real HCI applications. This work was done in collaboration with the TECFA (Educational Technologies Laboratory, Faculty of Psychology, University of Geneva).

Due to their capability to present information in an interactive and playful way, computer games have gathered increasing interest as tools for education and training [13]. Games are also interesting from a human-computer interaction (HCI) point of view, because they are an ideal ground for the design of new ways to communicate with machines. Affective computing [2] has opened the path to new types of human-computer interfaces that adapt to affective cues from the user. As one of the main goals of games is to provide emotional experiences such as fun and excitement, affective computing is a promising area of research to enhance game experiences. Affective information can be used to maintain involvement of a player by adapting game difficulty or content to induce particular emotional states. For this purpose, automatic assessment of emotions is mandatory for the game to adapt in real time to the feelings and involvement of the player, without interrupting his / her gaming experience (like it would be the case by using questionnaires). The present work thus focuses on emotion assessment from physiological signals in the context of a computer game application.

Games can elicit a lot of different emotional states but knowing all of them is not necessary to maintain involvement in the game. Many representations of the player's affective state have been used in previous studies like anxiety, frustration, engagement, distress and the valence-arousal space [17, 70]. According to emotion and flow theories [20, 161] strong involvement in a task occurs when the skills of an individual meet the challenge of a task (Figure 7.1). Too much challenge would raise anxiety and not enough would induce boredom. Both these situations would restrain the player's ability to achieve a "flow experience", leading to less involvement, engagement and possibly interruption of the game [162].

In a game, the change from an emotional state to another can occur due to two main reasons. First, the difficulty is increased because of progression in the different levels but too fast compared to the competence increase of the player (potentially giving rise to anxiety, see Figure 7.1). Secondly, the competence of the player has increased while the game remained at the same difficulty (potentially giving rise to boredom). In both cases, the challenge should be corrected to maintain a state of pleasure and involvement, showing the importance of having games that adapt their difficulty according to the competence and emotions of the player. Based on this theory, we

defined three emotional states of interest that corresponds to three well separated areas of the valence-arousal space: boredom (negative-calm), engagement (positive-excited) and anxiety (negative-excited).

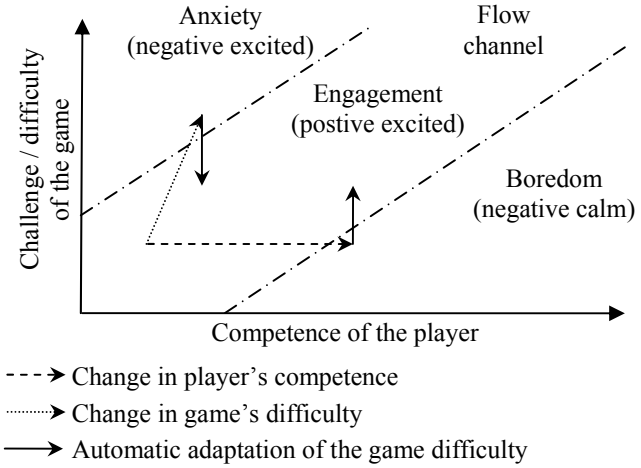


Figure 7.1. Flow chart and the suggested automatic adaptation to emotional reactions.

This work attempts to verify the validity and usefulness of the three defined emotional states by using a Tetris game where the challenge is modulated by changing the level of difficulty. Self-reports as well as physiological activity were obtained from players by using the acquisition protocol described in Section 7.2. Using those data, three analyses were conducted. The first aims at validating the applicability of the flow theory for games (see Section 7.3). In the second analysis, detailed in Section 7.4, physiological signals were used for the purpose of classification of the different states. In this case, since one of the goals of this study is to go toward applications, particular attention was paid to designing classifiers that could be used for any gamer without having to re-train it. The third analysis concerned the analysis of physiological signals after the occurrence of game-over events (see Section 7.5).

7.2 Data acquisition

7.2.1 Acquisition protocol

A gaming protocol was designed for acquiring physiological signals and gathering self-reported data. The Tetris game (Figure 7.2) was chosen in this experiment for the following reasons: it is easy to control the difficulty of the game (speed of falling blocks); it is a widely known game so that we could expect to gather data from players with different skill levels (which occurred, see Figure 7.3); and it is playable using only one hand, which is mandatory since the other hand is used for placement of some data acquisition sensors.

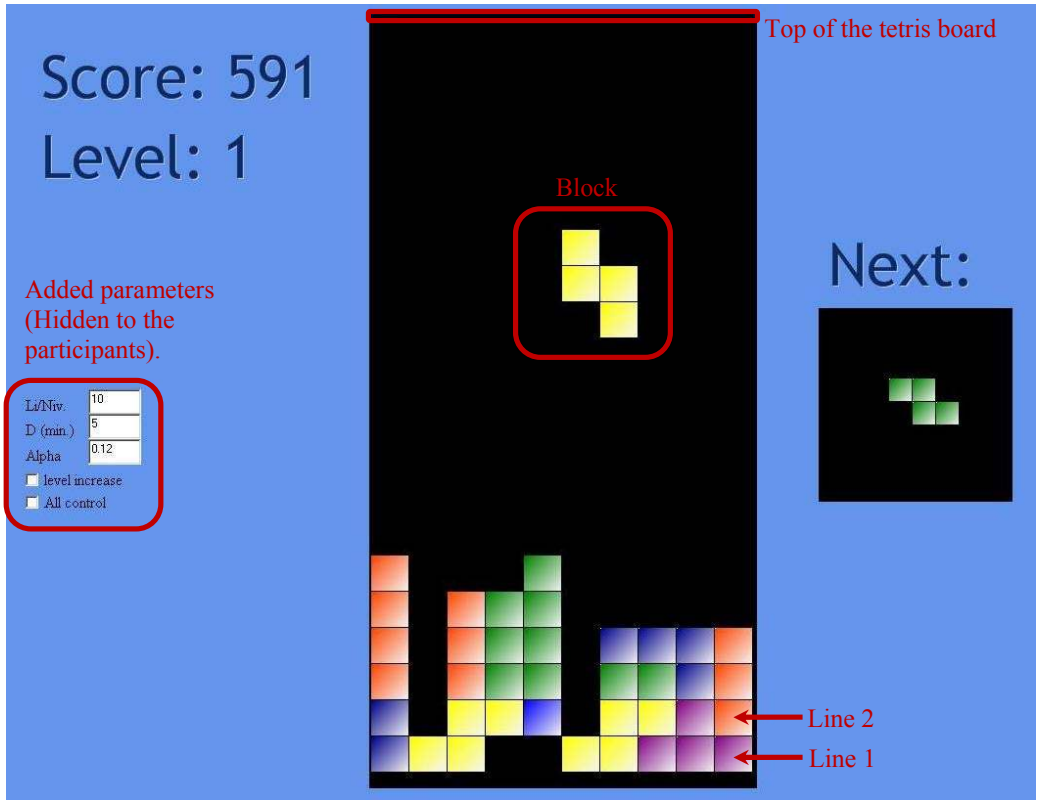


Figure 7.2. Screen shot of the Tetris (DotNETris) game.

The Tetris game used in this experiment is the DotNETris software[9] (Figure 7.2). The difficulty levels implemented in DotNETris were adapted to have a wider range of difficulties than in the original game. The new levels ranged from 1 to 25 in which a block goes down one line every Δt seconds with:

$$\Delta t = e^{-(0.12 \times \text{level} + 0.5)} \tag{7.1}$$

where the exponential function was chosen because a small change in Δt is not really perceivable when the blocks are falling slowly while it is a significant change in difficulty if the blocks are already falling fast. The 0.12 and 0.5 values where chosen empirically to have a "smooth" increase of the difficulty. As a consequence, the blocks were going down a line every 0.54 seconds at level 1 and 0.03 seconds at level 25. Other modifications to the original DotNETris include:

- the possibility to play the game at a given level, without change of a level when 10 lines are eliminated;

[9] available at http://sourceforge.net/projects/dotnetris/ (retrieved on 29 April 2009)

- the command to speed up the fall of the blocks was disabled so that the participants had to wait for the blocks to go down at the chosen speed;
- playing the game for a given duration. Each time the blocks reach the top of the Tetris board (Figure 7.2) during this duration, a game over event was reported, the board was automatically cleared and the participant could continue to play.

20 participants (mean age: 27, 13 males, all right handed) took part in this study. After signing the consent form (Appendix A), each participant played Tetris several times to determine the game level where he/she reported engagement. This was done by repeating three times the threshold method, starting from a low level and progressively increasing it until engagement was reported by the participant or starting from a high level and decreasing it. The average of the obtained levels was then considered as the participant skill level. Depending on this skill level, three experimental conditions were determined: medium condition (game difficulty equal to the player's skill level), easy condition (lower difficulty, computed by subtracting 8 levels of difficulty from the player's skill level), and hard condition (higher difficulty, computed by adding 8 levels). As can be seen from Figure 7.3, the participants to the study reported to be engaged at different levels ranging for most of them from 11 to 16, confirming that they add different Tetris skills. One of the participants was even a "Tetris expert" with a skill level of 20.

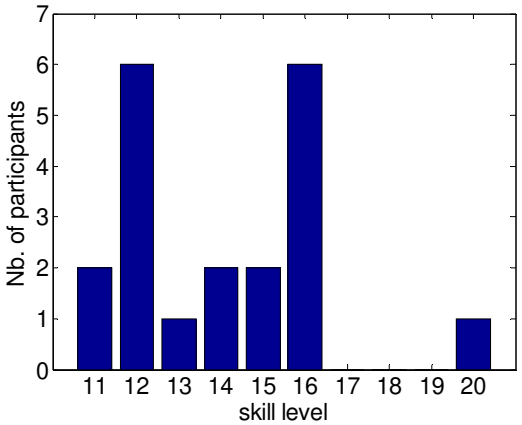


Figure 7.3. Histogram of the skill levels of the 20 participants.

Participants were then equipped with several sensors to measure their peripheral physiological activity: a GSR (Galvanic Skin Response) sensor to measure skin resistance, a plethysmograph to record relative blood pressure, a respiration belt to estimate abdomen extension and a temperature sensor to measure palmar changes in temperature. Those sensors are known to measure signals that are related to particular emotional activations as well as useful for emotion detection (see Section 2.2.2). In addition, an EEG system was used to record central signaling from 14 participants. In this study 19 electrodes were positioned on the skull of participants according to the 10-20 system (see Section 3.1.1). As demonstrated in other studies, EEG's can help to assess

emotional states and is also useful to provide an index of task engagement and workload [130-133]. All signals were recorded at a 1024Hz sampling rate using the Biosemi Active 2 acquisition system. The acquired signals were subsequently downsampled to 256Hz using the built-in Biosemi software [126]. This allows to keep the frequency bands of interest for this study and to speed up computations.

Once equipped with the sensors, the participants took part in 6 consecutive sessions (Figure 7.4). For each session the participants had to follow 3 steps: stay calm and relax for one minute, play the Tetris game for 5 minutes in one of the three experimental conditions (difficulty level) and finally answer a questionnaire. The first step was useful to let the physiological signals return to a baseline level, to record a baseline activity and to provide a rest period to the participants. For the second step, each experimental condition was applied twice and in a random order to account for side effects of time in questionnaires and physiological data. The goal of participants was to perform the highest possible score. To motivate them toward this goal, a prize of 20 CHF was offered to three of the participants having the highest score (the participants were divided in three groups according to their competence). The questionnaire was composed of 30 questions related to both the emotions they felt and their level of involvement in the game (see Appendix D). The answer to each question was given on a 7 points Likert scale. Additionally, participants rated their emotions in the valence-arousal space using SAM [73] scales.

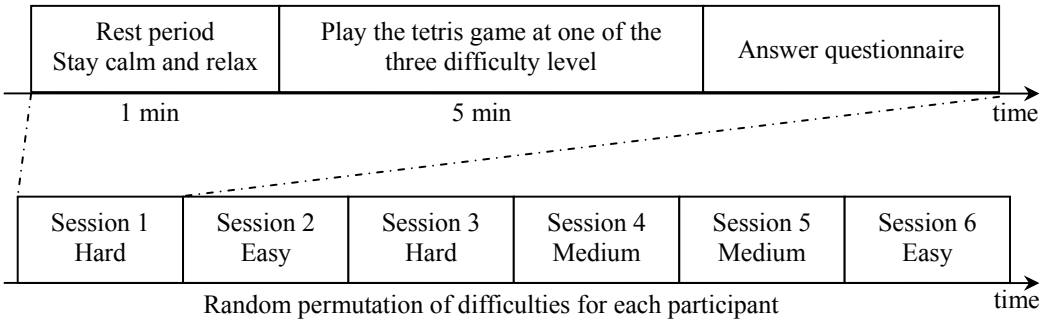


Figure 7.4. Schedule of the protocol.

7.2.2 Feature extraction

From the EEG signals the *EEG_FFT* feature set was computed. This feature set contains for each electrode the EEG signal energy in the θ, α and β bands over the complete duration of a trial. The *EEG_W* feature is also part of the *EEG_FFT* feature set and is related to cognitive processes such as workload and engagement. This last feature is particularly interesting for this protocol since the participants are expected to be more engaged at the medium difficulty than at the two others. The *EEG_FFT* feature set contains a total of $3 \times 19 + 1 = 58$ features (3 frequency bands and 19 electrodes plus the *EEG_W* feature).

Concerning the peripheral activity, features extracted from the corresponding signals are given in Table 7.1. A detailed explanation of the features can be found in Section 3.3.1 for the standard features. The "Reference" column indicates the section of Chapter 3 that corresponds to an explanation for the given advanced features. The peripheral features were computed from the complete duration of the 5 minute trial and concatenated in a feature vector. In this protocol, the duration of a trial was sufficiently long to allow for the computation of peripheral features (i.e. f_{HR}^{LF}, f_{HR}^{HF}, $f_{HR}^{LF/HF}$) that require signals of longer duration than the features computed in the protocols of Chapter 5 and 6. The HR signal was computed from the BVP signal of the plethysmograph as discussed in Section 3.2.4 and a-posteriori correction of the falsely detected beats was applied. The f_{Resp}^{Rate} feature was computed from the respiration spectrum obtained by using the Welch algorithm (Matlab *pwelch* function) with 20 seconds windows and 10 seconds overlap. For a given trial, all the peripheral features were concatenated in a unique feature vector containing a total of 18 features.

Peripheral signal	Standard features			Advanced features	Reference
	μ_x	σ_x	δ_x		
GSR	X		X	$f_{GSR}^{DecRate}$, $f_{GSR}^{DecTime}$, $f_{GSR}^{NbPeaks}$	Section 3.3.2.b
BVP	X	X			
Heart Rate (HR)	X	X	X	f_{HR}^{LF}, f_{HR}^{HF}, $f_{HR}^{LF/HF}$	Section 3.3.2.d
Respiration		X		f_{Resp}^{Rate}, f_{Resp}^{DR}	Section 3.3.2.e
Temperature	X		X		

Table 7.1. The features extracted from the peripheral signals.

In this study the collected data are not analyzed for each participant separately but as a whole. It is thus important to account for inter-participant variability in physiological activity (see Section 3.3). For this reason, the physiological signals acquired during the rest period were used to compute a baseline activity for each session (6 baselines per participant) that was subtracted from the corresponding physiological features. Depending on the feature, the following baseline strategies were applied (see Table 7.2):

- no baseline was subtracted for the features that are already a relative measure of physiological activity such as δ_x and $f_{GSR}^{DecRate}$;
- the last value of the baseline signal was subtracted (L) from some of the μ_x features; in that case a μ_x features represents the signal average change from the end of the baseline;

- the method used to compute the subtracted baseline activity is the same as the one used for the feature computation. For instance the EEG energy in the θ band for the Fz electrode was computed from the baseline signals and subtracted to the corresponding EEG feature.

Feature name	μ_{GSR}	δ_{GSR}	$f_{GSR}^{DecRate}$	$f_{GSR}^{DecTime}$	$f_{GSR}^{NbPeaks}$	μ_{BVP}	σ_{BVP}	μ_{HR}	σ_{HR}	δ_{HR}
Baseline strategy	L	-	-	-	-	L	F	F	F	-
Feature name	f_{HR}^{LF}	f_{HR}^{HF}	$f_{HR}^{LF/HF}$	σ_{Resp}	f_{Resp}^{Rate}	f_{Resp}^{DR}	μ_{Temp}	δ_{Temp}	EEG features	
Baseline strategy	F	F	F	F	F	-	L	-	F	

Table 7.2. The baseline subtraction strategy used for each feature. -: no subtraction of a baseline. L: last value of the baseline signal subtracted. F: the baseline is computed using the same method than for feature computation.

7.3 Analysis of questionnaires and of physiological features

In this section the data gathered from the questionnaires and from the computed physiological features is analyzed to control the applicability of the flow theory for games. For this purpose the validity of the following two hypotheses were tested:

- H1: playing in the three different conditions (difficulty levels) will give rise to different emotional states;
- H2: as the skill increases, the player will switch from an engagement state to a boredom state (see Figure 7.1).

7.3.1 Elicited emotions

a. Questionnaires

To test for hypothesis H1, a factor analysis was performed on the questionnaires to find the axes of maximum variance. The first two components obtained from the factor analysis account for 55.6% of the questionnaire variance and were found to be associated with higher eigenvalues than the other components (the eigenvalues of the first 3 components are 10.2, 8.2 and 1.7). The questionnaire answers given for each session were then projected in the new space formed by the two components and an ANOVA test was applied to those new variables to check for differences in distributions of judgment for the different conditions. By looking at the weights of the two components (see Appendix D) it was found that:

- the **first component** was positively correlated with questions related to **pleasure, amusement, interest and motivation**;
- the **second component** was positively correlated with question corresponding to levels of **excitation and pressure** and negatively correlated with calm and control levels.

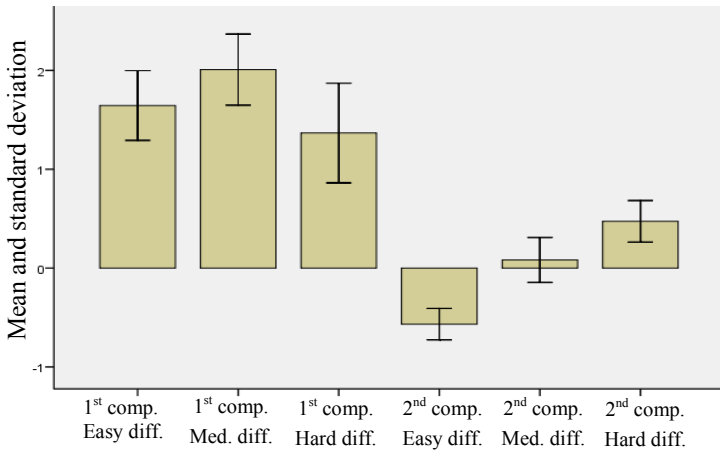


Figure 7.5. Mean and standard deviation of judgments for each axis of the two component (comp.) space and the different difficulties (diff.): easy, medium (med.) and hard.

The ANOVA test, applied on the data projected on the first component (see Figure 7.5), showed that participants felt lower pleasure, amusement, interest and motivation for the easy and hard conditions than for the medium one ($F=46$, $p<0.01$). Differences in the three distributions obtained from the second component demonstrated that increasing difficulty led to higher reported excitation and pressure as well as lower control ($F=232$, $p<0.01$). This demonstrates that an adequate level of difficulty is necessary to engage players in the game so that they feel motivated and pleased to play. Moreover those results also validate hypothesis H1 since they show that the different playing difficulties successfully elicited different emotional states with various levels of pleasure and arousal. According to the self-evaluations those states were defined as boredom for the easy condition, engagement for the medium condition and anxiety for the hard condition.

b. Peripheral features

The physiological features were subjected to an ANOVA test to search for differences in activations for the different conditions and analyze the relevance of those features for emotion assessment. For this purpose the ANOVA test was applied on the three distributions and the F-values and p-values are reported in Table 7.3. Moreover, the ANOVA test was also applied to check for differences between the easy and medium conditions as well as between the medium

and hard condition. If a difference is significant (p-value < 0.1) the trend of the mean from a condition to another is reported in Table 7.3.

Feature	F-value	p-value	Trend of the mean
μ_{GSR}	4.4	0.01	↘→
δ_{GSR}	2.7	0.07	↘→
$f_{GSR}^{DecRate}$	3.1	0.05	↘→
$f_{GSR}^{DecTime}$	6.7	< 0.01	→↗
$f_{GSR}^{NbPeaks}$	18.3	< 0.01	↗→

Feature	F-value	p-value	Trend of the mean
μ_{HR}	3.4	0.04	→↗
f_{HR}^{LF}	2.4	0.09	↘↗
σ_{Resp}	5.8	< 0.01	→↗
μ_{Temp}	9.4	< 0.01	↘↘
δ_{Temp}	10	< 0.01	↘↘

Table 7.3. F-values and p-values of the ANOVA tests applied on the peripheral features for the 3 difficulty levels. Only the relevant features are presented (p-value < 0.1). The "Trend of the mean" column indicates the differences between two conditions. For instance ↘↘ indicate a significant decrease of the variable from the easy to the medium condition (first ↘) and from the medium to the hard condition (second ↘), while →↗ indicate no significant differences between the easy and medium condition and a significant increase to the hard condition.

The decrease observed for the μ_{GSR}, δ_{GSR}, $f_{GSR}^{DecRate}$ features and the increase of the $f_{GSR}^{NbPeaks}$ between the easy and medium conditions indicate an increase of EDA when progressing from the easy to the medium difficulty level. Between the easy and medium conditions a significant decrease of temperature is also observed. Those results are in favor of an increase of arousal between the easy and the medium condition. When analyzing the GSR features changes between the medium condition and the hard condition, only the $f_{GSR}^{DecTime}$ feature (percentage of negative samples in the GSR derivative) is significantly increasing. An increase of mean HR and a decrease of temperature are also observed between the same conditions. Those results suggest that there is also an increase of arousal between the medium and hard conditions but to a lesser extent than between the easy and medium conditions. In summary, an increased arousal is observed for increasing game difficulty, supporting the results obtained from the analysis of the questionnaires.

As can be seen from Table 7.3 a total of ten features were found to have significantly different distributions among the three difficulties. This suggests that the conditions correspond to different emotional states and demonstrates the interest of those features for later classification of the three conditions. One feature of particular interest is f_{HR}^{LF}, the HR energy in low frequency bands, because it has a lower value for the medium condition than for the two others, showing that this condition can elicit particular peripheral activation. This is also one of the only features that can help to distinguish the medium condition from the two others.

c. EEG features

An ANOVA test was also performed on each EEG feature to test for differences between the three conditions. Table 7.4 gives a list of the EEG features that are relevant (p-value < 0.1). However, to illustrate the EEG activity we focused on the *EEG_W* feature since it is a combination of the other features and is known to be related to cognitive processes such as engagement and workload [130, 131, 133].

	Left electrodes	Midline electrodes	Right electrodes
Theta band	C3, T7, P3, P7, O1	Fz, Cz	F4, C4, T8, O2
Beta band	Fp1, P7, O1	Cz	C4, T8, P8, O2

Table 7.4. List of the relevant EEG features (p-value < 0.1) given by frequency band and electrode.

Significant differences were observed for the *EEG_W* feature between the three conditions ($F=5.5$, $p<0.01$). Figure 7.6 shows the median and quartiles of the *EEG_W* values for each condition. Since for the medium difficulty the participants reported higher interest and motivation than for the easy and hard difficulty, it was expected that the mean of the *EEG_W* values would be significantly higher for the medium condition. However, as can be seen from Figure 7.6, there is increase in the median of the *EEG_W* values as the difficulty increases. The differences between the medium and hard conditions as well as between the easy and hard conditions are significant according to the ANOVA test. In our view this reflects the fact that the *EEG_W* feature is more related to workload than to engagement. The participants involved more executive functions in the hard difficulty than the medium one, even if they were less engaged.

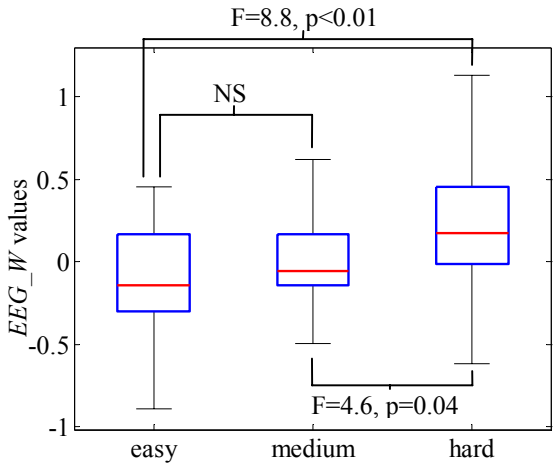


Figure 7.6. Boxplot of the EEG_W values for the three condition. The red line represent the median of the EEG_W values, the box the quartile and the whiskers the range. NS: non significant.

7.3.2 Evolution of emotions in engaged trials

Hypothesis H2 was tested by focusing on the data of the two sessions corresponding to the medium condition where the participant is expected to be engaged. Both physiological and

questionnaire data were analyzed using a pairwise t-test to verify that there was a decrease of engagement from the first to the second session.

The pairwise t-test on the variables of the questionnaire showed a significant decrease from the first to the second medium condition for the questions "I had pleasure to play" (t=-1.8, p=0.09) and "I had to adapt to the interface" (t=-3, p=0.06). From peripheral signals, a decrease in the number of GSR peaks $f_{GSR}^{NbPeaks}$ (t=-2.4, p=0.02) as well as an increase in the average of temperature μ_{Temp} (t=2.6, p=0.02) and average of temperature derivative δ_{Temp} (t=2.3, p=0.03) was found.

Those results are indicative of a decrease of arousal and pleasure while playing twice in the same condition, thus supporting hypothesis H2. The result obtained for the question "I had to adapt to the interface" gives a cue that this decrease could be due to an increase of player's competence. However the competence changes were not measured with other indicators to confirm this possibility. In any case, those results demonstrate the importance of having automatic adaptation of game's difficulty when the challenge of the game remains the same.

7.4 Classification of the gaming conditions using physiological signals

7.4.1 Classification methods

As shown in Section 7.3 several features computed from both the peripheral and central signals were found to significantly differ between the three gaming conditions. Moreover the participants reported to be in a different emotional state for each of these conditions. In this section, the next step is to investigate in more details the classification accuracy that can be expected from emotion assessment in gaming conditions. For this purpose classification methods were applied on the data gathered from the gaming protocol. The ground-truth labels were defined as the three gaming conditions, each one being associated to one of three states: boredom (easy condition), engagement (medium condition) and anxiety (hard condition).

Four classifiers were applied on this data set: a DLDA, a DQDA, a linear SVM and a SVM with RBF kernel. The diagonalized versions of the LDA and the QDA were employed because of the low number of samples, which sometimes gives rise to the problem of singular covariance matrices. The participant cross-validation method proposed in Section 4.1.2 was used to compute the accuracy of the classifiers. For each participant a classifier was trained using features of the other participants; accuracy was then computed by applying the trained model on the physiological data of the tested participant. The gamma parameter of the RBF SVM was chosen to maximize accuracy on the training set (see Section 4.1.3.c). Since the classifier is tested on the data of participants that are not present in the training set, this method allows evaluating the

performance of the classifier in the worst case where the model is not user-specific, i.e. no information about the specificity of the user's physiology is required for emotion assessment, except for a baseline recording of 1 min.

Three feature selection algorithms were applied on this problem to find the features that are relevant for classification and provide good generalization across participants. All those algorithms were applied on the training set to select features of interest and only the selected features were used for classification of the test set. The ANOVA feature selection was applied to keep only those that are relevant to the class concept. Only the features having a p-value below 0.1 were kept. The FCBF algorithm was applied to select relevant features and remove redundant ones. The δ_{FCBF} parameter was set to 0.2 because (*i*) it was shown in Chapter 6 that this value is relevant for FCBF EEG features selection and (*ii*) the number of features that have a correlation with the classes higher than 0.2 (7 for peripheral features and 23 for EEG features) is similar to the number of relevant features found using the ANOVA test (10 for peripheral features and 20 for EEG features, see Section 7.3.1). Finally, the SFFS algorithm (see Section 4.2.2) was also used to select features of interest, including potentially interacting features. To search for features that have good generalization across participants, the accuracy of a feature subset was estimated by computing the participant cross-validation accuracy on the training set. The maximum size F_{SFFS} of a feature subset was set to 18 for peripheral features and 20 for EEG features. We limited the maximum size of the EEG feature set to 20 because the SFFS algorithm is computationally expensive and 20 features were found to be relevant according to the ANOVA test.

Since the EEG signals were recorded only for 14 out of the 20 participants, the available number of samples for EEG based classification is not the same as for peripheral based classification. For this reason the results obtained from EEG and peripheral features are separated in two sections with classification algorithm applied on 14 participants for EEG and 20 participants for peripheral features. In Section 7.4.4 the classification accuracies obtained with EEG and peripheral features on different time scales are compared while the fusion of peripheral and EEG modalities is investigated in Section 7.4.5. In both cases, the classification accuracy was computed only on the 14 participants having EEG recorded.

7.4.2 Peripheral signals

Figure 7.7 presents the accuracies obtained by applying the classification methods on the features extracted from the peripheral signals. The result obtained for the linear SVM is omitted for the SFFS. When using the SFFS algorithm to search for the first best feature, the computations could not be completed, this presumably was caused by convergence problems or by an error in the libSVM toolbox implementation.

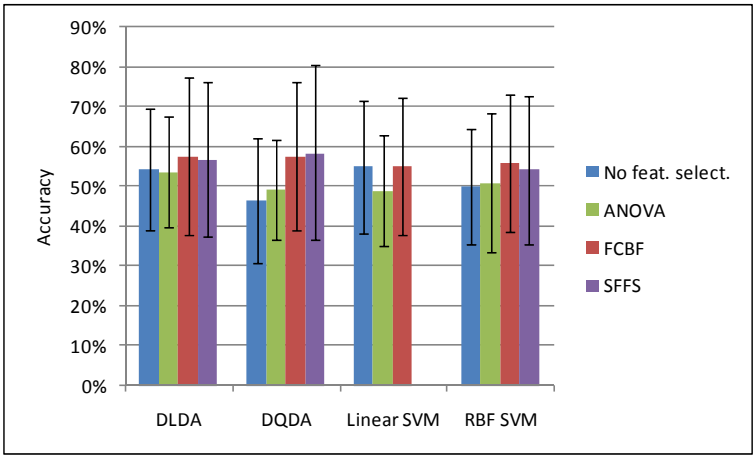


Figure 7.7. Accuracies of the different classifiers and feature selection metods on the peripheral features.

Without feature selection the linear classifiers obtained the best accuracies of 55% for the linear SVM and 54% for the DLDA showing their ability to find a boundary that generalizes well across participants. In any case, the accuracies are higher than the random level of 33%. Except for the ANOVA, the feature selection methods always improved the classification accuracies. The best accuracy of 59% is obtained with the DQDA combined with SFFS feature selection. However the FCBF results (58%) are not significantly different from those obtained with the SFFS algorithm because of the high variance of the accuracies. Moreover, the variance of the accuracies obtained with SFFS tends to be higher than those obtained with the FCBF which shows that the FCBF is more stable than the SFFS algorithm in selecting the proper features. According to the results and considering that the FCBF is much faster than the SFFS, the FCBF can be considered as the best feature selection algorithm for this classification scheme.

Since the participant cross-validation method was used, the feature selection algorithms were applied 20 times on different training sets. For this reason, the features selected at each iteration of the cross-validation procedure can be different. The histograms of Figure 7.8 show for each feature the number of times it was selected by a given feature selection algorithm. The average number of selected features is 3.5 for the FCBF, 9.35 for the ANOVA feature selection and 4.8 for the SFFS. The ANOVA nearly always selected the features that were found to be relevant in Section 7.3.1.b but with poor resulting accuracy (Figure 7.7). Thanks to the removal of redundant features, the FCBF strongly reduces the original size of the feature space with a good resulting accuracy. Moreover this algorithm nearly always selected the same features independently of the training set showing its stability. The SFFS also obtained good performance but as can be seen from Figure 7.8, some of the features were selected only on some of the training sets, showing that this algorithm is less stable than the FCBF.

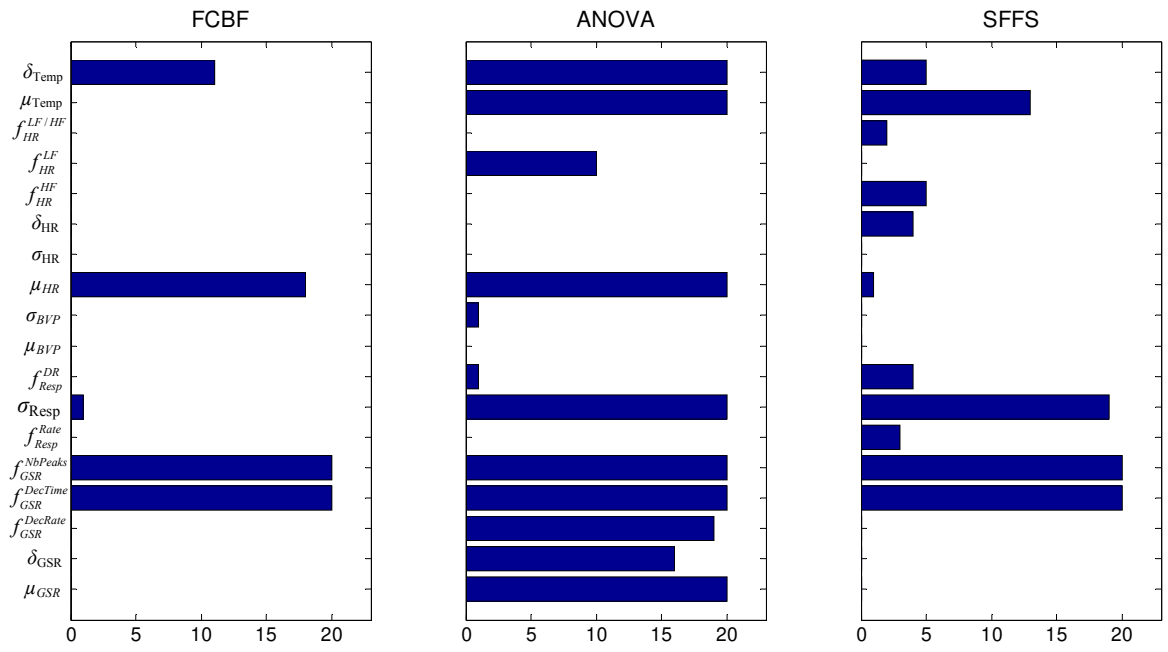


Figure 7.8. Histograms of the number of cross-validation iterations (over a total of 20) in which features have been selected by the FCBF, ANOVA and SFFS feature selection algorithms. The SFFS feature selection is displayed for the DQDA classification.

By inspecting the SFFS, FCBF and ANOVA selected features, the $f_{GSR}^{DecTime}$ and $f_{GSR}^{NbPeaks}$ features were always selected which shows their importance for classification of the three conditions from physiological signals. To our knowledge similar features have been used only in [109] for emotions assessment despite of their apparent relevance. The μ_{HR} feature was frequently selected by the FCBF but never by the SFFS and vice-versa for the σ_{Resp} feature. The σ_{Resp} feature was removed by the FCBF because it was correlated with μ_{HR}. However the SFFS kept the σ_{Resp} feature based on its predictive accuracy which suggests that this feature may be better than μ_{HR} for classification. Finally, the temperature features were also found to be frequently relevant.

Because of its good accuracy and low computational time the FCBF algorithm coupled with DQDA classification was used for further analyses involving the peripheral modality. Table 7.5 presents the confusion matrix for the 3 classes: it can be seen that the boredom condition was well classified, followed by the anxiety condition. Samples from the engagement condition tend to be classified mostly as bored samples and also as anxious samples. This is not surprising since this condition lies in between the others. Notice that 21% of the samples belonging to the anxiety class are classified as bored samples; this can be due to fact that some participants completely disengaged from the task because of its difficulty, reaching an emotional state close to boredom. In this case, the adaptive game we propose would increase the level of difficulty since the

detected emotion would be boredom, which is not the proper decision to take. A solution to correct this problem could be to use contextual information such as the current level of difficulty and the direction of the last change in difficulty (i.e. increase or decrease) to correctly determine the action to take.

Estimated / True	Easy (Boredom)	Medium (Engagement)	Hard (Anxiety)
Easy (Boredom)	80%	10%	10%
Medium (Engag.)	37%	33%	30%
Hard (Anxiety)	21%	19%	60%

Table 7.5. Confusion matrix for the DQDA classifier with FCBF feature selection.

7.4.3 EEG signals

The accuracies obtained for classification of the EEG features with the different classifiers and feature selection methods are displayed in Figure 7.9. Linear SVM results combined with the SFFS procedure are not displayed because of the problem described in Section 7.4.2.

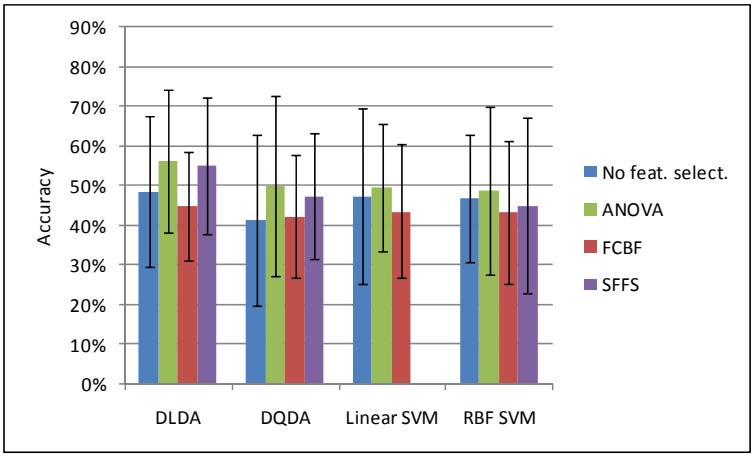


Figure 7.9. Accuracies of the different classifiers and feature selection metods on the EEG features.

All the classification methods obtained accuracy higher than the random level of 33%. Without feature selection the DLDA had the best accuracy of 49%, followed by the linear SVM classifier with 47.5% of accuracy and the RBF SVM with 47%. As with the peripheral features, these results demonstrate the ability of linear and support vector classifiers to well generalize across the participants. The best result of 56% was obtained by the DLDA coupled with ANOVA feature selection. The ANOVA feature selection method always had a better performance than the other methods. To our knowledge these are the first results concerning the identification of gaming conditions from EEG signals, especially considering that the classifiers were trained using a cross-participant framework.

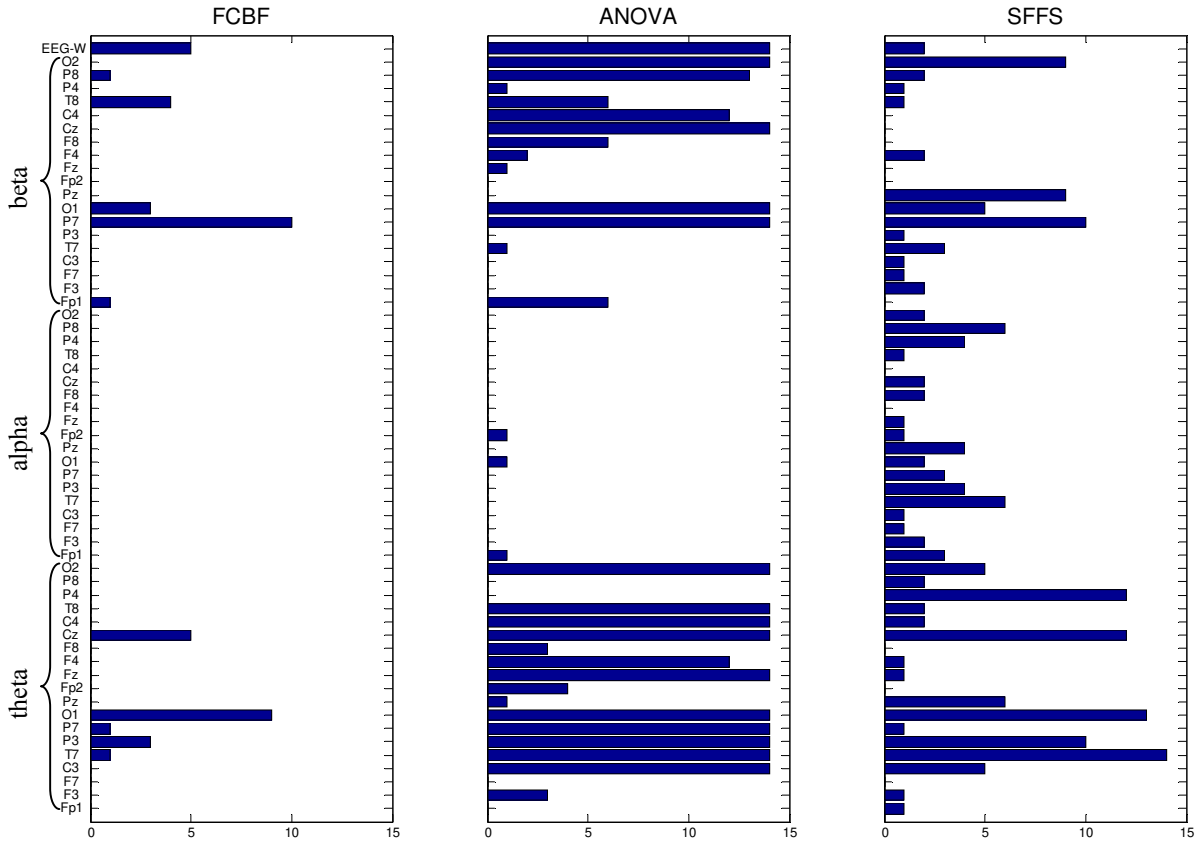


Figure 7.10. Histograms of the number of cross-validation iterations (over a total of 14) in which features have been selected by the FCBF, ANOVA and SFFS feature selection algorithms. The SFFS feature selection is displayed for the DLDA classification.

As can be seen from Figure 7.10, the FCBF selected less features than the two other feature selection methods. It selected 3.1 features in average compared to 20.3 for the ANOVA and 13.0 for the SFFS coupled with the DLDA. This explains the low accuracy obtained with the FCBF and shows that good accuracies on this problem can be obtained only by concatenating several features. The ANOVA algorithm often selected the features described in Section 7.3.1.c. The SFFS coupled with the DLDA had accuracies close to those of the ANOVA with DLDA but by selecting less features in average. For this reason the features selected by this method are of particular importance for accurate classification of the three gaming conditions. The more often selected features (selected more than 8 times) were the theta band energies of the T7, O1, Cz, P4 and P3 electrodes and the beta band energies of the P7, Pz and O2 electrodes. This result shows that the occipital and parietal lobes were particularly useful for differentiation of the three gaming conditions.

The confusion matrix displayed in Table 7.6 for the DLDA and FCBF methods shows that the different classes were detected with similar accuracies. The medium condition still has the lowest accuracy but is better detected than when using the peripheral features. On the other hand, the

easy condition is detected with less accuracy than with peripheral features. This indicates that the fusion of the two modalities should increase the overall accuracy.

Estimated / True	Easy (Boredom)	Medium (Engagement)	Hard (Anxiety)
Easy (Boredom)	57%	43%	0%
Medium (Engag.)	21%	50%	29%
Hard (Anxiety)	19%	19%	62%

Table 7.6. Confusion matrix for the DLDA classifier with ANOVA feature selection.

7.4.4 EEG and peripheral signals

In order to compare accuracies obtained using either EEG or peripheral signals, the best combinations of classifiers and feature selection methods were applied on the physiological database with the same number of participants for both modalities (the 14 participants for whom EEG was recorded). Moreover, the comparison was conducted for different time scales to analyze the performance of each modality as a function of the signal duration used for the features computation. For this purpose, each session (see Figure 7.4) was divided into 1 to 10 non-overlapping windows of $300/W$ seconds, where W is the number of windows and 300 seconds the duration of a session. An EEG and peripheral feature vector was then computed from each window and the label of the session was attributed to this feature vector. By using this method, a database of physiological features was constructed for each window size ranging from 30 to 300 seconds.

For a database in which the features were computed from W windows, the number of samples for each class is $20 \times 2 \times W$ (20 participants, 2 sessions per class and W windows per session). Thus the number of samples per class increases with W. Since the number of samples can influence classification accuracy and the goal of this study is to analyze the performance of EEG and peripheral features at different time scales, it is important that this comparison be conducted with the same number of samples for each window's length. To satisfy this constraint one sample was chosen randomly from each session using a uniform distribution to have $20 \times 2 = 40$ samples per class. The classification algorithms were than applied on this reduced database. This was repeated 1000 times for each value of W to account for the different possible combinations of the windows (except for W=1). Notice that it is not possible to perform classification for all windows combinations since there are W^{40} such combinations.

By using this method the average accuracies over the 1000 iterations are displayed in Figure 7.11. The small accuracy oscillations that can be observed for small time windows (less than 100 seconds) are likely due to the increase of the number of possible combinations of windows. As can be seen from Figure 7.11 the accuracy obtained for the peripheral signals with the original duration of the sessions (300 seconds) is not significantly different from the one obtained with all

of the 20 of participants (See Section 7.4.2). Thus having 13 or 19 participants for classifiers training (because of participant cross-validation) does not significantly change the classification performance. This suggests that adding more participants to the current database would not increase classification accuracies and that recording 14 to 20 participants is enough to obtain reliable accuracy estimations.

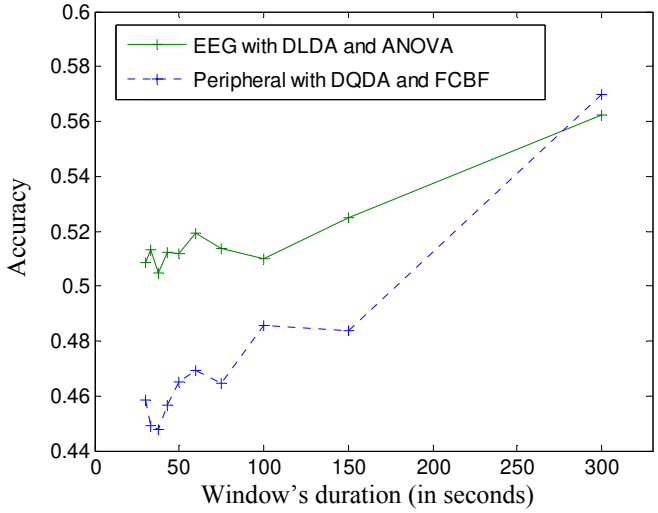


Figure 7.11. Classification accuracy as a function of the duration of a trial for EEG and peripheral features.

For both modalities, decreasing the duration of the window on which the features are computed leads to a decrease of accuracy. However, this decrease is stronger for peripheral features than for EEG features. For the EEG features, the accuracy drops from 56% for windows of 300 seconds to around 51% for windows of 30-50 seconds. For the peripheral features the accuracy is 57% for windows of 300 seconds and around 45% for windows of 30-50 seconds. Moreover, the EEG accuracy remains approximately the same for windows having duration inferior to 100 seconds while the peripheral accuracy continues to decrease. All those results demonstrate that the EEG features are more robust on short term assessment than the peripheral features. For our application, adapting the difficulty of the Tetris game based on the physiological signals gathered during the 5 precedent minutes may be undesirable since there is a high probability that the difficulty of the game has changed during this laps of time due to usual game progress. Having modalities, like EEG, that are able to estimate the state of the user on shorter time periods is thus of great interest.

7.4.5 Fusion

Fusion at the feature level (see Section 4.3.1) was performed and the different classifiers and feature selection methods were applied on the resulting feature sets. The results did not show any significant improvement of the classification accuracy. Similar results were obtained when performing fusion at the classifier level using the sum rule (see Section 4.3.2.a) and the best

classifiers for each feature set. This is due to the fact that the posterior probabilities outputed by the classifiers were generally higher for EEG features than for peripheral features so that the class estimated from the EEG features was often selected by the fusion.

As can be seen from the confusion matrices obtained from the classification based on the peripheral and EEG features (Table 7.5 and Table 7.6), the errors made in with these two feature sets are quite different. As explained in Section 4.3.2.b the Bayes belief integration is well suited for this type of problem, and thus was employed for fusion of the best classifiers found for each feature set (the DLDA couples with ANOVA for EEG features and DQDA coupled with FCBF for peripheral features). Another advantage of the Bayes belief integration is that the probabilities $P(\omega_i \mid \hat{y}_q)$ can be estimated independently for the two classifiers. It was thus possible to use the training data of 19 participants to compute probabilities for the peripheral features while only 13 participants were used for the EEG features. The resulting accuracy and confusion matrices were obtained by using the participant cross-validation applied on the 14 participants for whom both EEG and peripheral activity were recorded.

Estimated / True	Easy (Boredom)	Medium (Engagement)	Hard (Anxiety)
Easy (Boredom)	82%	14%	4%
Medium (Engag.)	29%	39%	32%
Hard (Anxiety)	4%	27%	69%

Table 7.7. Confusion matrix for the "Bayes belief integration" fusion.

The accuracy obtained after fusion was 63% which corresponds to an increase of 5% compared to the best accuracy obtained with the peripheral features. Table 7.7 presents the confusion matrix obtained after fusion. By comparing this table to Table 7.5 and Table 7.6 it can be observed that the detection accuracy of the easy and the hard classes was increased by 2% and 7% respectively compared to the accuracy obtained with the best feature set (peripheral features for the easy class and EEG features for the hard class). The accuracy obtained on the medium class with fusion (39%) is lower than the one obtained with EEG features (50%) but higher than with peripheral features (33%). When performing classification based either on EEG or peripheral features many of the hard samples were classified as easy while this problem was solved after fusion. All these results demonstrate the interest of peripheral and EEG fusion for a more accurate detection of the three conditions.

7.5 Analysis of game-over events

Analysis of the physiological signals should also be conducted on the basis of the events in the game and not only for the complete 5 min of a session since change of emotional states can also occur during any of the sessions. For this purpose the analysis of physiological signals after a

game-over event is performed in this section. Notice that since the game restarted automatically after each game-over event, the changes in physiological activity observed after these events can be due either to the loss of a game or to the start of the new one.

7.5.1 Method

In this section, the variations of the BVP, HR and GSR signals after game-over events were analyzed. Those signals were chosen to represent the peripheral activity because they have fast reaction time and they are known to be indicators of arousal and mental effort [7, 163]. Since the HR signal is not sampled at regular interval and with the same sampling rate as other signals, it was first interpolated using a cubic interpolation. All the signals were then segmented into windows of 5 seconds, each window starting at one of the game-over triggers recorded during the acquisition. Each window was also associated to the label of the corresponding session in order to distinguish between peripheral reactions of the different difficulty conditions.

During the easy sessions, none of the participants reached the top of the Tetris board (Figure 7.2) which shows that the difficulty level was appropriately chosen as easy but prevents any analysis of game-over events for this condition. For the two other conditions, the game-over events were present for all participants and occurred on average each 98 seconds for the medium condition and each 14 seconds for the hard condition. As a result, the number of samples obtained for the hard condition is much higher than the one of the medium condition. Due to the high frequency of game-over events in the hard condition, it is possible that an event fell in the window of the precedent event (two game-overs within 5 seconds). In that case, the window associated to the first event was rejected.

Within a given window, each signal is normalized by subtracting the amplitude of the first sample from all samples. Once all the signals were segmented and normalized, differences between the conditions were investigated. For this purpose an ANOVA test was applied every 0.5 second on the values of each type of signal.

7.5.2 Results

Figure 7.12 shows the HR and GSR averages over the windows of each experimental condition. No significant differences between the medium and hard conditions were found for the BVP signal. The GSR values were significantly different from 3 seconds after the game-over trigger with a higher value for the medium condition than for the hard condition. The most significant difference ($F=5.3$, $p=0.02$) was obtained at the fourth second. For the HR signal, a significant difference between the conditions was found between 1.5 to 3 seconds after the game-over trigger with a higher heart rate for the medium than for the hard condition. The most significant difference ($F=6.3$, $p=0.01$) occurred 2.5 seconds after game-over events.

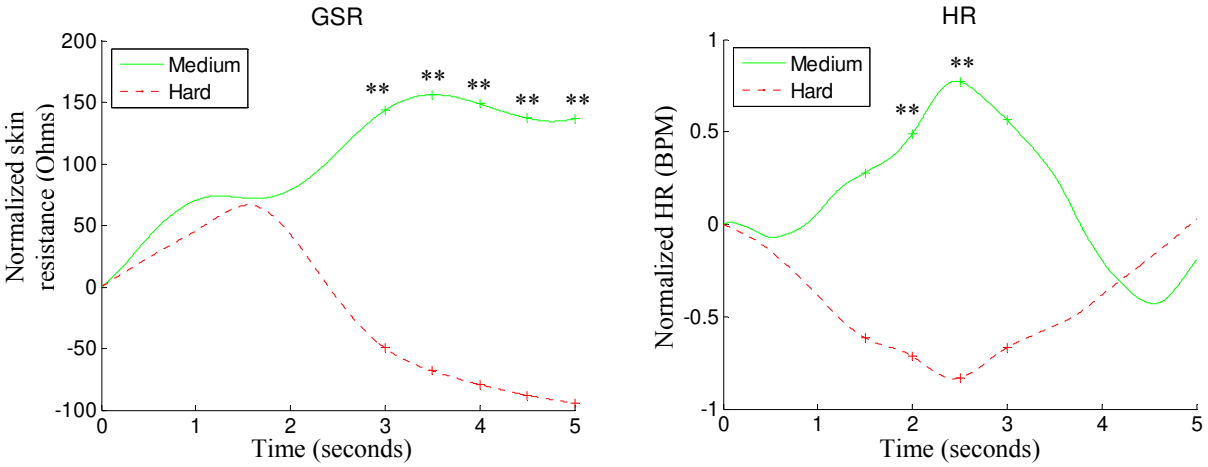


Figure 7.12.Averages of the normalized GSR and HR signals for the 5 seconds following the game-over triggers. Points that are marked with a '+' corresponds to the samples that were found to be significantly different (p-value < 0.1) among the two conditions. '' indicate that p-value < 0.05.**

The results obtained from the GSR signal indicate a higher EDA for the hard condition than for the medium condition which shows that the sympathetic activity was higher in the hard condition. Since from the questionnaire the hard condition was found to be related to higher excitation and pressure, a potential interpretation of this result is that the participants were more aroused and stressed because they had to start a new game that they know to be too hard relatively to their competences. The HR responses are significantly different only for a short period of time after the trigger. For this reason, those responses were assumed to be related to the game-over event and not to the new game. Higher HR was reported for unpleasant stimuli compared to pleasant stimuli in [62]. Since the participant reported higher pleasantness and amusement in the medium condition the difference in HR could be due to the deception of loosing a game in the medium condition.

Unfortunately, more variables should be gathered to confirm the interpretations given in the precedent paragraph, especially self-reports concerning the game-over events. Nevertheless, results demonstrate that there are different patterns of peripheral activity after game-over events between the sessions where the participants reported higher motivation and pleasantness and those were they reported high pressure and less motivation. This suggests that this activity could be used to distinguish engaged from stressed states in games where such events occurs frequently. Examples of such games are the first-person shooter games where the character driven by the gamer is frequently "killed". Further studies are needed to investigate if those responses are also present for other games than the modified Tetris used in this protocol.

7.6 Conclusion

This study investigated the possible use of emotion assessment from physiological signals to adapt the difficulty of a game. A protocol was designed to record physiological activity and gather self-reports of 20 participants playing a Tetris game at three different levels of difficulty. The difficulty levels were determined according to the competence of the players on the task. Three types of analysis were conducted on the data: statistical analysis of self-reports and physiological data was performed to control that different cognitive and emotional states were elicited by the protocol, classification was conducted to determine whether it is possible to detect those states from physiological signals, and an analysis of the changes in physiological activity after game-over events was performed.

The results obtained from the analysis of self-reports and physiological data showed that playing the Tetris game at different levels of difficulty gave rise to different emotional states. The easy difficulty was related to a state of low pleasure, low pressure, low arousal and low motivation which was determined as boredom. The medium difficulty elicited higher arousal than the easy difficulty, as well as higher pleasure, higher motivation and higher amusement. It was thus defined as engagement. Finally the hard difficulty was associated to anxiety since it elicited high arousal, high pressure and low pleasure. Moreover, the analysis of consecutive engaged trials showed that the engagement of a player can decrease if the game difficulty does not change. These results demonstrate the importance of adapting the game difficulty according to the emotions of the player in order to maintain his / her engagement.

The classification accuracy of EEG and peripheral signals to recover the three states elicited by the gaming conditions were analyzed for different classifiers, feature selection methods and durations on which the features were computed. Without feature selection the best classifiers obtained an accuracy around 55% for peripheral features and 48% for EEG features. The FCBF increased the best accuracy on the peripheral feature to 59% while the ANOVA selection increased the accuracy to 56% for EEG features. The analysis of the classification accuracy for EEG and peripheral features computed on different durations demonstrated that the EEG features are more robust to a decrease in duration than the peripheral features, which confirms the importance of EEG features for short term emotion assessment.

From the analysis of the game-over events, distinct patterns of GSR and HR activity for the medium and hard conditions were found in the 5 seconds following the event. The distinct patterns were supposedly due to the differences in arousal and pleasantness relative to the game-over event and to the starting of a new game. Nevertheless, those distinct peripheral patterns suggest that peripheral signals recorded after events occurring during the game could be used to determine the state of the player.

Future work will focus on the improvement of the detection accuracy. Fusion of physiological information with other modalities such as facial expressions and speech would certainly improve the accuracy. Including game information such as the evolution of the score can also help to better detect the three states. Another question of interest is to determine the number of classes to be detected. Since boredom and anxiety are detected with higher confidence than engagement it might be enough to use those two classes for adaptation to the game difficulty. Moreover, from the observation of Figure 7.1, one can conclude that it is more interesting to adapt the difficulty of the game solely based on the increase of competence because it leads to a stronger change of state in the flow chart (Figure 7.1) and stimulates learning. In this case only the detection of boredom is of importance to modulate difficulty. This also implies to more clearly define what are the relations between emotions, competence and learning. A future study would be to implement an adaptive Tetris game and verify that it is more fun and enjoyable than the standard one. Finally, analysis of physiological signals for different types of games is also required to see if the results of this study can be extended to other games.

Chapter 8 Conclusions

8.1 Outcomes

This thesis concentrated on the study of physiological signals in the context of affective computing. It aimed at demonstrating and comparing the usefulness of two categories of physiological signals for emotion assessment: those that reflect the activity of the central nervous system (EEG's) and those reflecting the peripheral nervous system activity (GSR, temperature, BVP, HR and respiration).

Chapter 2 presented a state of the art of the topics related to emotion assessment from physiological signals. Firstly, the most well-known representations and models of emotions were given together with their implications for emotion assessment studies. The valence-arousal space was used as the representation of emotions throughout this thesis since it was considered as being the most general, flexible and less dependent on the application. The role of context and the multimodal aspects of emotions were also emphasized from the analysis of the described models of emotions. Secondly, the physiological processes related to emotions were described for both the peripheral and central activities. A non exhaustive list of the devices usable to record those activities was given and the signals features that are known to be related to emotional processes were discussed. Thirdly, several studies concerning the assessment of emotions from physiological signals were reviewed. Six criterions were proposed to organize, evaluate and compare those studies as well as to discuss important aspects of physiological emotion assessment.

Chapter 3 started by a description of the material used to monitor the physiological activity of several participants during emotional experiences. While this part is important for the reproducibility of the results, it also provides a guide to help persons that are not used to this type of apparatus for setting up an acquisition process. The algorithms to pre-process the signals and extract the features that characterize the physiological activity were then presented. For EEG's, features based on the energy of the signals were computed according to the literature and a new a feature set based on the MI between pairs of electrodes was proposed. An algorithm to compute HR from a BVP signal based on the detection of the foot of the systolic upstroke was detailed. The proposed peripheral features were explained and discussed according to the literature. This chapter concluded on the ethical aspects that should be (and have been) considered when acquiring physiological signals.

Chapter 4 presented the supervised learning methods used to assess emotions from the extracted physiological features. It stressed the importance of the ground-truth definition and provided different possibilities to determine the (supposedly) true emotional state elicited by an event. The

accuracy and the confusion matrix obtained on test data, by using the defined cross-validation strategies, were chosen to measure the performance of the emotion assessment. The classifiers (Naïve-Bayes, QDA, LDA, SVM and RVM) used to find models that map the physiological responses to an estimated emotional state were given. Since some of the extracted feature sets were of high dimensionality, filter (ANOVA, Fisher based and FCBF) and wrapper (SFFS) feature selection algorithms were detailed to later select the features of interest. The last part of Chapter 4 presented the information fusion methods. These methods were used in the next chapters to investigate the usefulness of fusion between the peripheral and central information.

Chapter 5, 6 and 7 applied the previously described methods on physiological and emotional data acquired in different elicitation contexts (elicitation method, emotions elicited and duration of trial). In Chapter 5 the emotions were elicited by using images from the IAPS and the classifiers were independently trained to recover valence or arousal classes. In Chapter 6, the participants had to self-generate emotions by remembering past emotional episodes of their life belonging to three areas of the valence-arousal space. Chapter 7 described a gaming protocol that aimed at testing emotion assessment in a context that is closer to HCI applications. The main conclusions and outcomes drawn from those chapters are given below.

One of the main outcomes of this thesis is the conclusion that EEG's are useful for emotion assessment. For all the protocols and classes formulations, the average classification accuracy was higher than the random level. Moreover, compared to the peripheral features, the classification based on EEG features generally led a higher accuracy for the assessment of the valence dimension of emotions. The experiments performed in chapter 6 and 7 also demonstrated that the EEG modality was more adequate than the peripheral modality to assess emotions on a short time scale.

Fusion of peripheral and EEG features was shown to be effective as it increased the classification accuracy, especially for the data presented in Chapter 6 and 7. The fusion of the classifiers outputs was always more effective than simple concatenation of the feature vectors. The MI feature set proposed in Chapter 3, was less effective than the energy feature set. However when the two feature sets were fused, an increase of the accuracy was observed showing the interest of the MI features for emotion assessment.

Applying the feature selection algorithms to reduce the dimensionality of the extracted features led to a strong decrease of the size of the feature sets with only a reasonable decrease (Chapter 6) or increase (Chapter 6 and 7) of the accuracy depending on the employed classifiers and datasets. This is of interest to improve the computational speed of the classification algorithms. Moreover, these algorithms assisted in finding the EEG and peripheral features that are useful for inter-participant classification as detailed in Chapter 7.

Concerning the classification algorithms, none of the classifiers systematically performed better than the others. However, the analysis conducted in Chapter 6 confirmed the interest of the SVM classifiers for classification of emotions in high dimensional feature spaces.

An important outcome of this thesis is the production of three emotional databases containing the peripheral and EEG signals of several participants, acquired in different emotional elicitation contexts. While those databases are not publicly available because of ethical aspects, it remains that further studies can be performed on this data within the University of Geneva.

This thesis has shown the usability of EEG and peripheral signals for emotion assessment and therefore encourages their use for affective computing. An example of an affective Tetris game that adapts its difficulty to the emotions felt by the user is given in Chapter 7. The results obtained from the analysis of physiological signals gathered in Chapter 7 showed that these signals are useful for the game adaptation. Many other applications in areas such as health and information retrieval can be targeted by this research.

8.2 Future prospects

As demonstrated in this thesis, physiological signals can be used to assess emotions. However, there are still some research issues that have to be investigated in order to improve the accuracy of the assessment and apply it to concrete HCI applications.

Emotions can be elicited in several contexts influencing the emotional expression. It is thus important that the methods developed to assess emotions take into account those contextual elements. Since new human-computer interfaces will more and more involve all of the human senses it can be interesting to analyze the performance of emotion assessment algorithms according to the sense used for the emotion elicitation. Analyzing the combinations of such stimuli can also be of interest. The emotional model proposed by Ortony [57] is a possible direction to follow in order to better determine an emotional state according to the course of events that elicited the emotion. Emotion assessment from physiological signals (or from other sources) can then be added to this model to add personal emotional information. The mood of the user and the persistence of the precedent emotional state are also context related issues that can influence the elicitation of an emotion and thus should be taken into account.

All the experiments conducted in this thesis were done in an “ideal” environment where the participants were instructed to accomplish given tasks and to avoid movements. Switching from this type of experiment to the real environments gives rise to several issues. Physiological signals are very sensitive to movements, for instance if a user stands up his / her blood pressure will change. Moreover, the user can be disturbed by external events that are not related to the

application of interest. It is thus very important to develop algorithms that are able to detect signals changes that are not related to emotional processes.

While this work focused on the identification of emotional classes defined as areas of the valence-arousal space, going further toward the identification of a point in this space is of high interest to determine emotional states with higher precision. This could be useful to infer the intensity of the emotion, generally defined as the distance of the point to the center of the space. Having enough resolution in this space is also mandatory to map a valence-arousal estimation to a given emotional label. Assessing the dominance (or control) dimension of emotions can be useful to well differentiate emotional states like fear and anger. Continuous estimations of emotional points in those spaces can be done by using supervised regression algorithm. In that case, since the valence and arousal variables seem to be dependent, the regression should be performed accordingly. Going forward to other continuous representations of emotions, like the SEC proposed in Scherer's [10] model is still an opportunity but requires the evaluation of many continuous variables.

As demonstrated in this thesis, EEG signals can be used for emotion assessment. However, a lot of effort should be put on the design of new EEG caps that are less obtrusive to go toward applications. For parts of this thesis, 64 electrodes were employed to monitor brain activity. This high number of electrodes is problematic regarding prices aspects and leads to high dimensional feature spaces. Developing algorithms that are able to select the electrode positions of interest for emotion assessment is of major importance to solve those issues. A possible solution to this problem could be to use the MI computed between pairs of electrodes (as proposed in Chapter 3) to regroup electrodes that recorded similar information and choose one of them as the representative of the group.

Increasing the accuracy obtained from EEG features is of major importance for practical use of this device. For this purpose, new features should be investigated possibly inspired from the BCI community like features based on common spatial patterns. The MI feature set used with success in this study encourages the investigation of interactions between brain areas during emotional processes. The synchronization of brain processes could be used to determine new features. Some of the brain structures involved in emotional processes lie deep in the brain and it is thus difficult to assess their activity from surface EEG signals. Solving the inverse problem (i.e. finding the brain sources corresponding to a given EEG) to estimate deep sources and using this information as new features for emotion assessment could be promising.

Fusion with other sensors and sources of emotional information could lead to improvements of the emotion assessment accuracy. Several sensors can be used to acquire signals originating from the same sources. For instance, EEG measurements can be coupled with fNIRS measurements to

better estimate brain activity. However, since emotions are multi-modal processes that involve several component of the organism it is certainly most valuable to perform the fusion of different sources of information. For instance, this could be achieved by combining facial expression and speech identification with physiological measurements of emotions. Those fusions would necessitate more studies, especially concerning time aspects. The time resolution of different sensors is not the same and the different components involved in emotional processes do not have the same reaction time.

Most of the studies concerning emotion assessment from physiological signals (including this thesis) are done on emotional data that are not available to the whole research community. As a consequence it is difficult to compare the methods used for emotion assessment since their performances strongly depend on the protocol used for data acquisition. There is thus an important need for databases that are freely accessible. Such databases should ideally be multimodal, include contextual information and meet the constraints imposed by the law / ethical rules. A freely available multimodal database[10] of emotionally driven brain and peripheral signals was constructed in collaboration with partners of the European Network of excellence Similar. Unfortunately, this data was not analyzed in this thesis but we strongly encourage the use of this database for further research on the topic. A similar effort is now underway in the context of the EU project Petamedia.

Taken together, the above suggestions should lead to the development of a robust emotion assessment system. Once such a system is developed the next step would be to determine how the machine should adapt to the user's emotional state. Some propositions were given in Chapter 7 for computer games but this strategy is highly dependent on the gaming application. Finally, the investigation of how the user perceives the complete system (i.e. emotion assessment and adaptation) is mandatory to control how it would be received by the general public.

[10] available at http://enterface.tel.fer.hr/index.php?frame=results (retrieved on 19 May 2009)

Appendix A Consent form

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

Battelle bâtiment A
7, Route de Drize,
1227 Carouge

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES

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

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

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

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

Déroulement des expériences

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

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

Respect de la sphère privée, conservation des données

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

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

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

Conditions d'arrêt de l'expérience

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

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

Informations supplémentaires

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

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

Expérimentateurs actuels: Guillaume Chanel (Guillaume.chanel@cui.unige.ch), Mohammad Soleymani, mohammad.soleymani@cui.unige.ch, Joep Kierkels, Joep.Kierkels@cui.unige.ch.

Acquisition de signaux physiologiques

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

Acquisition de signaux électro-encéphalographiques (EEG)

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

Déroulement des expériences d'enregistrement des EEGs

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

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

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

Acquisition d'autres types de signaux physiologiques

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

Questionnaire d'expérience

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

Coordonnées

Nom et prénom(s)

Adresse

Numéro postal/Ville

Adresse email

Téléphone

Renseignements généraux

Vous êtes un/une ☐ homme ☐ femme

Votre date de naissance (jour/mois/année)

Renseignements influençant les signaux EEG

Main prédominante: ☐ gaucher ☐ droitier ☐ ambidextre

Formulaire de consentement

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

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

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

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

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

Appendix B Neighborhood table for the Laplacian filter

The neighborhood of an electrode was defined as proposed in [164]. The following table gives for each electrode the list of the associated neighbors.

Electrode	Neighbors electrodes	Electrode	Neighbors electrodes
Fp1	F7, F5, AF7, AFz, Fpz	Fpz	Fp1, AFz, Fp2
AF7	Fp1, F5, F3, AF3, AFz	Fp2	Fpz, AFz, AF8, F6, F8
AF3	AFz, AF7, F3, F1, Fz, AF4	AF8	Fp2, AFz, AF4, F4, F6
F1	F3, FC3, FC1, Fz, AF3	AF4	AF8, AFz, AF3, Fz, F2, F4
F3	F5, FC5, FC3, F1, AF3, AF7	AFz	Fpz, Fp1, AF7, AF3, AF4, AF8, Fp2
F5	F7, FT7, FC5, F3, AF7, Fp1	Fz	AF3, F1, FC1, FCz, FC2, F2, AF4
F7	FT7, F5, Fp1	F2	AF4, Fz, FC2, FC4, F4
FT7	T7, FC5, F5, F7	F4	F6, AF8, AF4, F2, FC4, FC6
FC5	FT7, T7, C5, FC3, F3, F5	F6	F8, Fp2, AF8, F4, FC6, FT8
FC3	C5, C3, C1, FC1, F1, F3, FC5	F8	Fp2, F6, FT8
FC1	F1, FC3, C1, FCz, Fz	FT8	F8, F6, FC6, T8
C1	FCz, FC1, FC3, C3, CP1, Cz	FC6	FT8, F6, F4, FC4, C6, T8
C3	C1, FC3, C5, CP3, CP1	FC4	FC6, F4, F2, FC2, C2, C4, C6
C5	CP5, CP3, C3, FC3, FC5, T7	FC2	F2, Fz, FCz, C2, FC4
T7	TP7, CP5, C5, FC5, FT7	FCz	Fz, FC1, C1, Cz, C2, FC2
TP7	P9, P7, CP5, T7	Cz	FCz, C1, CP1, CPz, CP2, C2
CP5	TP7, P7, P5, CP3, C5, T7	C2	FC4, FC2, FCz, Cz, CP2, C4
CP3	CP5, P5, P3, CP1, C3, C5	C4	C6, FC4, C2, CP2, CP4
CP1	Cz, C1, C3, CP3, P3, Pz, CPz	C6	T8, FC6, FC4, C4, CP4, CP6
P1	POz, P2, Pz, P3, P5, PO3	T8	FT8, FC6, C6, CP6, TP8
P3	P5, P1, Pz, CP1, CP3	TP8	T8, CP6, P8, P10
P5	PO3, P1, P3, CP3, CP5, P7	CP6	TP8, T8, C6, CP4, P6, P8
P7	PO7, PO3, P5, CP5, TP7, P9	CP4	CP6, C6, C4, CP2, P4, P6
P9	O1, PO7, P7, TP7	CP2	C4, C2, Cz, CPz, Pz, P4, CP4
PO7	Iz, Oz, PO3, P7, P9, O1	P2	PO4, P6, P4, Pz, P1, POz
PO3	Oz, POz, P1, P5, P7, PO7	P4	P6, CP4, CP2, Pz, P2
O1	Iz, PO7, P9	P6	P8, CP6, CP4, P4, P2, PO4
Iz	O2, PO8, Oz, PO7, O1	P8	TP8, CP6, P6, PO4, PO8, P10
Oz	Iz, PO8, PO4, POz, PO3, PO7	P10	TP8, P8, PO8, O2
POz	Oz, PO4, P2, P1, PO3	PO8	O2, P10, P8, PO4, Oz, Iz
Pz	P2, P4, CP2, CPz, CP1, P3, P1	PO4	P8, P6, P2, POz, Oz, PO8
CPz	Cz, CP1, Pz, CP2	O2	P10, PO8, Iz

Appendix C List of IAPS images used

Liste of IAPS images for the **arousal** experiment

IAPS image number	Associated class	IAPS image number	Associated class	IAPS image number	Associated class
1050	High arousal	4659	High arousal	7050	Low arousal
1201	High arousal	4800	High arousal	7060	Low arousal
1300	High arousal	5130	Low arousal	7080	Low arousal
1310	High arousal	5390	Low arousal	7090	Low arousal
1931	High arousal	5470	High arousal	7100	Low arousal
2190	Low arousal	5500	Low arousal	7110	Low arousal
2381	Low arousal	5510	Low arousal	7140	Low arousal
2440	Low arousal	5520	Low arousal	7150	Low arousal
2480	Low arousal	5530	Low arousal	7175	Low arousal
2570	Low arousal	5621	High arousal	7185	Low arousal
2580	Low arousal	5623	High arousal	7187	Low arousal
2620	Low arousal	5626	High arousal	7205	Low arousal
2661	High arousal	5700	High arousal	7217	Low arousal
2691	High arousal	5731	Low arousal	7224	Low arousal
2840	Low arousal	5740	Low arousal	7233	Low arousal
2850	Low arousal	5910	High arousal	7234	Low arousal
2870	Low arousal	5920	High arousal	7235	Low arousal
2880	Low arousal	5940	High arousal	7380	High arousal
2890	Low arousal	5950	High arousal	7490	Low arousal
3030	High arousal	6570	High arousal	7491	Low arousal
3053	High arousal	7000	Low arousal	7640	High arousal
3071	High arousal	7004	Low arousal	7950	Low arousal
3080	High arousal	7006	Low arousal	8030	High arousal
3150	High arousal	7010	Low arousal	8080	High arousal
3170	High arousal	7020	Low arousal	8160	High arousal
3261	High arousal	7025	Low arousal	8161	High arousal
4220	High arousal	7031	Low arousal	8170	High arousal
4290	High arousal	7035	Low arousal	8200	High arousal
4658	High arousal	7040	Low arousal	8300	High arousal
8400	High arousal	9050	High arousal	9600	High arousal
8490	High arousal	9360	Low arousal	9622	High arousal
8500	High arousal	9405	High arousal	9810	High arousal
8501	High arousal	9570	High arousal		
9040	High arousal	9571	High arousal		

Appendix C

Liste of IAPS images for the **valence** experiment

IAPS image number	Associated class	IAPS image number	Associated class	IAPS image number	Associated class
1710	Positive	5621	Positive	8501	Positive
1811	Positive	5623	Positive	8502	Positive
2053	Negative	5629	Positive	8503	Positive
2160	Positive	5660	Positive	8531	Positive
2710	Negative	5700	Positive	8540	Positive
2800	Negative	5910	Positive	9006	Negative
3051	Negative	6212	Negative	9050	Negative
3060	Negative	6230	Negative	9140	Negative
3063	Negative	6300	Negative	9181	Negative
3064	Negative	6360	Negative	9252	Negative
3100	Negative	6830	Negative	9253	Negative
3110	Negative	6831	Negative	9300	Negative
3120	Negative	7230	Positive	9340	Negative
3130	Negative	7260	Positive	9400	Negative
3180	Negative	7270	Positive	9405	Negative
3220	Negative	7330	Positive	9421	Negative
3230	Negative	7380	Negative	9433	Negative
3400	Negative	7502	Positive	9500	Negative
4220	Positive	8030	Positive	9520	Negative
4599	Positive	8034	Positive	9560	Negative
4607	Positive	8080	Positive	9570	Negative
4608	Positive	8090	Positive	9571	Negative
4610	Positive	8170	Positive	9600	Negative
4640	Positive	8180	Positive	9611	Negative
4641	Positive	8190	Positive	9620	Negative
4660	Positive	8200	Positive	9630	Negative
4680	Positive	8210	Positive	9800	Negative
5260	Positive	8230	Negative	9810	Negative
5270	Positive	8350	Positive	9910	Negative
5450	Positive	8370	Positive	9911	Negative
5460	Positive	8380	Positive	9920	Negative
5470	Positive	8400	Positive	9921	Negative
5480	Positive	8420	Positive		
5600	Positive	8496	Positive		

The images that are common to the arousal and valence experiment are: 4220, 5470, 5621, 5623, 5700, 5910, 7380, 8030, 8080, 8170, 8200, 8400, 8501, 9050, 9405, 9570, 9571, 9600, 9810.

Appendix D Questionnaire results for the game protocol

The following table gives the list of the 30 statements that the participants had to evaluate using Likert scales. The participants also had to evaluate their emotion using the valence, arousal and dominance SAM scales.The weights higher than 0.7 are highlighted in gray.

	Statement	1st component	2nd component
Q1	J'ai apprécié le jeu	0.90	-0.09
Q2	J'ai été intéressé(e)	0.90	0.13
Q3	J'ai essayé de nouvelles commandes	0.34	-0.03
Q4	J'ai été motivé(e) à faire le meilleur score possible	0.74	-0.17
Q5	J'ai eu du plaisir à jouer	0.90	-0.17
Q6	J'ai dû m'adapter aux commandes	0.35	0.41
Q7	J'ai trouvé la partie difficile	-0.12	0.87
Q8	J'ai été absorbé(e) par le jeu	0.60	0.56
Q9	J'ai pris en compte d'avantage d'information qu'auparavant	0.55	-0.16
Q10	J'ai dû réfléchir	0.75	0.00
Q11	J'ai été stressé(e)	0.06	0.88
Q12	Mon attention s'est focalisée sur la partie	0.53	0.50
Q13	J'ai eu l'impression que je m'améliorais	0.66	0.24
Q14	J'aurais pu obtenir un meilleur score	0.54	0.00
Q15	J'ai été amusé(e)	0.86	0.07
Q16	Je n'ai pas vu le temps passer	0.50	0.45
Q17	J'ai joué au maximum de mes capacités	0.36	0.48
Q18	J'ai oublié ce qui se passait autour de moi	0.32	0.44
Q19	J'ai trouvé la partie trop facile	0.11	-0.84
Q20	J'ai ressenti de l'ennui	-0.46	-0.21
Q21	J'ai senti que j'avais le contrôle	0.40	-0.78
Q22	J'ai trouvé la partie plaisante	0.88	-0.10
Q23	J'ai été calme	0.19	-0.67
Q24	J'aurais volontiers continué à jouer	0.86	-0.11
Q25	J'ai été excité(e)	0.31	0.70
Q26	J'ai su clairement ce que je devais faire	0.35	-0.49
Q27	J'ai pu diriger les pièces comme je le voulais	0.34	-0.82
Q28	J'ai été concentré(e) totalement sur le jeu	0.60	0.46
Q29	J'ai été mis(e) sous pression	-0.02	0.81
Q30	J'ai pensé à d'autres choses qu'au jeu	-0.12	-0.45
Q31	valence (SAM)	0.73	0.01
Q32	arousal (SAM)	0.17	0.77
Q33	dominance (SAM)	0.37	-0.59

Appendix E Publications

Refereed journals, special issues

G. Chanel, J.J.M. Kierkels, M. Soleymani, T. Pun, "Short-term emotion assessment in a recall paradigm", *International Journal of Human-Computer Studies*. Accepted for publication, March 2009, DOI:10.1016/j.ijhcs.2009.03.005

M. Soleymani, G. Chanel, J.J.M. Kierkels, T. Pun, "Affective characterization of movie scenes based on content analysis and physiological changes", *International Journal on Semantic Computing*; submitted, Feb. 2009.

J. Kronegg, G. Chanel, S. Voloshynovskiy, T. Pun, "EEG-based synchronized brain-computer interfaces: a model for optimizing the number of mental tasks", *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol. 15, No. 1, March 2007, pp. 50-58.

T. Pun, T. I. Alecu, G. Chanel, J. Kronegg, S. Voloshynovskiy, "Brain-computer interaction research at the Computer Vision and Multimedia Laboratory, University of Geneva", *IEEE Transactions Neural Systems and Rehabilitation Engineering*, Special Issue on Brain-Computer Interaction, T. M. Vaughan and J. R. Wolpaw, Eds., Vol. 14, No. 2, June 2006, pp. 210-213.

International conferences with refereed full length articles

G. Chanel, C. Rebetez, M. Betrancourt, T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games", *11th MindTrek Conference, MindTrek 2008: Entertainment and Media in the Ubiquitous Era*, Tampere, Finland, October 7-9, 2008.

G. Chanel, K. Ansari-Asl, T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm", 2007 IEEE SMC, *International Conference on Systems, Man and Cybernetics, Smart cooperative systems and cybernetics: advancing knowledge and security for humanity*, Montreal, Canada, Oct. 7-10, 2007.

G. Chanel. J. Kronegg, D. Grandjean, I. Alecu, T. Pun, "Pattern recognition in peripheral and central signaling", *Workshop on Multimodal synchronization in affective expressions*, Humaine European NOE 3rd Summer School and Affective Sciences NCCR, Genova, Italy, September 22-24, 2006. Invited long presentation.

G. Chanel, J. Kronegg, D. Grandjean, T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals", Proceedings of the *International Workshop on Multimedia Content Representation, Classification and Security (MRCS)*, Special Session: Multimodal Signal Processing, Istanbul, Turkey, Sept. 11-13, 2006, B. Gunsel, A. K. Jain, A. M. Tekalp, B. Sankur, Eds., Lecture Notes in Computer Science, Vol. 4105, Springer, 530-537.

Appendix E

M. Soleymani, J.J.M. Kierkels, G. Chanel, T. Pun, "A Bayesian framework for video affective representation", ACII 2009, *International Conference on Affective Computing and Intelligent Interaction*, Amsterdam, The Netherlands, Sept. 10-12, 2009. Subm. April. 2009.

M. Soleymani, G. Chanel, J.J.M. Kierkels, T. Pun, "Affective characterization of movie scenes based on multimedia content analysis and user's physiological emotional responses", ISM2008, *IEEE International Symposium on Multimedia*, Berkeley, California, USA, December 15-17, 2008, 228-235.

M. Soleymani, G. Chanel, J.J.M. Kierkels, T. Pun, "Affective ranking of movie scenes using physiological signals and content analysis", ACM Multimedia 2008, *Workshop on Many Faces of Multimedia Semantics (MS'08)*, Vancouver, BC, Canada, Oct. 27-31, 2008, pp. 32-39.

M. Soleymani, G. Chanel, J.J.J. Kierkels, T. Pun, "Valence-arousal representation of movie scenes based on multimedia content analysis and user's physiological emotional responses", *MLMI 2008, 5th Joint Workshop on Machine Learning and Multimodal Interaction*, Utrecht, The Netherlands, 8-10 Sept. 2008, (extended abstract and poster).

K. Ansari-Asl, G. Chanel, T. Pun, "A channel selection method for EEG classification in emotion assessment based on synchronization likelihood", *Eusipco 2007, 15th European Signal Processing Conference*, Poznan, Poland, Sept. 3-7, 2007.

A. Benoit, L. Bonnaud, A. Caplier, P. Ngo, L. Lawson, D. Trevisan, V. Levacik, C. Mancas, and G. Chanel, "Multimodal Focus Attention and Stress Detection and Feedback in an Augmented Driver Simulator", *3rd IFIP Conference on Artificial Intelligence Applications and Innovations (AIAI)*, Athens, June 7-9, 2006.

A. Savran, K. Ciftci, G. Chanel, J. Cruz Mota, L. Hong Viet, B. Sankur, L. Akarun, A. Caplier and M. Rombaut, "Emotion Detection in the Loop from Brain Signals and Facial Images", *Proceedings of eNTERFACE 2006 Workshop*, Dubrovnik, Croatia, July - August 2006.

J. Kronegg, T. Alecu, G. Chanel, S. Voloshynovskiy, T. Pun, "Analyse des mesures de débit pour interfaces cerveau-ordinateur", *TAIMA'05, Traitement et Analyse de l'Information : Méthodes et Applications*, Hammamet, Tunisie, 26 sep - 1 oct, 2005.

Other conferences

G. Chanel, J. Kierkels, M. Soleymani, D. Grandjean, T. Pun, "Short-term emotion assessment in a recall paradigm", *Joint (IM)2-Interactive Multimodal Information Management and Affective Sciences NCCRs Workshop*, Riederalp, Switzerland, Sept. 1-3, 2008.

G. Chanel, S. Delplanque, "Olfactory-elicited emotions assessment through psychophysiological signals: statistical and classification approaches", *Joint (IM)2-Interactive Multimodal Information Management and Affective Sciences NCCRs Workshop*, Riederalp, Switzerland, Sept. 1-3, 2008.

G. Chanel, C. Rebetez, M. Betrancourt, T. Pun, "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games", *Joint (IM)2-Interactive Multimodal Information Management and Affective Sciences NCCRs Workshop*, Riederalp, Switzerland, Sept. 1-3, 2008.

G. Chanel, K. Ansari-Asl, T. Pun, "From thoughts to emotions: emotional state assessment using physiological recordings", *HUMABIO EU Project Workshop, Securing infrastructures and enhancing safety in critical operations; Humabio physiological and behavioural biometrics for unobtrusive authentication and monitoring*, Basel, Switzerland, February 2, 2007.

M. Soleymani, G. Chanel, J. Kierkels, T. Pun, "Valence-arousal representation of movie scenes based on multimedia content analysis and user's physiological emotional responses", *2008 PetaMedia Workshop on Implicit, Human-Centered Tagging (HCT)*, Queen Mary University London, London, UK, Sept. 5, 2008.

M. Soleymani, J. Kierkels, G. Chanel, E. Bruno, S. Marchand-Maillet, T. Pun, "Estimating emotions and tracking interest during movie watching based on multimedia content and physiological responses", *Joint (IM)2-Interactive Multimodal Information Management and Affective Sciences NCCRs Workshop*, Riederalp, Switzerland, Sept. 1-3, 2008.

K. Ansari-Asl, G. Chanel, T. Pun, "A Channel selection method for EEG classification in emotion assessment based on synchronization likelihood", *Similar NOE Workshop*, University of Magdeburg, Germany, June 4-5, 2007.

J. Kronegg, G. Chanel, S. Voloshynovskiy, T. Pun, "Information-transfer rate based performance optimization for brain-computer interfaces", *Similar NOE Workshop*, Heraklion, Greece, June 8-9, 2006.

T. Pun, T. I. Alecu, G. Chanel, J. Kronegg, S. Voloshynovskiy, "Research in Brain-computer interaction, Multimodal Interaction Group, Computer Vision and Multimedia Laboratory, University of Geneva", *BCI 2005, Brain-Computer Interface Technology: Third International Meeting*, Rensselaerville, NY, USA, June 14-19, 2005.

References

[1] D. Goleman, *Emotional Intelligence: Why It Can Matter More Than IQ*: Bantam, 1997.

[2] R. W. Picard, *Affective computing*: The MIT Press, 1997.

[3] S. Brave, C. Nass, and K. Hutchinson, "Computers that care: investigating the effects of orientation of emotion exhibited by an embodied computer agent," *International Journal of Human Computer Studies*, vol. 62, No. 2, 2005, pp. 161-178.

[4] Z. Xu, D. John, and A. C. Boucouvalas, "Expressive image generation: Towards expressive Internet communications," *Journal of Visual Languages and Computing*, vol. 17, No. 5, Oct 2006, pp. 445-465.

[5] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic Nervous-System Activity Distinguishes among Emotions," *Science*, vol. 221, No. 4616, 1983, pp. 1208-1210.

[6] P. Rainville, A. Bechara, N. Naqvi, and A. R. Damasio, "Basic emotions are associated with distinct patterns of cardiorespiratory activity," *International Journal of Psychophysiology*, vol. 61, No. 1, 2006, pp. 5-18.

[7] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, No. 3, May 1993, pp. 261-273.

[8] D. A. Norman, *The Psychology of Everyday Things*. New York: Doubleday / Currency, 1990.

[9] D. Sander, D. Grandjean, and K. R. Scherer, "A systems approach to appraisal mechanisms in emotion," *Neural Networks*, vol. 18, No. 4, May 2005, pp. 317-352.

[10] K. R. Scherer, *Appraisal considered as a process of multi-level sequential checking*. Oxford: Oxford University Press, 2001.

[11] R. R. Cornelius, *The Science of Emotion*. Upper Saddle River, NJ: Prentice-Hall, 1996.

[12] R. L. Mandryk and M. S. Atkins, "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies," *International Journal of Human-Computer Studies*, vol. 65, No. 4, Apr 2007, pp. 329-347.

[13] M. Prensky, "Computer games and learning : digital game-based learning," in *Handbook of computer games studies*, J. Raessens and J. Goldstein, Eds. Cambridge, MA: The MIT Press, 2005.

[14] J. M. Carroll, "Learning in Communities, Interdisciplinary Persepectives on Human Centered Information Technology," in *Human-computer interaction series*, J. Karat, Ed. London: Springer, 2009.

[15] M. Kankaanranta and P. Neittaanmäki, "Design and Use of Serious Games," in *Intelligent systems, control, and automation: science and engineering*, vol. 37, S. G. Tzafestas, Ed. London: Springer, 2009.

[16] W. Burleson and R. W. Picard, "Gender-specific approaches to developing emotionally intelligent learning companions," *IEEE Intelligent Systems*, vol. 22, No. 4, Jul-Aug 2007, pp. 62-69.

[17] S. H. Fairclough, "Psychophysiological inference and physiological computer games," Brainplay'07: Brain-Computer Interfaces and Games, Workshop at the Int. Conf. on Advances in Computer Entertainment, Salzburg, Austria June 13-15, 2007.

[18] B. Cowley, D. Charles, and M. Black, "Toward an Understanding of Flow in Video Games," *ACM Computers in Entertainment* vol. 6, No. 2, July 2008.

References

[19] S. K. Sutton and R. J. Davidson, "Prefrontal brain asymmetry: A biological substrate of the behavioral approach and inhibition systems," *Psychological Science*, vol. 8, No. 3, May 1997, pp. 204-210.

[20] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, No. 1, Jan 2001, pp. 32-80.

[21] S. Helal, W. Mann, H. El-Zabadani, J. King, Y. Kaddoura, and E. Jansen, "The Gator Tech Smart House: A programmable pervasive space," *Computer*, vol. 38, No. 3, Mar 2005, pp. 50-+.

[22] A. Hanjalic, "Adaptive extraction of highlights from a sport video based on excitement modeling," *IEEE Transactions on Multimedia*, vol. 7, No. 6, Dec 2005, pp. 1114-1122.

[23] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on Multimedia*, vol. 7, No. 1, February 2005, pp. 143-154.

[24] H. L. Wang and L. F. Cheong, "Affective understanding in film," *Ieee Transactions on Circuits and Systems for Video Technology*, vol. 16, No. 6, Jun 2006, pp. 689-704.

[25] C. C. Liu, K. Conn, N. Sarkar, and W. Stone, "Online affect detection and robot behavior adaptation for intervention of children with autism," *IEEE Transactions on Robotics*, vol. 24, No. 4, Aug 2008, pp. 883-896.

[26] J. R. Wolpaw, N. Birbaumer, W. J. Heetderks, D. J. McFarland, P. H. Peckham, G. Schalk, E. Donchin, L. A. Quatrano, C. J. Robinson, and T. M. Vaughan, "Brain-computer interface technology: A review of the first international meeting," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, No. 2, Jun 2000, pp. 164-173.

[27] T. M. Vaughan, W. J. Heetderks, L. J. Trejo, W. Z. Rymer, M. Weinrich, M. M. Moore, A. Kubler, B. H. Dobkin, N. Birbaumer, E. Donchin, E. W. Wolpaw, and J. R. Wolpaw, "Brain-computer interface technology: A review of the second international meeting," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, No. 2, Jun 2003, pp. 94-109.

[28] J. Kronegg, "Mesures et optimisation des performances pour une interface de communication neuronale", Computer science department, University of Geneva, Geneva, 2006.

[29] F. Lotte, M. Congedo, A. Lecuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," *Journal of Neural Engineering*, vol. 4, No. 2, Jun 2007, pp. R1-R13.

[30] J. Kronegg, G. Chanel, S. Voloshynovskiy, and T. Pun, "EEG-based synchronized brain-computer interfaces: A model for optimizing the number of mental tasks," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 15, No. 1, Mar 2007, pp. 50-58.

[31] A. Choppin, "EEG-Based Human Interface for Disabled Individuals: Emotion Expression with Neural Networks", Master thesis, Information processing, Tokyo institute of technology, Yokohama, Japan, 2000.

[32] R. S. Lazarus, "Progress on a Cognitive Motivational Relational Theory of Emotion," *American Psychologist*, vol. 46, No. 8, Aug 1991, pp. 819-834.

[33] R. R. Cornelius, "Theoretical approaches to emotion," Proc. Int. Speech Communication Association (ISCA) Workshop on Speech and Emotion, Belfast, Ireland, 2000.

[34] A. Ortony and T. J. Turner, "Whats Basic About Basic Emotions," *Psychological Review*, vol. 97, No. 3, Jul 1990, pp. 315-331.

[35] C. M. Whissell, "The dictionary of affect in language," in *Emotion: Theory, Research and Experience*, vol. 4, R. Pluchik and H. Kellerman, Eds. New York: Academic Press, Inc., 1989, pp. 113-131.
[36] G. Chanel, J. Kronegg, D. Grandjean, and T. Pun, "Emotion assessment: Arousal evaluation using EEG's and peripheral physiological signals," in *Multimedia Content Representation, Classification and Security*, vol. 4105, A. K. J. B. Gunsel, A. M. Tekalp, B. Sankur, Ed., Lecture Notes in Computer Science ed. Istanbul, Turkey: Springer LNCS, 2006, pp. 530-537.
[37] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Boredom, Engagement and Anxiety as Indicators for Adaptation to Difficulty in Games," 12th International MindTrek Conference: Entertainment and Media in the Ubiquitous Era, ACM, Tampere, Finland, Oct., 2008.
[38] G. Chanel, K. Ansari-Asl, and T. Pun, "Valence-arousal evaluation using physiological signals in an emotion recall paradigm," IEEE SMC and International Conference on Systems, Man and Cybernetics, Smart cooperative systems and cybernetics: advancing knowledge and security for humanity, Montreal, Canada, October 7-10, 2007.
[39] G. Chanel, J. J. M. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *accepted for publication in International Journal of Human-Computer Studies, DOI:10.1016/j.ijhcs.2009.03.005*, 2009.
[40] W. James, "What is an emotion ?," *Mind*, vol. 9, 1884, pp. 188-205.
[41] M. B. Arnold, "Emotion and personality." New York: [42] P. Ekman, W. V. Friesen, M. O'Sullivan, A. Chan, I. Diacoyanni-Tarlatzis, K. Heider, R. Krause, W. A. LeCompte, T. Pitcairn, and P. E. Ricci-Bitti, "Universals and cultural differences in the judgments of facial expressions of emotion," *Journal of Personality and Social Psychology*, vol. 53, No. 4, Oct. 1987 1987, pp. 712-717.
[43] J. A. Gray, "The Neuropsychology of Anxiety - an Inquiry into the Functions of the Septo-Hippocampal System," *Behavioral and Brain Sciences*, vol. 5, No. 3, 1982, pp. 469-484.
[44] J. Panksepp, "Toward a General Psycho-Biological Theory of Emotions," *Behavioral and Brain Sciences*, vol. 5, No. 3, 1982, pp. 407-422.
[45] W. McDougall, *An introduction to social psychology*. Boston: Luce, 1926.
[46] O. H. Mower, *Learning theory and behavior*. New York: Wiley, 1960.
[47] R. Pluchik, "A general psychoevolutionary theory of emotion," in *Emotion: Theory, Research and Experience*, vol. 3-31, R. Pluchik and H. Kellerman, Eds. New York: Academic Press, Inc., 1980.
[48] R. Pluchik, *Emotion: a psychoevolutionary synthesis*. New York: Harper and Row, 1980.
[49] R. Pluchik, "A psychoevolutionary theory of emotions," *Social Science Information*, vol. 21, 1982, pp. 529.553.
[50] P. Shaver, J. Schwartz, D. Kirson, and C. O'Connor, "Emotion Knowledge: Further Exploration of a Prototype Approach," in *Emotions in Social Psychology*, W. Parrott, Ed. Philadelphia: Psychology Press, 2001, pp. 26-56.
[51] J. A. Russell, "A Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, No. 6, Dec. 1980, pp. 1161-1178.
[52] J. A. Russel, "Affect Grid: A Single-Item Scale of Pleasure and Arousal," *Journal of Personality and Social Psychology*, vol. 57, No. 3, Sep. 1989, pp. 493-502.
[53] J. A. Russel, M. Lewicka, and T. Niit, "A Cross-Cultural Study of a Circumplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 57, No. 5, 1989, pp. 848-856.

References

[54] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): digitized photographs, instruction manual and affective ratings," University of Florida, Gainesville, FL 2005.

[55] M. Isomursu, M. Tahti, S. Vainamo, and K. Kuutti, "Experimental evaluation of five methods for collecting emotions in field settings with mobile applications," *International Journal of Human-Computer Studies*, vol. 65, No. 4, Apr 2007, pp. 404-418.

[56] N. H. Frijda, *The emotions*. Cambridge ; New York Cambridge Univ. Press 1986.

[57] A. Ortony, G. L. Clore, and A. Collins, *The cognitive structure of emotions*. Cambridge, MA: Cambridge University Press, 1988.

[58] D. Sander and O. Koenig, "No inferiority Complex in the Study of Emotion Complexity: A Cognitive Neuroscience Computational Architecture of Emotion," *Cognitive Science Quaterly*, vol. 2, No. 3/4, 2002, pp. 249-272.

[59] C. Elliott, "The Affective Reasoner: A process model of emotions in a multi-agent system", Northwestern university, Evanston, Illinois, 1992.

[60] C. Adam, B. Gaudou, A. Herzig, and D. Longin, "OCC's emotions: a formalization in a BDI logic," International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA 2006), Varna, Bulgaria, Sept 13-15, 2006.

[61] D. Grandjean, D. Sander, and K. R. Scherer, "Conscious emotional experience emerges as a function of multilevel, appraisal-driven response synchronization," *Consciousness and Cognition*, vol. 17, No. 2, June 2008, pp. 484-495.

[62] S. Delplanque, D. Grandjean, C. Chrea, G. Coppin, L. Aymard, I. Cayeux, C. Margot, M. I. Velazco, D. Sander, and K. R. Scherer, "Sequential unfolding of novelty and pleasantness appraisals of odors: evidence from facial electromyography and autonomic reactions," *Emotion, In press*.

[63] A. Ortony, W. Revelle, and R. Zinbarg, "Why Emotional Intelligence needs a fluid component," in *The science of Emotional Intelligence*, G. Matthews, M. Zeidner, and R. D. Roberts, Eds.: Oxford University Press, 2007.

[64] K. R. Scherer, "Emotions as Episodes of Subsystem Synchronization Driven by Nonlinear Appraisal Processes," in *Emotion, Development and Self-Organization, Dynamic System Approaches to Emotional Development*, M. D. Lewis and I. Granic, Eds.: Cambridge University press, 2000.

[65] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial expression recognition from video sequences: temporal and static modeling," *Computer Vision and Image Understanding*, vol. 91, No. 1-2, Jul-Aug 2003, pp. 160-187.

[66] Z. Zeng, J. Tu, B. M. Pianfetti, and T. S. Huang, "Audio–Visual Affective Expression Recognition Through Multistream Fused HMM," *IEEE Transactions on multimedia*, vol. 10, No. 4, June 2008, pp. 570-577.

[67] Z. H. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, No. 1, Jan 2009, pp. 39-58.

[68] Y. Wang and L. Guan, "Recognizing human emotional state from audiovisual signals," *IEEE Trans. on multimedia*, vol. 10, No. 5, Aug. 2008, pp. 936-946.

[69] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, 2005, pp. 407-422.

[70] P. Rani, N. Sarkar, and C. Liu, "Maintaining Optimal Challenge in Computer Games through Real-Time Physiological Feedback," 11th HCI International, Las Vegas, USA, July 22-27, 2005.

[71] Z. Hammal, "Segmentation des Traits du Visage, Analyse et Reconnaissance d'Expressions Faciales par les Modèles de Croyance Transférable", Sciences Cognitive, Joseph Fourier Grenoble, Grenoble, 2006.

[72] P. Rani and N. Sarkar, "Operator Engagement Detection for Robot Behavior Adaptation," *Int. Journal of Advanced Robotic Systems*, vol. 4, No. 1, 2007, pp. 1-12.

[73] J. D. Morris, "SAM:The Self-Assessment Manikin, An Efficient Cross-Cultural Measurement of Emotional Response," *Journal of Advertising Research*, vol. 35, No. 6, November 1995.

[74] D. Purves, G. J. Augustine, D. Fitzpatrick, and L. C. Katz, *Neuroscience (french edition)*, 1997.

[75] G. Pfurtscheller and F. H. L. da Silva, "Event-related EEG/MEG synchronization and desynchronization: basic principles," *Clinical Neurophysiology*, vol. 110, No. 11, Nov 1999, pp. 1842-1857.

[76] R. Oostenveld and P. Praamstra, "The five percent electrode system for high-resolution EEG and ERP measurements," *Clinical Neurophysiology*, vol. 112, No. 4, Apr 2001, pp. 713-719.

[77] J. E. LeDoux, "Emotion Circuits in the Brain," *Annual Review of Neuroscience*, vol. 23, March 2000, pp. 155-184.

[78] R. J. Davidson, D. C. Jackson, and N. H. Kalin, "Emotion, Plasticity, Context, and Regulation: Perspectives From Affective Neuroscience," *Psychological Bulletin*, vol. 126, No. 6, 2000, pp. 890-909.

[79] E. T. Rolls, "Précis of The brain and emotion," *Behavioral and Brain Sciences*, vol. 23, No. 2, April 2000, pp. 177-233.

[80] R. Adolphs, D. Tranel, and A. R. Damasio, "Dissociable neural systems for recognizing emotions," *Brain and Cognition*, vol. 52, No. 1, Jun 2003, pp. 61-69.

[81] R. J. Davidson, "Affective neuroscience and psychophysiology: Toward a synthesis," *Psychophysiology*, vol. 40, No. 5, 2003, pp. 655-665.

[82] L. I. Aftanas, N. V. Reva, A. A. Varlamov, S. V. Pavlov, and V. P. Makhnev, "Analysis of Evoked EEG Synchronization and Desynchronization in Conditions of Emotional Activation in Humans: Temporal and Topographic Characteristics," *Neuroscience and Behavioral Physiology*, vol. 34, No. 8, 29 May 2002 2004, pp. 859-867.

[83] T. Costa, E. Rognoni, and D. Galati, "EEG phase synchronization during emotional response to positive and negative film stimuli," *Neuroscience Letters*, vol. 406, No. 3, Oct 9 2006, pp. 159-164.

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

[85] C. Féré, "Note sur les modifications de la resistance electrique sous l'influence des excitations sensorielles et des emotions," *Compt. Rend. Soc. Biol.*, vol. 5, 1888, pp. 217-219.

[86] D. C. Fowles, "The eccrine system and electrodermal activity," in *Psychophysiology*, M. G. H. Coles, E. Donchin, and S. W. Porges, Eds. New York: Guilford Press, 1986, pp. 51-96.

[87] J. A. Healey, "Wearable and Automotive Systems for Affect Recognition from Physiology", PhD Thesis, EECS Dpt., MIT, USA, 2000., Electrical Engineering and Computer Science Dept., MIT, 2000.

References

[88] M. E. Dawson, A. M. Schell, and D. L. Filion, "The electrodermal response system," in *Principles of psychophysiology: Physical, social and inferential elements*, J. T. Cacioppo and L. G. Tassinary, Eds. Cambridge: Cambridge University Press, 1990, pp. 295-324.

[89] H. Sequeira, P. Hot, L. Silvert, and S. Delplanque, "Electrical autonomic correlates of emotion," *International Journal of Psychophysiology*, No. In press, 2008.

[90] S. Delplanque, D. Grandjean, C. Chrea, L. Aymard, I. Cayeux, B. L. Calve, M. I. Velazco, K. R. Scherer, and D. Sander, "Emotional Processing of Odors: Evidence for a Nonlinear Relation between Pleasantness and Familiarity Evaluations," *Chemical Senses*, vol. 33, 2008, pp. 469-479.

[91] D. Shapiro, L. D. Jamner, J. D. Lane, K. C. Light, M. Myrtek, Y. Sawada, and A. Steptoe, "Blood pressure publication guidelines," *Psychophysiology*, vol. 33, No. 1, Jan 1996, pp. 1-12.

[92] R. Sinha, W. R. Lovallo, and O. A. Parsons, "Cardiovascular differentiation of emotions," *Psychosomatic Medicine*, vol. 54, No. 4, 1992, pp. 422-435.

[93] C. L. Lisetti and F. Nasoz, "Using Noninvasive Wearable Computers to Recognize Human Emotions from Physiological Signals," *Journal on Applied Signal Processing*, vol. 11, 2004, pp. 1672-1687.

[94] R. A. Wright, R. J. Contrada, and M. J. Patane, "Task-Difficulty, Cardiovascular-Response, and the Magnitude of Goal Valence," *Journal of Personality and Social Psychology*, vol. 51, No. 4, Oct 1986, pp. 837-843.

[95] R. A. Wright and J. C. Dill, "Blood-Pressure Responses and Incentive Appraisals as a Function of Perceived Ability and Objective Task Demand," *Psychophysiology*, vol. 30, No. 2, Mar 1993, pp. 152-160.

[96] G. G. Berntson, J. T. Bigger, D. L. Eckberg, P. Grossman, P. G. Kaufmann, M. Malik, H. N. Nagaraja, S. W. Porges, J. P. Saul, P. H. Stone, and M. W. VanderMolen, "Heart rate variability: origins, methods, and interpretive caveats," *Psychophysiology*, vol. 34, No. 6, Nov 1997, pp. 623-648.

[97] J. R. Jennings, W. K. Berg, J. S. Hutcheson, P. Obrist, S. Porges, and G. Turpin, "Publication Guidelines for Heart-Rate Studies in Man," *Psychophysiology*, vol. 18, No. 3, 1981, pp. 226-231.

[98] T. Ritz, B. Dahme, A. B. Dubois, H. Folgering, G. K. Fritz, A. Harver, H. Kotses, P. M. Lehrer, C. Ring, A. Steptoe, and K. P. van de Woestijne, "Guidelines for mechanical lung function measurements in psychophysiology," *Psychophysiology*, vol. 39, No. 5, Sep 2002, pp. 546-567.

[99] F. H. Wilhelm, M. C. Pfaltz, and P. Grossman, "Continuous electronic data capture of physiology, behavior and experience in real life: towards ecological momentary assessment of emotion," *Interacting with Computers*, vol. 18, No. 2, 2006, pp. 171-186.

[100] J. Kim, "Emotion Recognition from Physiological Measurement," Humaine European Network of Excellence Workshop, Santorini, Greece, September 18-21, 2004.

[101] A. C. Guyton and J. E. Hall, *Textbook of medical physiology 11th edition*: Elsevier Inc., 2006.

[102] R. A. McFarland, "Relationship of skin temperature changes to the emotions accompanying music," *Applied Psychophysiology and Biofeedback*, vol. 10, No. 3, sept. 1985, pp. 255-267.

[103] C. Puri, L. Olson, I. Pavlidis, J. Levine, and J. Starren, "StressCam: non-contact measurement of users' emotional states through thermal imaging," Conference on Human Factors in Computing Systems (CHI'05), Portland, Oregon, April 2-7, 2005.

[104] I. Pavlidis, J. Levine, and P. Baukol, "Thermal image analysis for anxiety detection," IEEE International Conference on Image Processing, Thessaloniki, Greece, Oct. 7-10, 2001.

[105] S. Jarlier, D. Grandjean, K. N'Diaye, S. Delplanque, D. Sander, P. Vuilleumier, and K. R. Scherer, "Thermal imaging of facial expressions: investigating thermal correlates of Facial Action Units activities," 10th International Conference on Cognitive Neuroscience, Bodrum, Turkey, Sept. 1-5, 2008.

[106] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, No. 10, Oct 2001, pp. 1175-1191.

[107] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion Recognition Using Bio-Sensors: First Step Toward an Automatic System," Affective Dialog Systems: Tutorial and Research Workshop, Kloster Irsee, Germany, June 14-16, 2004.

[108] G. Stemmler, M. Heldmann, C. A. Pauls, and T. Scherer, "Constraints for emotion specificity in fear and anger: The context counts," *Psychophysiology*, vol. 38, No. 2, Mar 2001, pp. 275-291.

[109] C. D. Katsis, N. Katertsidis, G. Ganiatras, and D. I. Fotiadis, "Toward emotion recognition in car racing drivers: a biosignal processing approach," *IEEE Trans. Systems, Man and Cybernetics - Part A: Systems and Humans*, vol. 38, No. 3, May 2008, pp. 502-512.

[110] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short term monitoring of physiological signals," *Medical Biological Engineering and computing*, vol. 42, 2004, pp. 419-427.

[111] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: implementing and comparing selected methods for features extraction and classification," IEEE International Conference on Multimedia & Expo, 6-8 July, 2005.

[112] R. Sinha and O. A. Parsons, "Multivariate response patterning of fear and anger," *Cognition & Emotion*, vol. 10, No. 2, Mar 1996, pp. 173-198.

[113] K. Takahashi, "Remarks on Emotion Recognition from Bio-Potential Signals," Proceedings of the 2nd International Conference on Autonomous Robots and Agents, Palmerston North, New Zealand, December 13-15, 2004.

[114] E. Leon, G. Clarke, V. Callaghan, and F. Sepulveda, "A user-independent real-time emotion recognition system for software agents in domestic environments," *Engineering Applications of Artificial Intelligence*, vol. 20, No. 3, Apr 2007, pp. 337-345.

[115] T. Sakata, S. Watanuki, H. Sakamoto, T. Sumi, and Y.-K. Kim, "Objective evaluation of Kansei by a complementary use of physiological indexes, brain wave and facial expressions for user oriented designs," Proceedings of the10th Qmod conference, Quality Management and Organisational Development: Our Dreams of Excellence, Helsingborg, Sweden, June 18-20, 2007.

[116] P. Rani, C. Liu, and N. Sarkar, "An Empirical study of machine learning techniques for affect recognition in human-robot interaction," *Pattern Analysis & Applications*, vol. 9, 2006, pp. 58-69.

[117] M. Pelzer, J. D. Schipke, D. Horstkotte, and G. Arnold, "Effect of Respiration on Short-Term Heart-Rate-Variability," *Faseb Journal*, vol. 8, No. 5, Mar 18 1994, pp. A846.

[118] R. Bailón, P. Laguna, L. Mainardi, and L. Sörnmo, "Analysis of Heart Rate Variability Using Time-Varying Frequency Bands Based on Respiratory Frequency," IEEE 29th International Conference on EMBS, Lyon, France, Aug. 23-26, 2007.

References

[119] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, second ed: Wiley Interscience, 2001.

[120] C. M. Bishop, *Pattern recognition and machine learning*: Springer, 2006.

[121] E. L. van den Broek, M. H. Schut, J. H. D. M. Westerink, J. van Herk, and K. Tuinenbreijer, "Computing emotion awareness through facial electromyography," *Computer Vision in Human-Computer Interaction*, vol. 3979, 2006, pp. 52-63.

[122] K. Takahashi, "Remarks on SVM-based emotion recognition from multi-modal bio-potential signals," Proceedings of the 13th IEEE International Workshop on Robot and Human Interactive Communication Sept. 20-22, 2004.

[123] B. Herbelin, P. Benzaki, F. Riquier, O. Renault, and D. Thalmann, "Using physiological measures for emotional assessment: a computer-aided tool for cognitive and behavioural therapy," 5th International Conference on Disability, Oxford, 2004.

[124] C. Lee, S. Yoo, Y. Park, N. Kim, K. Jeong, and B. Lee, "Using Neural Network to Recognize Human Emotions from Heart Rate Variability and Skin Resistance," Engineering in Medecine and Biology 27th Annual Conference, Shangai, China, September 1-4, 2005.

[125] R. L. Mandryk, K. M. Inkpen, and T. W. Calvert, "Using psychophysiological techniques to measure user experience with entertainment technologies," *Behaviour & Information Technology*, vol. 25, No. 2, Mar-Apr 2006, pp. 141-158.

[126] Biosemi, "http://www.biosemi.com/. Biosemi website 30/01/2009."

[127] D. C. Fowles, M. J. Christie, R. Edelberg, W. W. Grings, D. T. Lykken, and P. H. Venables, "Publication Recommendations for Electrodermal Measurements," *Psychophysiology*, vol. 18, No. 3, May 1981, pp. 232-239.

[128] M. Aboy, J. McNames, T. Thong, D. Tsunami, M. S. Ellenby, and B. Goldstein, "An automatic beat detection algorithm for pressure signals," *IEEE Transactions on Biomedical Engineering*, vol. 52, No. 10, Oct 2005, pp. 1662-1670.

[129] K. Ansari-Asl, G. Chanel, and T. Pun, "A channel selection method for EEG classification in emotion assessment based on synchronization likelihood," 15th European Signal Processing Conference (Eusipco 2007), Poznan, Poland, Sept. 3-7, 2007.

[130] C. Berka, D. J. Levendowski, M. M. Cvetinovic, M. M. Petrovic, G. Davis, M. N. Lumicao, V. T. Zivkovic, M. V. Popovic, and R. Olmstead, "Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset," *International Journal of Human-Computer Interaction*, vol. 17, No. 2, 2004, pp. 151-170.

[131] A. T. Pope, E. H. Bogart, and D. S. Bartolome, "Biocybernetic System Evaluates Indexes of Operator Engagement in Automated Task," *Biological Psychology*, vol. 40, No. 1-2, May 1995, pp. 187-195.

[132] M. Besserve, M. Philippe, G. Florence, F. Laurent, L. Garnero, and J. Martinerie, "Prediction of performance level during a cognitive task from ongoing EEG oscillatory activites," *Clinical Neurophysiology*, vol. 119, No. 4, April 2008, pp. 897-908.

[133] F. G. Freeman, P. J. Mikulka, M. W. Scerbo, and L. Scott, "An evaluation of an adaptive automation system using a cognitive vigilance task," *Biological Psychology*, vol. 67, No. 3, Nov. 2004, pp. 283-297.

[134] R. Moddemeijer, "On Estimation of Entropy and Mutual Information of Continuous Distributions," *Signal Processing*, vol. 16, No. 3, Mar 1989, pp. 233-248.

[135] E. Pauwels, "Ethics for researchers, facilitating research excellence in FP7," 2007.

[136] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," 14th IJCAI, Montréal, Canada, August 20-25, 1995.

[137] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, No. 3, Sep 1995, pp. 273-297.
[138] J. Hua, Z. Xiong, J. Lowey, E. Suh, and E. R. Dougherty, "Optimal number of features as a function of sample size for various classification rules," *Bioinformatics*, vol. 21, No. 8, 2005, pp. 1509-1515.
[139] J. Platt, "Probabilities for SV Machines," in *Advances in Large Margin Classifiers*, A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. Cambridge, MA: MIT Press, 2000, pp. 61-64.
[140] H. T. Lin, C. J. Lin, and R. C. Weng, "A note on Platt's probabilistic outputs for support vector machines," *Machine Learning*, vol. 68, No. 3, Oct 2007, pp. 267-276.
[141] T. F. Wu, C. J. Lin, and R. C. Weng, "Probability estimates for multi-class classification by pairwise coupling," *Journal of Machine Learning Research*, vol. 5, Aug 2004, pp. 975-1005.
[142] C.-C. Chang and C.-J. Lin, "LIBSVM : a library for support vector machines," 2001.
[143] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, No. 3, Sum 2001, pp. 211-244.
[144] H. Zhang and J. Malik, "Selecting Shape Features Using Multi-class Relevance Vector Machine," EECS Department, University of California, Berkeley UCB/EECS-2005-6, October 10 2005.
[145] J. Ye, T. Li, T. Xiong, and R. Janardan, "Using Uncorrelated Discriminant Analysis for Tissue Classification with Gene Expression Data," *IEEE/ACM Transactions on computational biology and bioinformatics*, vol. 1, No. 4, Oct-Dec 2004, pp. 181-190.
[146] J. Yang, J. Y. Yang, and D. Zhang, "What's wrong with Fisher criterion?," *Pattern Recognition*, vol. 35, No. 11, Nov 2002, pp. 2665-2668.
[147] G. H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," Proc. 11th Int. Conf. on Machine Learning, San Francisco, USA, 1994.
[148] P. Pudil, F. Ferri, J. Novovicová, and J. Kittler, "Floating search methods for feature selection with nonmonotonic criterion functions," *Proc. of the IEEE Intl. Conf. on Pattern Recognition*, vol. 2, 1994, pp. 279-283.
[149] L. Yu and H. Liu, "Efficient Feature Selection via Analysis of Relevance and Redundancy," *Journal of Machine Learning Research*, vol. 5, 1205-1224 2004.
[150] P. Viola and M. Jones, "Fast and robust classification using asymmetric AdaBoost and a detector cascade," *Advances in Neural Information Processing Systems 14, Vols 1 and 2*, vol. 14, 2002, pp. 1311-1318.
[151] J. Kludas, E. Bruno, and S. Marchand-Maillet, "Can Feature Information Interaction help for Information Fusion in Multimedia Problems?," *Multimedia Tools and Applications Journal special issue on "Metadata Mining for Image Understanding"* vol. 42, No. 1, 2009, pp. 57-71.
[152] A. A. Freitas, "Understanding the crucial role of attribute interaction in data mining," *Artificial Intelligence Review*, vol. 16, No. 3, Nov 2001, pp. 177-199.
[153] C. Sanderson and K. K. Paliwal, "Identity verification using speech and face information," *Digital Signal Processing*, vol. 14, No. 5, Sep 2004, pp. 449-480.
[154] D. Ruta and B. Gabrys, "An Overview of Classifier Fusion Methods," *Computing and Information Systems*, vol. 7, No. 1, Feb. 2000.
[155] G. Bologna, B. Deville, and T. Pun, "On the use of the auditory pathway to represent image scenes in real-time," *Neurocomputing*, vol. 72, No. 4-6, Jan 2009, pp. 839-849.

[156] K. R. Scherer, E. S. Dan, and A. Flykt, "What Determines a Feeling's Position in Affective Space? A Case for Appraisal," *Cognition and emotion*, vol. 20, No. 1, 2006, pp. 92-113.

[157] L. Collet and R. Duclaux, "Hemispheric Lateralization of Emotions - Absence of Electrophysiological Arguments," *Physiology & Behavior*, vol. 40, No. 2, 1987, pp. 215-220.

[158] A. P. R. Smith, K. E. Stephhan, M. D. Rugg, and R. J. Dolan, "Task and Content Modulate Amygdala-Hippocampal Connectivity in Emotional Retrieval," *Neuron*, vol. 49, No. 4, Feb. 2006 2006, pp. 631-638.

[159] G. Barret, "Event-related potentials (ERPs) as a measure of complex cognitive function," *Electroencephalography and clinical neurophysiology, Supplement*, vol. 46, 1996, pp. 53–63.

[160] C. Guger, G. Edlinger, W. Harkam, I. Niedermayer, and G. Pfurtscheller, "How many people are able to operate an EEG-based brain-computer interface (BCI)?," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 11, No. 2, Jun 2003, pp. 145-147.

[161] M. Csikszentmihalyi, *Flow: the psychology of optimal experience*: New York: Harper Collins, 1991.

[162] K. Salen and E. Zimmerman, *Rules of play: game design fundamentals*. Cambridge: MIT Press, 2004.

[163] G. H. E. Gendolla and M. Richter, "Ego involvement and effort: Cardiovascular, electrodermal, and performance effects," *Psychophysiology*, vol. 42, No. 5, 2005, pp. 596-603.

[164] T. I. Alecu, "Robust Focalized Brain Activity Reconstruction using ElectroEncephaloGrams", Computer science department, University of Geneva, Geneva, 2005.

List of figures

List of figures

List of figures

List of tables

List of tables