THÈSE DE DOCTORAT DE
l'UNIVERSITÉ SORBONNE UNIVERSITÉ

Spécialité

Informatique
École doctorale Informatique, Télécommunications et Électronique (Paris)

Présentée par

Wenlu YANG

Pour obtenir le grade de

DOCTEUR de l'UNIVERSITÉ SORBONNE UNIVERSITÉ

Sujet de la thèse :

Personalized physiological-based emotion recognition and implementation on hardware

soutenance prévue le 27 février 2018
devant le jury composé de :

M. MARTIN Jean-claude	LIMSI-CNRS	Rapporteur
M. STRAUSS Olivier	Université Montpellier II	Rapporteur
M. BENNANI Younès	Université Paris 13	Examinateur
M. ESSID Slim	Télécom ParisTech	Examinateur
Mme OCHS Magali	Aix-Marseille Université	Examinatrice
M. MARSALA Christophe	Sorbonne Université	Co-Directeur
M. PINNA Andrea	Sorbonne Université	Co-Directeur
Mme RIFQI Maria	Université Paris 2	Co-Directrice
M. GARDA Patrick	Sorbonne Université	Invité

This thesis has been funded under the SENSE project within the Labex SMART (ANR-11-LABX-65) supported by French state funds managed by the ANR within the Investissements dAvenir programme under reference ANR-11-IDEX-0004-02.

CONTENTS

1 INTRODUCTION

1.1 CONTEXT AND MOTIVATION

Recent years have witnessed a steady growth of computer games industry, which has become one of the most popular leisure activities and has won great economical success. The worldwide video game market, which includes online, mobile and PC games with a wide variety of game types and a large consumer group spreading across the world, has made the game industry profiting.

Due to the competitive industry and high demand for novelty, there have been increasing interests in the field of Human Computer Interaction (HCI) towards video game. HCI is a multidisciplinary topic which involves cognitive sciences, psychology, engineering, and computer science. Emotion is an important factor in HCI. It has been proven that emotion plays an important role in perception, decision-making and behavior (Holbrook et al. 1984). In the complex context of video games, affective factors can greatly influence the player's experience. Traditional communication between human and computing system is purposeful where the intention of user is conveyed to the computing system with the controller such as keyboard or mouse. However, the lack of information related to the users' psychological state (e.g., emotions) provides little opportunity for the computing system to react and adapt in a dynamic fashion in order to meet the need of users. Therefore, the realization of a game system which is aware of the psychological state of player, is a prerequisite for the development of adaptive game systems that are capable of responding to the needs of the player.

Many modalities can be used to evaluate one's psychological state. Among them physiological information is viewed as an effective one, as physiological responses are viewed as a major component of the emotion response (Kreibig 2010). Physiological signals include the ones originated from *central nervous system* measurements and *peripheral nervous system* measurements. *Central nervous system* signals include electroencephalography (EEG), functional MRI (fMRI), functional Near-infra-red Spectroscopy (fNIRS). *Peripheral nervous system* measurements include electrocardiography (ECG), Electrooculography (EOG), electromyography (EMG), electrodermal activity (EDA), blood volume (BVP), blood oxygenation, respiration (RESP), skin temperature (SKT). It

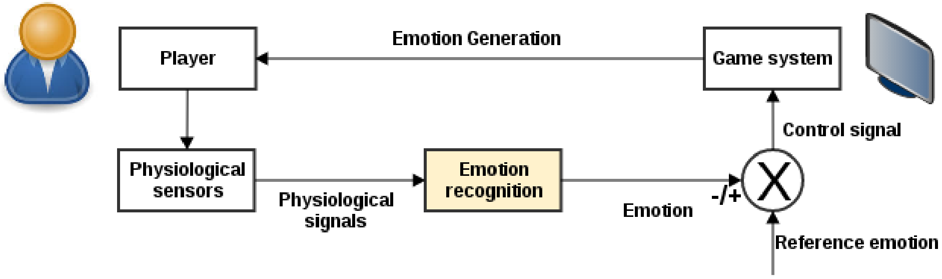


Figure 1.1: biocybernetic loop for affective game

is suggested that there exist a considerable physiological response specificity in emotion when considering subtypes of distinct emotions (Kreibig 2010).

Emotion-aware games are reported to be able to improve player's engagement, immersion, excitement, and challenge by dynamically adapting game features according to the physiological response (Yannakakis et al. 2008). These games operate by transforming user's physiological data into a control signal, which will be further used as an input to the biocybernetic loop for adaptation (Parnandi et al. 2015; Pope et al. 1995). The biocybernetic loop is based on concepts from classical control theory. The loop (Figure 1.1) is initiated by the collection of psychophysiological data from the user via physiological sensors. Then, the system analyzes the data in order to quantify the user's psychophysiological indicator level (e.g. heart rate), psychological state intensity (e.g. level of stress, engagement) or recognize psychological state (e.g. happy, sad). Next, an appropriate adaptation (e.g. game difficulty level, audio-visual effects) is determined by the control signal. The loop can be designed as negative loop for maintaining a desirable state (Parnandi et al. 2015), a positive loop for reinforcing certain emotion (Lindley et al. 2006) or can be a hybrid loop containing both negative loop and positive loop (Fairclough 2008). The biocybernetic loop has a wide application. For example, measuring a physiological EDA response of the player to maintain a desired level of arousal (Parnandi et al. 2015), measuring electroencephalographic (EEG) signals to maintain pilot's engagement (Pope et al. 1995). In the field of video game, the primary goals are to entertain the player and improve the player's game experience. It has been proposed that the targets of adaptation for affective game are: "assist me, challenge me, emote me" (Gilleade et al. 2005). Specifically, the objective is to offer assistance if the user is frustrated or in a state of "stuck", adapt the level of challenge in sustain task engagement and incorporate emotional element into the user's interface (Fairclough 2008).

One critical step in the biocybernetic loop is the analysis of the physiological and recognition of the emotional state (Figure 1.1, emotion recognition block). The general process of emotion recognition includes: measurement physiological signals, extraction of features from the measured modalities, selection of the most relevant features and finally use of all this information for psychophysiological inference. In some applications, the extracted signal features are directly mapped to the adapting mechanism of the affective game without any recognition of emotions (Dekker et al. 2007; Tijs et al. 2008a). This method offers intuitive adaptability but it lacks of evaluation on the player's real psychological state and is more likely to react to artifact. In other contexts, the adaptation is based on the result of the emotion inference, so that the adaptation is more robust.

Meanwhile, affective computing, coupled with new intelligent hardware, is a promising and profiting field. As it has been written by Glen Martin on Forbes in a paper on "wearable intelligence" [1] in 2014: *"intelligent devices other than phones and screens — smart headsets, glasses, watches, bracelets — are insinuating themselves into our daily lives. The technology for even less intrusive mechanisms, such as jewelry, buttons, and implants, exists and will ultimately find commercial applications"*. From this point of view, we can easily imagine the possibility of implementation of an emotion recognition module on an intelligent device such as on a game stick or a wearable device such as headsets. The potential advantage of this application is promising: to afford better emotional experience in HCI, to predict client's preference, to help maintain good mental health, etc.

The main objectives of this work are:

- Realization of an emotion recognition system for affective gaming;
- Evaluation of the computation resources required for such a system, and implementation on an embedded system.

1.2 CHALLENGE

Automatic recognition of emotion using physiological signal is a popular topic in the affective computing community (Picard 1995). However, physiological signals are non-stationary and recognition from these signals can be

[1] http://www.forbes.com/sites/oreillymedia/2014/04/01/wearable-intelligence/

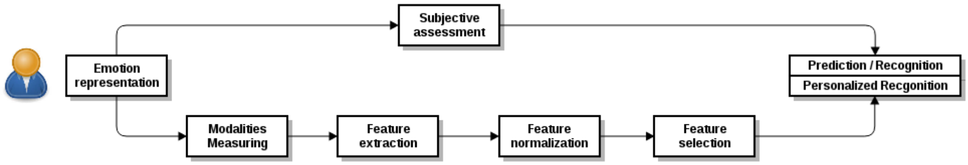


Figure 1.2: Main steps to create emotion recognition system

suffered from individual differences, day variation, user's mental states, electrode impedance or even noises from electrode failures. Despite of the high recognition accuracy achieved under strictly controlled laboratory conditions, challenges have been raised to create a flexible module which can provide user independent and condition independent emotion recognition in the real world. An argument (Fairclough 2008) has been that psychophysiological measures may be insufficient for the recognition of internal psychological states, such as emotions, due to (1) the absence of sufficient correspondence between the psychological state and associated physiological changes, (2) the lack of representative of the physiological signal to large range of psychological states, (3) the fuzzy boundary between psychological states, and (4) the variable and idiosyncratic experience of psychological states (Picard 2003). As has been mentioned in the reviews (Fairclough 2008; Kreibig 2010), psychophysiological experience is complex and there is no "literal, isomorphic representation of a given thought, intention or emotion"(Fairclough 2008). Even though there exist evidences of relation between psychological state and physiological response (Kreibig 2010), the quality of which may vary from context to context, measure to measure and between different internal state.

Despite of this, if we accept the fact that physiology computing can only provide a less-than-perfect representation of internal states, we come to a new critical questions: in a given context (in our circumstance, a video game), are we able to reveal some interesting relationship between the physiological measures and emotion? Is the model built on selected measures or features sensitive and diagnostic enough? Moreover, as our final objective of creating an embedded emotion recognition system, a further question is, can the model be implemented on the embedded system considering the computing resource constraints?

In this thesis, we present a design process of physiological-based emotion recognition embedded system, which compromises a wide, multidisciplinary domain of research. Figure 1.2 displays the main steps that compose a traditional emotion/mental state recognition system. Each step faces specific challenges:

- **representation:** Researches on emotion have been flourishing for long, yet debate continues about how to represent them. There exist various theories on how to represent emotions. The two most widely accepted approaches for modeling emotions are the *categorical approaches* and the *dimensional approaches*. Both approaches provides different representation and have both advantages and drawbacks (refer to Chapter 2 for a detailed review). The occurrence, frequency and combination of different emotions have a huge influence on the player's engagement, and motivation and thus greatly impact player's game experience (Bontchev 2016). The complexity of emotion makes it challenging to select appropriate methods to represent emotions in video game context.

- **modality measurement:** Modality measurement is closely related to data collection. The research interest in affective computing has motivated the creation of novel databases for affective computing (refer to Chapter 3 for a detailed review). Common simulations for emotions are image, music, film clips (Koelstra et al. 2012; Abadi et al. 2015), which are flawed in the sense that they are essentially passive so that the emotional experience may not generalize to active tasks such as video game. Therefore, challenge raises to collect relevant data to the application context. In the active context such as video game, more challenging issues should be taken into account: which game should be used as stimulation to elicit relatively rich and repetitive emotions? how to make the measure of physiological signal less intrusive to the participant? how to segment the game in order to label the emotional moments? Properly solving these problems is a necessity for obtaining reliable data for further analysis.

- **subjective assessment:** Despite the variable and idiosyncratic experience of psychological states (Picard 2003), subjective assessment of emotion remains to be an effective measure to evaluate participants' internal state. There are different methods of obtaining subjective assessment such as questionnaire, interview, focus-group. Each method has its own application cases (refer to Chapter 2 for a detailed review). In an experimental study, acquiring the subjective assessment which truly reflect players' in game emotion assessment without disturbing their game experience is also challenging.

- **feature extraction and selection:** Researchers have not reached a consensus on the most effective features of physiological signals for affective computing. The most presented physiological features in the literature (Kreibig 2010) are heart rate, followed by skin conductance level and other cardiovascular variables. The common features are extracted from the time domain, the frequency domain on the raw signal, preprocessed signal or the transformed signal. Physiological signals are high-dimensional data. The selection of appropriate signals and features plays a vital role in the emotion recognition model.

- **feature normalization:** Considering the great individual difference, the normalization is widely applied in the physiological-based affective computing (Koelstra et al. 2012; Soleymani et al. 2012). Extracted features may be optionally normalized for reducing both intra-and inter-subject variability. The objective is to reduce the effect of irrelevant variables other than emotion by setting a referencing baseline or offering a numerical range. The selection of a normalization method and a referencing baseline are dependent of the context. Therefore, researchers must make a wise choice of feature normalization.

- **prediction/recognition:** Based on the extracted features, researchers usually implement several classifiers and compare their performances by using cross-validation. (Novak et al. 2012) compared different classification methods such as k-Nearest Neighbour(kNN), Bayesian Networks (BNT), Regression Trees (RT) and Decision Trees (DT), Naïve Bayes Classifier (NB), Linear Discriminant Analysis (LDA)], and Support Vector Machines (SVM). Authors claimed that the accuracy rates are dependent on types of extracted features, feature selection and normalization. Therefore, researcher should choose the learning model most relevant to the faced problem.

- **model personalizing:** Traditional approach of usability test holds universal and static view of players which focus on developing the best setting possible for all the players. Most emotion recognition modules learn a model from a group of users and apply the same learned model for everyone (Tijs et al. 2008b; Chanel et al. 2011; Negini et al. 2014; Liu et al. 2009). However, considering the diversity of players and variability of the game context, such a simple approach is clearly unable to obtain satisfactory results. Different players have different skill levels, preference for game, motivation for playing and eventually different physiological response. An identical setting is unable to meet the demand of all of players. In order to achieve a better recognition accuracy, the model should be able to adapt to the player.

- **embedded system implementation:** Once the personalized physiological-based emotion recognition model is created, the remaining question is whether it is possible to implement it on an embedded system while meeting the computation time and memory constraints. The required computation resource for the previous step such as feature extraction, model prediction should be evaluated and compared in order to select a overall solution for the final implementation.

1.3 CONTRIBUTIONS

In this work, we made contributions on several steps of creating physiological-based emotion recognition embedded system. We achieve our goal by dividing and conquer the problem in three aspects: affective gaming, emotion recognition model, and implementation on an embedded system.

- Concerning **affective gaming** aspect, we introduce a new multi-modal database for affective gaming. In order to achieve this objectives, the achieved works are presented as follows:
  - In Chapter 2, we review related work concerning emotion theory, emotion measuring, assessment and affective gaming. The reviewed works served as important base to settle our study method.
  - In Chapter 3, we describe how we collected physiological signals and self assessment data in a video game context. we introduce the DAG database[2] - a multi-modal **D**atabase for **A**ffective **G**aming. We focus on peripheral physiological signals (ECG, EDA, Respiration, EMG). Two kinds of self-assessed evaluations are available in the database: minor scope evaluation on *game event* and global scope evaluation on *game sequence*. This database is made publicly available to support the affective game research. In the end, we also present *statistical analysis* on player's self-assessed annotation of evaluation on game event and game sequence.
- Concerning the creation of **emotion recognition model**, we carried out a set of analysis on the proposed DAG database. We present the training method and evaluation of general, user-specific model and

[2] http://erag.lip6.fr

personalized group-based model. The achieved works concerning this part are presented as follows:

  - In Chapter 4, we present how we process the physiological signals and extract features.
  - In Chapter 5, we present a set of analyses to create general model concerning:
    1. *emotional moment detection*: realized by classification of the sequences with and without annotations. Effects of segmentation lengths and relevant features are discussed;
    2. *emotion recognition*: realized by classification of emotions on game events. Effects of segmentation lengths and relevant features, as well as three normalization methods (standard normalization, neutral baseline referencing normalization, precedent moment referencing normalization) aiming to reduce individual variability are discussed;
    3. *game experience evaluation*: realized by preference learning on match rankings.
  - In Chapter 6, we present the performance of a user-specific model. Different feature selection methods (filter and wrapper, nested LOOCV and non nested LOOCV), optimal size of feature set size were investigated and user-specific model was trained on each subject. We confirm the existence of individual variability among subjects and presented the flaw of using user-specific model.
  - In Chapter 7, we present how to find physiological similar user by using clustering techniques. We show that by clustering users, the performance of recognition can be improved.
  - In Chapter 8, we propose a new group-based model that both takes into consideration the individual variability and makes the most use of existing data. We show that the proposed group-based model performs better than the general model and the user-specific model. We also investigate some characteristics of the proposed model.

- Concerning **implementation on embedded system**, we implemented the proposed model on an embedded system.
  - In Section 8.4, we evaluate the computation time for the personalized emotion recognition model. We realize a simulation on computer and an implementation on a real ARM A9 embedded

system. And we present that the proposed method can meet the time requirement.

Finally, we summarize the achievement of the thesis in Chapter 9 and propose some future work.

Part I

BACKGROUND AND AFFECTIVE DATASET

In this part, we present the contribution of this thesis concerning the affective gaming aspect.

We firstly review related work concerning emotion theory, emotion measuring, assessment and affective gaming (Chapter 2). The reviewed works served as important base to settle our study method.

Then, we describe how we collected physiological signals and self assessment data in a video game context. we introduce the DAG database[3] - a multi-modal **D**atabase for **A**ffective **G**aming. We focus on peripheral physiological signals (ECG, EDA, Respiration, EMG). Two kinds of self-assessed evaluations are available in the database: minor scope evaluation on *game event* and global scope evaluation on *game sequence* (Chapter 3).

In the end, we also present *statistical analysis* on player's self-reported assessment (Chapter 3).

[3] http://erag.lip6.fr

2 AFFECTIVE GAMING

Affective Gaming (AG) is a relatively new field of research that exploits human emotion for the enhancement of player's experience during video game play (Christy et al. 2014). It is an interdisciplinary science of emotion theory, psychophysiology, game research, user experience research.

In order to integrate emotion factors into video games, one major challenge is the realization of an emotion recognition system. The objective of an emotion recognition system is by using relevant emotion measures (e.g. facial expression, physiological signals), trying to deduce subject's emotion assessment. The emotion assessment can be represented using various methods based on different emotion theories. In this chapter, we first review various emotion representation methods (Section 2.1). Then, we overview the different emotion measuring modalities in the emotion research (Section 2.2) as well as various methods of emotion assessment approaches (Section2.3). In the end, we present a review of affective gaming research (Section 2.4).

2.1 EMOTION REPRESENTATION

Researches on emotion have been flourishing for long, yet debate continues about the nature of emotions, their biological mechanisms, their categories and their role in our daily activities (Izard 2007). Among these topics, emotion representation is one of the fundamental question of emotion researches.

Emotion representation characterizes the way of modeling emotions. In the task of emotion recognition, it defines the emotion labels that can be used to describe the emotional states. There exist various theories on how to represent emotions. The two most widely accepted approaches for modeling emotions are the *categorical approach* and the *dimensional approach*. In this section, we present these two approaches as well as their advantages and drawbacks.

2.1.1 *Categorical approach*

The categorical approach claims that there exist a small set of basic emotions. From a biological perspective, this idea is based on the belief that there might be neurophysiology and anatomical substrates corresponding to the basic emotions (Ortony et al. 1990). From a psychological perspective, basic emotions are often held to be the primitive building blocks of other complex emotions (Ortony et al. 1990).

Tomkins (Tomkins 1962; Tomkins 1963) is typically cited as the modern inspiration for the ”basic emotion” approach. He claimed that all instances of emotion that bear the same name are supposed to show the same pattern of reaction such as behavior, bodily activation, expression, etc. They are controlled by neural programs or circuits, which are ”hardwired, preprogrammed, genetically transmitted mechanisms that exist in each of us”. Nine affects are proposed in this theory - enjoyment/joy, surprise/startle, anger/rage, disgust, dissmell, distress/anguish, fear/terror, shame/humiliation.

(Ekman et al. 1982) based their assumptions mainly on the facial expression of emotions. The authors set up a standard, that all emotions which share the nine characteristics can be viewed as basic emotions. In their studies, facial expressions of emotions were recognized by people from very different cultures. Expressions they found to be universal included anger, disgust, fear, happiness, sadness, and surprise. Their latter work has included more emotions such as excitement, amusement, fiero and sensory pleasure (Ekman et al. 2011).

OCC Model (Ortony 1990) states the 3 factors of the generation of emotions - “consequence of events”, “action of agents”, and “aspects of objects”. The model specifies about 22 categories of emotions and the processes to follow in order to decide one’s emotion. It is a widely used model of emotion and easy to implement.

In a review (Ortony et al. 1990), the authors gave a summary of a representative set of emotion theorists and their basic emotion theory (Table 2.1). Although many researchers share the view that some emotions are basic, there is still little agreement on what these basic emotions are and why they are basic.

From literature discussing categorical methods, we inspect some of its advantages and drawbacks. The advantages of this model are:

1. Intuitive
   Once emotion is identified, the process of generation, perception as well as recognition is intuitive. For example, when thinking about happiness, one can easily think about the event to trigger happiness, the feeling when we are happy or recognize happiness from one’s face, voice or behavior.

Reference	Basic emotions	Basis for inclusion
(Arnold 1960)	Anger, aversion, courage, ejection, desire, despair, fear, hate, hope, love sadness	Relation to action tendencies
(Ekman et al. 1982)	Anger, disgust, fear, joy, sadness, surprise	Universal facial expression
(Frijda 1986)	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
(Gray 1982)	Rage and terror, anxiety, joy	Hardwired
(Izard 1971)	Anger, contempt, disgust, distress, fear guilt, interest, joy, shame, surprise	Hardwired
(James 1884)	Fear, grief, love, rage	Bodily involvement
(McDougall 1926)	Anger, disgust, elation, fear, subjection, tender-emotion, wonder	Relation to instincts
(Mowrer 1960)	Pain, pleasure	Unlearned emotional states
(Oatley et al. 1987)	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
(Panksepp 1982)	Expectancy, fear, rage, panic	Hardwired
(Plutchik 1980)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
(Tomkins 1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame	Density of neural firing
(Watson et al. 1925)	Fear, love, rage	Hardwired
(Weiner et al. 1984)	Happiness, sadness	Attribution independent

Table 2.1: A selection of lists of "basic" emotions (cited and extracted from (Ortony et al. 1990))

2. Concise
   As all the emotions are combined with a few basic ones, the number of major emotions is limited. They are therefore widely used in the field of emotional computing, as they can be served as labels for emotion recognition.

Despite these advantages, the categorical representation of emotions have some drawbacks:

1. Vagueness of language
   Same emotion may be labeled differently by different researchers (for example, some theorists use the term ″anger″ and others the word ″rage″ while presumably referring to the same emotion).

2. Cognitive bias
   As a subjective evaluation, perception of emotion may be different among different persons.

3. Complex attribute of emotion
   Cannot cover all the emotional aspects. People exhibit subtle and complex mental/ affective states which may be too difficult to handle.

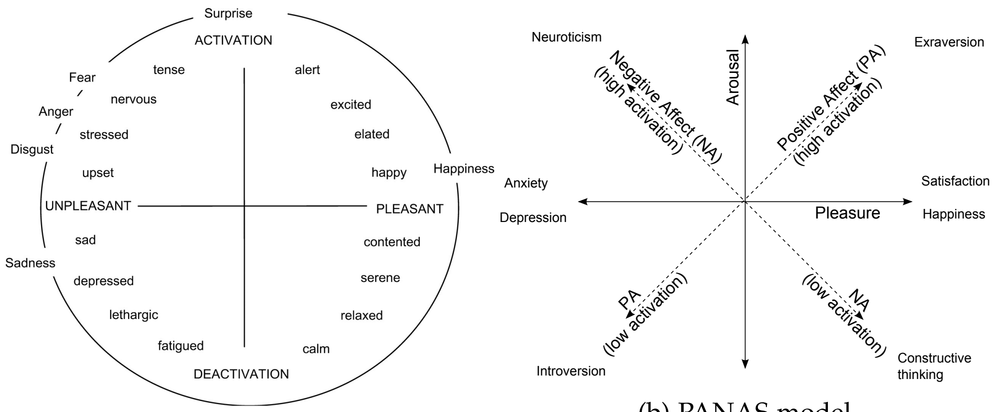
4. Evolving attribute of emotion
   Human affective state evolves continuously, categorical approach may be not enough to describe the transition of emotion.

2.1.2 *Dimensional approach*

Dimensional emotion theories use dimensions rather than discrete categories to describe the structure of emotions. There exist one or more major dimensions, and all the emotions can be represented in the space of these dimensions.

Regarding 2D - dimension models, Russell's Circumplex model of affect (Russell 1980a) (shown in Figure 2.1a) is one of the most widely used model. This model suggests that emotions are distributed in a two-dimensional circular space, containing arousal and valence dimensions. Arousal represents the general excitation and is presented on the vertical axis, ranging from deactivation to activation. Valence means the intrinsic attractiveness/″good″-ness or averseness/″bad″-ness of an event. It represents the horizontal axis, ranging from unpleasant to pleasant.

Another well known 2D - dimension model is the PANAS model (Watson et al. 1988) (shown in Figure 2.1b). The PANAS model suggests that positive



(a) Circumplex model

(b) PANAS model

Figure 2.1: Two dimensional models

affects and negative affects are two separate systems, that one can experience positive and negative affects at the same time. In the PANAS model, the vertical axis represents low to high positive affect and the horizontal axis represents low to high negative affect, while valence and arousal axis lay at a 45-degree rotation over these axes.

Concerning the 3D - dimensions models, (Schlosberg 1954) presented a reverse conical model with dimensions: (a) pleasantness vs. unpleasantness, (b) attention vs. rejection and (c) level of activation (Figure 2.2). In (Plutchik et al. 1997), they proposed the following four bipolar pairs of basic emotions: (a) joy vs. sadness, (b) anticipation vs. surprise, (c) anger vs. fear, and (d) disgust vs. trust.

Another 3-dimension model is the PAD emotional state model (Mehrabian 1996) (Figure 2.3), which is an extension of the 2 - dimension Circumplex (Arousal vs. Valence) model. The PAD model contains the dimensions (a) *pleasure* (b) *arousal* and (c) *dominance* (dominance means the controlling and dominant nature of the emotion. For instance while both fear and anger are unpleasant emotions, anger is a dominant emotion, while fear is a submissive emotion).

Dimensional model has several advantages:

1. reliability to the vagueness of language
   In comparison to categorical approaches, the representation of the same emotion is always consistent despite of the difference of wording.

2. Impregnability to the limitation of language
   It is possible to represent emotions without using labels. Therefore, the

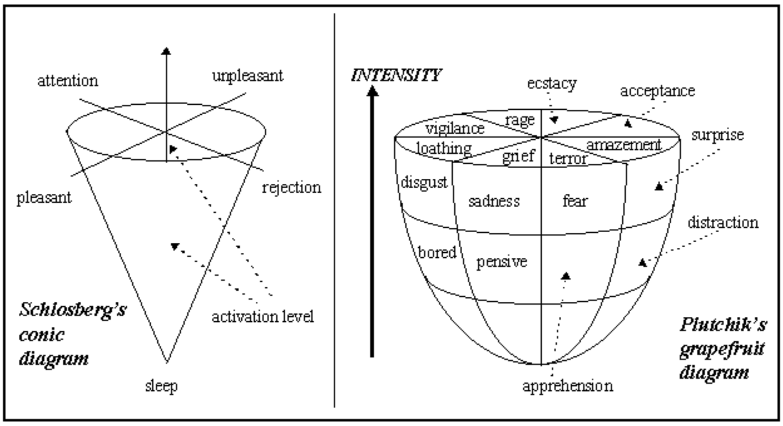


Figure 2.2: Scholosberg model (left) & Plutchik model (right) (Krech et al. 1974)

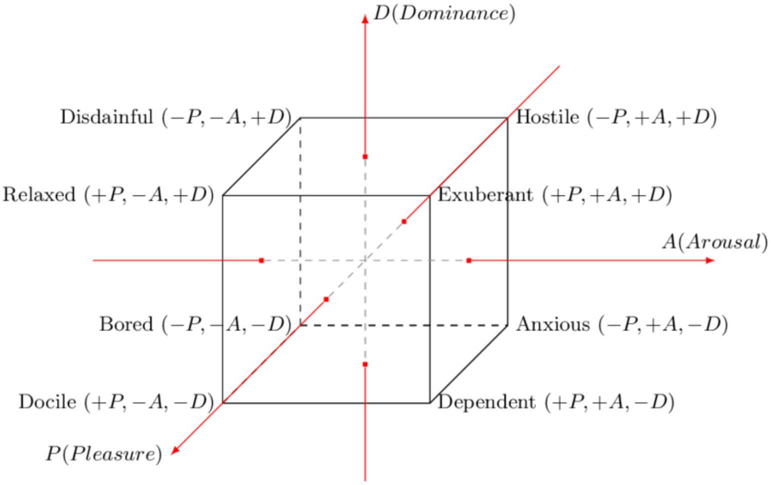


Figure 2.3: PAD model (Tarasenko 2010)

representation of an emotional state which is hard to describe is easily handled using the dimensional method.

3. Continuous and quantitative
   The change of emotion is continuous and there is not a clear barrier between the emotions. The dimensional approach makes the measurement of emotion changes possible.

4. Compatibility to categorical approach
   It is possible to associate dimensional representation with categorical ones. So the conversion from dimensional approach to categorical approach is easy, whereas the opposite conversion is difficult.

Despite these advantages, dimensional representation has received a number of criticisms:

1. Loss of information
   While dimensional approaches cover the full space of defined dimensions, some aspects of emotion which are beyond the defined dimension could be lost. This can result in an overlap of some emotions which share the same degrees on each dimension, but are totally different (for example,when using arousal-valence plan, one can hardly distinguish fear from anger, that is to say: part of information is lost. To solve this problem, one should add the neglected dimension - dominance (Mehrabian 1996)).

2. Omission
   Some emotions can be outside of the space of two or three dimensions (such as surprise, which could be either positive or negative).

(Reading 2004; Nicolaou et al. 2011) have shown that in the context of HCI or affects representation in everyday life, a single label or any small number of discrete classes may not reflect the complexity of affective states. Instead, in the dimensional approach, emotion transitions can be easily captured, and can be represented in continuous scales. Hence, a number of researchers advocate the use of dimensional description of human affect.

2.1.3 *Emotion classes in games*

Although the categorical and dimensional representations have their own advantages and drawbacks, the validity of different approaches is still controversial. They have been widely used as research target in the affective game domain, which are often referred to as classes or labels. The choice of representation approach depends strongly on the primary *purpose* and the *type of the game*. There are mainly two ***purposes*** of analyzing emotions in games: **evaluation** and **adaptation**.

For the **evaluation** purpose, the aim is to identify the emotions states within the game and to evaluate how they evolve in order to get better design of game setting. In this perspective, the emotions are closely related with game events and can be independently evaluated. Emotion categories closely related with game context are chosen. (Lazzaro 2004) evaluated the categorical emotions: fear, surprise, disgust, naches/ kvell (pleasure or pride at the accomplishment of a child or mentee), fiero (personal triumph over adversity), schadenfreude (gloat over misfortune of a rival competitive players enjoy beating each other especially a long-term rival), wonder using facial gestures, body language, and verbal comments. The author reported that by playing favorite games, participants enjoyed many emotions such as Fear and Surprise in Halo , the combination of Disgust with Naches in Odd World Schadenfreude and Fiero in head to head Top Spin Tennis, and Wonder in Myst's linking books. In another study, by using survey and statistical analysis method, (Bateman et al. 2011) evaluated the following emotions: positive emotions (contentment, relief, bliss), negative emotions (sadness, disgust, contempt, guilt, embarrassment), social emotions (gratitude, naches, envy, belonging), excitement (excitement, surprise), anger (anger, schadenfreude), curiosity (curiosity, wonderment), amusement, fiero to realize the player satisfaction modeling. Researches for evaluation purpose mostly contribute to game usability, game experience or player model.

For the **adaptation** purpose, the goal is to identify the important state which influences game experience and find an effective way to adjust the game settings. In this perspective, the important state can be determined by a series of evaluations and is more related with player's experience. (Csikszentmihalyi 1990) developed the fundamental concept of **Flow theory**, which described a balance between the inherent challenge of the game and the player's ability required to accomplish a task. High challenges provoke worry, anxiety while low challenges fail in engaging the player and evoke boredom. In general, people like being in a flow zone for love of the security and the hate of boredom. Inspired by the flow theory, a lot of works have considered only frustration, engagement and boredom (Chanel et al. 2008; Schwartz et al. 2017). Some have focus on different level of stress (Picard 1995; Rugg et al. 1995).

Another factor which influences the selection of the emotion representation is the **type of the game**. The possible induced emotion types depend on the game context, for example: a horror game are more likely to induce fear

Behavior	Physiological
• Vision-based – facial expressions – gestures/postures • Audio-based – voice modulation – dialogue with agent ... • HCI method: – Event log – Pressure on button	• Central nervous system – EEG electroencephalography • Peripheral nervous system – ECG Electro-cardiography – EOG Electro-oculography – EMG Electromyography – EDA Electrodermal activity – BVP Blood volume – Blood oxygenation – RESP Respiration – SKT Skin temperature

Table 2.2: Modalities for evaluating affective states

instead of giving player happiness (Watson et al. 1988). A Tetris (Fairclough et al. 2012) is less likely to evoke emotions as rich as in a FPS game (Dekker et al. 2007). Therefore, the selection of emotion representation approach in game research should take into account the game content.

2.2 EMOTION MEASURING MODALITIES

Methods of evaluating range from rigorous to casual, and can be qualitative or quantitative, subjective or objective, or hybrid (Mandryk 2005). The modalities to measure emotions can be classified into behavioral modalities and physiological modalities. Table 2.2 presents the methods used in each modality. In this section, a brief introduction of each modality is given as well as its advantages and drawbacks.

2.2.1 *Behavioral modalities*

Common behavioral modalities of the player when interacting with games are: vision-based methods, such as facial expressions, gestures/postures; audio-based method, such as voice modulation, dialogue with agent.

Facial expression

Facial expression information are acquired using camera or video camera, and the analysis is based on the static image, or the dynamic facial video. The pioneer work of (Ekman et al. 1997) formed the basis of automatic facial expression recognition systems. They created the facial action coding system (FACS) which classified human's facial expressions into many action units and described six basic facial emotions, joy, anger, surprise, disgust, fear and sadness. Since then a lot of effort has been made to build more reliable automatic facial expression recognition. There are basically two methods reported in the literature (Zavaschi et al. 2013), *geometry analysis* and *appearance-based method*. *Geometric analysis* takes into account some predefined fiducial points and refer their geometrical relation as features to represent facial expressions. The *appearance-based method* analyzes the faces through an holistic spatial analysis by using methods such as Principal Component Analysis (PCA), Independent Component Analysis, Gabor filters, Local Binary Patterns(LBP).

Though much progress has been made, recognizing facial expressions with a high accuracy remains difficult due to the complexity of facial expressions and the variability of individual and application context. (Krumhuber et al. 2013) reviewed the dynamic quality of the facial behavior. (Russell 1994) reviewed emotion from facial expression across different cultures and confirm the variation between different culture background. Others proposed method on specific groups of users such as children, patient with schizophrenia or autism (Edwards et al. 2002; Harms et al. 2010).

Body gesture/posture

Body gesture/posture represents positions of body joints and their changes with time. It can be obtain by a video camera, Kinect[1] or a motion capture system (e.g. VICON[2]). The emotion recognition is based on geometrical position or the motion of the body joints. (Kleinsmith et al. 2013) present a survey of affective body expression perception and recognition. (Glowinski et al. 2011) used upper-body movements to recognize valence and arousal from 10 actors. (Aigrain et al. 2015) used Kinect and took features such as quantity of movement, period of high body activity, posture changes, detection of self-touching to recognize different levels of stress.

[1] https://msdn.microsoft.com/en-us/library/microsoft.kinect.jointtype.aspx

[2] https://www.vicon.com/

Audio modalities

Audio-based information are acquired by using a microphone. The analysis include voice modulation and speech content analysis. The former focuses on using the acoustic features, prosody features, such as pitch variables ($F0$ level, range, contour and jitter), or speaking rate (Dellaert et al. 1996; Petrushin 2000). The latter focuses on the speech content, such as the selection of words, phrases and syntactic structures which can make lots of emotional expressions (Massaro et al. 1999).

As a whole, behavioral modalities based on video or audio information are non-intrusive and can provide rich intuitive information on the emotions. On the other hand, even though the visual expressions are intuitive to human, they are not efficient to machine-vision, for the large resources and processing powers it requires to process video stream. Thus, this method is time consuming and needs specialized model training (Kaplan et al. 2013). For example, facial expression modalities requires accurate and reliable facial feature detection and tracking, which is difficult to accommodate in many situations due to scale and resolution issues, illumination changes, pose variations. Moreover, their validity in some cases might be problematic in terms of subjective issues such as gender and personal character (Mauss et al. 2009) as well as cultures, races and social environments (Russell 1994). Also, inferring human emotions based on facial expression or body posture recognition may be problematic when emotions are intentionally expressed, suppressed or even hidden during the observation. Meanwhile, one should consider the application context of these modalities. For example, in a context of driving, drivers always don't speak and their emotional body gesture are limited, which makes the audio modalities and body posture/gesture modalities not applicable in this context. Also outward expressions such as facial expression, body languages or voice are less expected when playing computer game than in human to human interaction (Christy et al. 2014). That means we can not get much information about facial expression, posture or voice during playing.

In conclusion, behavior modalities even though rich and intuitive, are more applicable in expressive application context. One should also pay attention to its validity and the computing cost.

HCI modalities

HCI modalities refered to information generated during interaction with machine, such as event log, task performance, pressure on mouse/keyboards, frequency and speed of certain movement and other HCI patterns. This information can be gathered by using automatic or manual event logging (Nacke et al. 2008), or measured by pressure (Sykes et al. 2003b), gyro sensor or accelerometer.

For example, players have harder pressure on button and higher frequency of operation when their arousal level is high. Depending on the task, a number of task performance indicators can be used, such as task completion time, user's error rate, percent tasks completed, range of function used, etc (Mandryk 2005; Sweeney et al. 1993).

HCI information acts as an important complement for vision-based and audio-based modalities as they provide continuous, objective and quantitative measuring of player's behavior. However, their realization often requires extra work on programming the event logging tool (Nacke et al. 2008), installing sensors on mouse/keyboard and analyzing data which sometimes could be time consuming.

2.2.2 *Physiological modalities*

Physiological modalities include all the physiological signals or measures, they provides an objective, continuous, quantitative, real-time, sensitive way to assess the user's inner state (Kivikangas et al. 2011). Physiological modalities include *central nervous system* measurements and *peripheral nervous system* measurements. *Central nervous system* signals include electroencephalography (EEG), functional MRI (fMRI), functional Near-infra-red Spectroscopy (fNIRS); *peripheral nervous system* measurements include electrocardiography (ECG), Electro-oculography (EOG), electromyography (EMG), electrodermal activity (EDA), blood volume (BVP), blood oxygenation, respiration (RESP), skin temperature (SKT). Physiological signals are measured by special designed sensors and devices such as BioPac[3], BioRadio[4], NeuroSky[5]. Physiological data are translated to emotional states by extraction of features from measured physiological modalities. Next, the most popular features are selected in order to be used for emotion classification by means of various machine learning methods or statistical approaches.

Electroencephalography (EEG) and other methods such as MRI measure electrical activity caused spontaneously by functioning of the central nervous system (CNS). The measure of EEG requires special purpose scalp with many electrodes measuring the potentials provoked in different brain regions. It reflects directly the activity of the CNS and can contain rich information on

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

[4] http://glneurotech.com/

[5] http://neurosky.com/

the brain activity. It is demonstrated that the positive and negative emotions can be distinguished by evaluating the left and right frontal lobes' asymmetry (Li et al. 2009; Bos 2006). However, the calibration and measuring of EEG is time-consuming, as experimenter needs to verify the functionality of all the electrodes during the whole process, and the measures always contain noise triggered by artifacts such as eye blinking. Moreover, the measuring equipment is expensive.

Peripheral nervous system (PNS) activity is viewed as a major component of the emotion response in many recent theories of emotion. (Kreibig 2010) reviewed 134 publications that report experimental investigations of emotional effects on peripheral physiological and reported considerable PNS response specificity in emotion. For example, anger-eliciting contexts increased respiratory activity, particularly faster breathing; fear emotion caused cardiac acceleration and increased electrodermal activity. Sensor for PNS modalities measuring are less intrusive than the EEG measuring, meanwhile researchers are making an effort to make smaller, wireless wearable device to make the physiological signals measuring even less intrusive. Smart devices such as smart clothes, wristband, joystick (Sykes et al. 2003a; Christy et al. 2013; Bonarini et al. 2011; Oliver et al. 2006) or non-contact measurements using video processing technology (Lewandowska et al. 2011; Tan et al. 2010) have been able to non-intrusively measure signals from PNS (ECG, HR, EDA, etc.) which proves to be a good solution in a practical context.

In conclusion, compared with non-contact behavioral modalities, physiological measures required relevant sensors which are relatively more intrusive. However, they provide an objective, continuous, quantitative, real-time way to assess the user's inner state, and cost less computation resources (time series analysis cost less than video analysis). The development of the smart device also alleviate their weakness of intrusiveness, which makes them even better for practical affective computing application.

2.3 EMOTION ASSESSMENT APPROACHES

Emotion assessment is an importance issue in affective computing (Coan et al. 2007) that ranges from assessment approaches, applications to its understanding. Emotion recognition tries to find the relation between the different measuring modalities with the subjective self-reported emotion assessment. Even thought subjective self-reported is prone to bias due to cognitive error or memory limitations (Fairclough 2008), it is still viewed as an effective method of evaluating user's internal psychological state. Common emotion assessment of human computer interaction (HCI) typically include *questionnaires*, *interviews*, and *focus groups* (Mandryk 2005). In this section, we briefly introduce these methods as well as their advantages and drawbacks.

Questionnaire

Techniques such as questionnaire require from users to give their opinions or rate their experience through a series of statements and questions. Question types such as text, check box, multiple choice, list, scales can be used, which can cover wide range of questions. They are considered to be generalizable, convenient, amenable to rapid statistical analysis. With the development of web technology, on-line survey tools such as Google Forms[6], SurveyMonkey[7] are emerging which making its creation, distribution and analysis process even more convenient.

Some drawbacks of using questionnaires lie in the fact that results only involve asked questions and hence make finding hidden or complex patterns difficult; results may not correspond to the player's actual experience due to the fact that they can not recall all the details (Gow et al. 2010); results may also be distorted because the participants are inclined to cater to the experimenter, perhaps even without realizing it.

Interview

Interviews are about asking questions which involve one-to-one interactive contact with a participant. This is a more flexible method of gathering information which cover participant's opinion, perception, attitudes, thoughts, extracted from his/her answer or inferred from his/her behaviors.

Some drawbacks of interviews are: the task is time consuming; it is harder to analyze quantitatively, because of the unstructured nature of the resulting data; the answers might be incompleted because the participants don't recall all the details, or biased for some reason, or mis-interpreted by experimenter because of the perception differences (Oppenheim 1992); also interviewers must be careful to ask questions in a neutral, non-leading manner to avoid the potential bias (Isbister et al. 2008).

Focus group

Focus groups are a technique that involves bringing a small group of participants together with a moderator to discuss user needs and feelings (Nielsen 1994). Focus groups are sometimes preferred over interviews due to the time

[6] https://www.google.com/intl/en/forms/about/

[7] https://www.surveymonkey.com

saved by interviewing multiple people simultaneously, but also because of the spontaneous reactions and ideas that emerge through the participant's interactions (Nielsen 1994).

Limitations for focus groups are also that the results are always qualitative and subjective. In addition, participant's opinions may be swayed by other, more vocal participants in the group (Nielsen 1994).

Think-aloud

The basic approach in think-aloud studies is to ask participants to perform a given task and to verbalize their thought process while they proceed. Think aloud during task performance is called concurrent think-aloud, however this manner might influence task performance, change it, or distract user's attention (Isbister et al. 2008; Oppenheim 1992). For these reasons, some researchers advocate the use of retrospective think-aloud which allows participants watch their video recording during their task and try to verbalize the thoughts they had during the interaction.

Disadvantages think aloud protocols are: the process has a high cost in terms of time commitment; also, concurrent think-aloud introduces interferences to the users while retrospection think-aloud manner may lose some fidelity that would be present when discussing the task in real time (Mandryk 2005; Isbister et al. 2008).

In conclusion, concerning timeliness, think-aloud method allows to get in-time assessment, while questionnaire, focus group, interview provide posterior assessment. Concerning the information processing , interview, focus group, and think-aloud approaches contain rich information on the player's game experience but is time-consuming to process. On the contrary, questionnaire is given in the predefined format, thus more convenient for acquiring information and statistical analysis. Therefore, method such as interview can be applied in the preliminary study in order to explore all the interesting questions, and then contribute to make a relevant questionnaire for the formal study.

2.4 AFFECTIVE GAMING STUDIES REVIEW

Affective gaming refers to the new generation of games which will not only react to player's control, but also will take into account their emotional states, in order to adjust game plot accordingly, and offer more satisfactory gaming experience (Kotsia et al. 2013). In this section, we propose a review of the literature on how physiological signals are applied to affective gaming and try to figure out a research path for our work.

2.4.1 *Physiological response application in game*

Physiological responses reflect players' inner psychological state. Based on a review of commercial affective games (Kotsia et al. 2013), applications of physiological responses in game can be divided into two categories: *spontaneous control* and *automatic adaptation*.

In the use of *spontaneous control*, affective information are used as a novel input to control the game system. Biofeedback such as heart rate, electrodermal activity are used as criteria to adjust the game parameters. Therefore players can try to control spontaneously their emotion, in order to reach certain goal in game. For example, Brainball[8] is a two-player game using EEG singals to control a movement of a ball. The alpha and theta brainwaves which can move the ball forward are generated in the brain when one is calm and relaxed. A considerably stressed player will therefore lose. This game can help player to learn to regulate stress.

In *automated adaptation* context, affect information serves as supplementary information to adapt game automatically. The goal is to develop a video game that can adapt to suit individual players while they play in order to more effectively entertain them. Application includes: missile command (1980), Oshiete Your heart (1997), Left 4 Dead 2 (2008) which use the measures such as heart rate, sweat level to automatically recognize the psychological state adapt the game settings to improve the game experience.

In our work, we focus more on the automatic adaptation application, more specifically the emotion recognition system in this application.

2.4.2 *Adaptation for affective game*

Traditional approach of dynamic adjustment are based on performance, which are sometimes not applicable in the game context. One can still be interested in gaming when they are doing bad and can also lose interest when they are doing well. An emotion-centric adjustment could offer a more effective solution to this problem than that one based on performance in regard of an immersive and challenging gameplay. Dynamic Difficulty Adjustment (DDA) is a hot research topic for offering this sort of adjustment.

[8] https://www.tii.se/projects/brainball

The target of DDA ranges from game level, the AI of NPC to the game content in the level (Bontchev 2016).

(Tijs et al. 2008b) extracted features from behavioral (key-press force) and physiological responses (EDA, HR, RESP, EMG) and tried to realize DDA schema for different emotional states (bored / frustrated / enjoyed). Statistical test has been used to evaluate each feature in different modes.

(Liu et al. 2009) reported their efforts in developing a physiology-based affect recognition and real-time DDA in a closed-loop manner to allow a computer game to infer and respond to the affective state while interacting with the players. Features calculated on signals such as cardiovascular activity, EMG, EDA, temperature, and selected by correlation rate (> 0.3) have been used, and regression tree has been applied to identify 3 anxiety levels. After each epoch of 2 minutes, the difficulty level is changed based on performance (performance-based DDA) or anxiety level (affect-based DDA).

(Chanel et al. 2011) tested the validity to maintain player's engagement by adapting game difficulty according to player's emotions assessed from physiological signals. Questionnaire responses, electroencephalogram (EEG) signals, and peripheral signals of the players playing a Tetris game at 3 difficulty levels have been analyzed for different classifiers, feature-selection methods, and durations on which the features have been computed. A best accuracy of 63% was achieved by using signal fusion.

(Negini et al. 2014) create an affective game engine that uses affective states to adapt a FPS game. By using GSR signal and parameters decided in the prior test, equations for changing game parameters are given. Authors evaluated and compared the effects of design choice of adapting character, NPCs (Non-Player Character) and the environment. Results showed that affectively-adapting games were more arousing than the non-adapted version and adapting NPCs reduced player's enjoyment as it reduces the opportunity for the players to experience challenge.

While DDA has potential benefits, application of DDA has long way to go. Questions such as: what types of game elements should be adapted to affective state? When and how to adjust them? These questions remain to be solved by game designers.

As a first step of psychological state measuring, most of these work are dependent on the physiological modalities. Physiological-based emotion recognition for affective game and game adaptation has raised a lot of interest in the research area. Two relevant domains are non-intrusive physiological modalities measuring and physiological based emotion recognition model for game.

2.4.3 *Non-intrusive physiological modalities measuring*

In this field, efforts have been made to create less invasive devices, such as wearable systems or portable devices. These devices integrate different powerful sensitive sensor to evaluate player's affective state with minimum interference to their game experience. (Sykes et al. 2003a) made an analogue button on the game pad to measure finger pressure. Result showed that the game pads were pressed significantly harder in a more difficult level, thus proving the possibility to determine the level of a game player's arousal by the pressure they use when controlling the game pad. The author also asked about the possibility to detect valence through the player's use of game pad.

(Christy et al. 2013) propose a mouse capable of streaming real-time physiological data (pulse, EDA, SKT). They test the mouse on a custom game. The off-line analysis of amplitude of signals EDA, SKT and HR indicates the inclination to distinguish "calm" and "agitated" state. Making real-time classifier of state and integrating more affective modalities on mouse are raised for future research.

(Bonarini et al. 2011) developed a wearable device placed on the forehead to measure BVP, EDA and SKT during playing. Traditional data acquisition system requires the subject to be instrumented with sensors on fingers and chest, which could be invasive, and may affect both their performance and their emotional state. Authors claimed that a device on forehead can get the most interesting signals (such as BVP, EDA and SKT), won't occupy hands, and has relatively less impact on movement. This project focused on the wearability of the device and wearability impact on signal quality in order to develop easy to wear device which provides reliable and artifact-free signals.

To conclude, all these devices, ranging from control equipment such as game pad, mouse to wearable devices, are all dedicated to create hardware to access emotion information in a most viable and least invasive manner.

2.4.4 *Physiological-based emotion recognition model for game*

Compared with simulations such as image, music and video clips, video game provides a more active and dynamic emotional experience so that has several specialties. In the following section, we put forward the important elements of creating the emotion recognition system for game.

Physiological-based affective game is an interesting research topic (Bontchev 2016; Fairclough 2008). Table 2.3 presents key elements of physiological-based affective game research in the past decade. The comparing dimensions

Table 2.3: Physiological-based affective game review

Ref.	Game	Type	Modalities	Assessment	Time window (s)	Classification
(Toups et al. 2006)	PhysiRogue	2D action	EMG, EDA	stress level	1-300	Direct mapping
(Dekker et al. 2007)	Half-life 2	3D FPS	HRV(ECG), HR, EDA	horror	2	Direct mapping
(Tijs et al. 2008a)	Pacman	2D arcade	BVP, EDA, EMG, RESP, KeyB	boredom, frustration, enjoyment	180	Direct mapping
(Chanel et al. 2008)	Tetris	2D puzzle	EDA, BVP, HR, RESP, TEMP	boredom, anxiety, engagement	300	SVM
(Liu et al. 2009)	Pong	2D arcade	ECG, PPG, EDA, EMG, TEMP	anxiety level	120	RT, KNN, BNT, SVM
(Ayaz et al. 2009)	MindTactics	3D strategy	fNIR	attention level	16.5	KNN, NB
(Tognetti et al. 2010)	TORCS	3D car racing	BVP, ECG, EDA, RESP, TEMP	preference level	60	LDA
(Nacke et al. 2011)	FPS	2D FPS	BVP, EDA, ECG, EMG, RESP, TEMP	preference level		Direct mapping
(Fairclough et al. 2012)	Tetris	2D puzzle	EEG(Alpha&theta)	boredom, engagement, flow, overload	2	linear regression
(Liao et al. 2012)	Archery	shooting	EEG(Alpha)	focus level	2	Direct mapping
(Nogueira et al. 2013)	Slenderman	3D FPS	BVP, EDA, EMG	arousal & valence level	10-20	linear/non linear regression
(Parnandi et al. 2015)	Car racing	3D car racing	EDA	arousal	30	Direct mapping
(Emmen et al. 2014)	BioPong	2D arcade	EDA, HR	arousal		Direct mapping
(Negini et al. 2014)	Half-life 2	3D FPS	EDA	arousal		Direct mapping
(Tijs et al. 2008b)	Pacman	2D arcade	EDA, HR, RESP, EMG, key-press force	bored, frustrated, enjoyed		Direct mapping

we show include *game selection, modalities, emotion representation, time window*, and *classification*.

Game selection

Game selection refers to the selection of game used in the study. In the review work, there are simple game such as Pong, Pac-Man, Tetris, car racing game. These games provide relatively less possible game operations and generate small number of emotions. On the contrary, games such as strategy games or FPS provide more operation liberty and have the potential to generate more types of emotions which represent a better option to analyze physiological response of different emotions during game playing.

Modalities

Most physiological modalities presented in the literature are PNS, which confirmed the popularity of PNS over CNS in the affective game research. PNS modalities used in these literature are modalities related with cardiovascular system such as ECG, HR, BVP, HRV, Respiration, EDA and EMG. Among them, EDA and HR are the most frequently used signals. For measuring, the device can be custom self-made devices or the commercial systems such as ProComp Infinity or Biopac. The self-made devices are cheap and flexible while the commercial systems provide better quality for the signals.

Representation

The emotion representation largely depends on the genre of game and the research purpose. There are basically 3 types of representation methods:

- representation based on the emotion theory, using dimensional representation such as arousal/valence score (Nogueira et al. 2013; Parnandi et al. 2015) or using categorical emotions such as critical emotion in game such as horror, anxiety (Liu et al. 2009; Dekker et al. 2007);

- representation base on the flow theory (Csikszentmihalyi 1985) such as boredom, engagement and frustration (Tijs et al. 2008a; Fairclough et al. 2012);

- representation based on other dimensions such as preference, attention (Liao et al. 2012; Tognetti et al. 2010) which represent specific characteristics interested by the researchers.

Representation based on the emotion theory is the most frequently used and also the most flexible. Regarding the dimensional representation, as it covers a wide range of emotion space, it is applicable to most emotion evaluation cases. Concerning the categorical representation, a little difference from the emotion theory is that not all the basic emotions are used in the game research. Selection of the critical emotion is dependent on the type of game. For example, (Liu et al. 2009) only used "horror", as it is the most frequently occurred emotion in the game they used.

Representation based on the flow theory proposed a high level adaptation objective. Flow (Csikszentmihalyi 1985) describes a balance between the inherent challenge of the game activity and the player's ability to achieve the task. When skill required in game goes beyond player's skills, the task becomes too challenging and thus provoke anxiety and frustration. On the contrary, if the task is too simple, the game will fail to engage player and thus evoking boredom. The objective of flow-based adaptation is to try to detect and avoid the frustration or boredom emotion in game to make sure the player stays in the "flow" zone. However, video games often provide complex emotional experiences, so that the high level state of "flow" may depend on a series of low level events and emotions. An overall evaluation based on Flow theory often fail to reveal the mechanism how "flow" is generated. We can notice that the researchers who adopt this method of representation (Tijs et al. 2008a; Fairclough et al. 2012) have only applied adaptation on the simple games such as Tetris and Pac-man by adapting the speed parameter. When the game becomes more complex, the realization of "flow" state should be based on more detailed low level events and emotions.

Time window

Time window refers to the size of time window based on which the physiological signals are evaluated for emotion. It represents the temporal sensibility of the emotion recognition. Modalities such EEG or facial expression are reported to have good temporal sensibility (Ringeval et al. 2013). However, the response on PNS are relatively longer. There is still no consensus for the optimal time window for each PNS signal because of the complexity of the physiological signals and the variety of the application context. Based on a psychophysiological review (Kreibig 2010), the most frequently used time windows are 60, 30, 10s. In the Table 2.3, time window chosen for game research varies from 1s to 300s.

It should be mentioned that several researches applied direct mapping approach rather than emotion based approach. The direct mapping approach maps the physiological features to a certain game parameter directly in order to adapt them without recognizing the player's psychological state. This form of adaptation brings fun to the game by providing more randomness. The size of time window only control the frequency of adaptation and intensity of randomness which can have a wide range (Toups et al. 2006). However, this method is not robust as it fails to determine the real psychological state of the user. The emotion-based approach relies more on the appropriate window size for a better emotion recognition. Too short time window may fail to capture the physiological response, while the too long ones fail to capture the dynamic of emotional experience during the gameplay. An appropriate selection of time window is critical for the physiological-based affective game research.

Most emotion-based adaptive affective game research using PNS evaluate the game sequence as a whole (Chanel et al. 2008; Liu et al. 2009), which result in long evaluating time window and neglect dynamics of the emotion in game. As has been put by (Chanel et al. 2008), analysis of the physiological signals should also be conducted on the basis of the in-game events which can provides better sensitivity.

Classification

The emotion-based approach recognizes the player's psychological state using machine learning algorithms. The resulting output can be categorical by using the classification algorithms (Chanel et al. 2008; Liu et al. 2009; Tognetti et al. 2010) or numerical by using the regression model (Fairclough et al. 2012; Nogueira et al. 2013). The classification methods are more frequently applied. The selection of learning algorithm depends on the application context, the selection of signals, features and normalization method. Therefore, researchers often test several learning algorithms and select the one with the best performance.

2.5 CONCLUSION

In this chapter, we have presented some background knowledge involving emotion recognition: what to measure as emotions (emotion representation) and how to measure them (measuring modalities) and assess them (subjective assessment). We then present a literature review towards the critical elements of physiological-based affective game research. We investigate how the background knowledge involving emotion recognition have been used in the affective game research.

Based on the review above, we highlight the several points as follows:

- A lot of pioneering works have been dedicated to simple games such as Tetris. Their results can hardly be generalized to complex game with richer emotional experience. For evaluating richer emotions, one should choose a more complex video game. An ideal game should dispose the events which can generate different repetitive emotions, so that we have access to richer emotions and can use the event type as an objective emotion reference.

- PNS is a widely accepted form of measuring modality in affective game research. In order to alleviate the work of create self-made measuring device and ensure the quality of data collection, a reasonable choice will be using the commercial measuring system.

- Emotion representation with categorical method or dimensional method have both their strengths and weaknesses. In affective game research, both representation have been applied as they offer flexible evaluation of the participants emotion. Concerning categorical approach, the critical categorical emotions related with game should be decided. Concerning the dimensional approach, arousal/valence representation is the most widely used one. One should choose the appropriate range of measurement by finding a balance between precision and convenience.

- Flow theory based emotion representation may not be able to apply directly on the complex video game, as the high-level "flow" is dependent on a series of different low-level event or emotion. A wiser choice to evaluate the emotion is from the both minor scope and overall scope.

- In order to determine the psychological state, a safer granite for the time window should be no more less than 10 s. It can't be too long neither in order to capture the dynamics of the emotion in game.

- The selection of learning algorithm is largely dependent on the data. Therefore one should try different algorithms to settle the most appropriate one for the given problem.

- Among all the reviewed work on physiological-based affective game, none of the them realized personalized optimization for the model, which can be a promising research path.

- Among all the reviewed works on physiological-based affective game, none of them have evaluated the computation resource required. Emotion recognition system is an important auxiliary module for the realization of adaptive affective game. The computational resource evaluation for the overall process is important for the final implementation on an embedded system.

3 EXPERIMENT ON GAMERS FOR AFFECTIVE DATA COLLECTION

The creation of emotion recognition model should be based on high quality context related data. Even though there has been a growing number of public physiological-based affective computing databases, by far, to the best of our knowledge, none of them are acquired under video game context. The usability of a database to an application depends on its characteristics. In this chapter, we first review several most popular affective computing databases using physiological signals as modalities by highlighting their characteristics (Section 3.1). Then, we clarify the characteristics of the database needed for our research and present how we run an experiment to collect the data (Section 3.2). In the end, we propose a statistical analysis of the acquired database (Section 3.3).

3.1 REVIEW OF THE EXISTING AFFECTIVE DATABASE

The research interest in affective computing has motivated the creation of novel databases for affective computing. In this section, we present some related affective computing work and affective databases in the following aspects: (i)Affective stimulation, (ii)Objective modalities, (iii)Assessment, (iv)Time-scale. Table 3.1 summaries the databases related to our context. In the end, we highlight the characteristics of the proposed DAG database.

3.1.1 *Affective stimulation*

Affective stimulation refers to the material used to stimulate emotion. In affect computing research, lots of efforts have been dedicated to effective emotion stimulation (elicitation) (Coan et al. 2007). Stimulation can be further categorized as *social/individual*, *spontaneous/posed*. In social situations (e.g. job interview, debating, conversation, speech), subjects have a strong need to express themselves, consequently their emotional state can be deduced from expressive modalities such as facial expression, gesture, intonation. Many such databases provide speech, visual, or audiovisual data in natural interactions such as conversation and TV talk show (to name a few: Belfast database (Schröder et al. 2000), Vera am Mittag (VAM) (Grimm et al. 2008), HUMAINE

Table 3.1: Summary on existing affective databases

Database	Part.	Stimulation				Objective modalities		Assessment			Time-scale		
		Type	Ind.	Spon.	Nat.	Type	Physio.	Type	self.	obser.	seg.	event.	real.
Healey (Healey 2000)	1 (20 days)	"Sento-graph", self	✓	(✓)	✗	PPS	✓	categ.	✓	✗	✓	✗	✗
DEAP (Koelstra et al. 2012)	32	Music video	✓	✓	✗	EEG, PPS, video	✓	AVD, like, familiarity	✓	✗	✓	✗	✗
MAHNOB-HCI (Soleymani et al. 2012)	27	Video clips	✓	✓	✗	EEG, PPS, eye gaze, audiovisual	✓	AVD, categ., predictability	✓	✗	✓	✗	✗
DECAF (Abadi et al. 2015)	30	Music video, movie clips	✓	✓	✗	MEG, PPS, video	✓	AVD	✓	✓	✓	✗	✓
RECOLA (Ringeval et al. 2013)	46	Meeting	✗	✓	✓	PPS, audiovisual	✓	AV, social info	(✓)	✓	✓	✗	✓
PED (Karpouzis et al. 2015)	58	Game: Super Mario	✓	✓	✓	game content, behaviour, visual	✗	experience	✓	✗	✓	✗	✗
Mazeball (Yannakakis et al. 2010)	36	Game: Mazeball	✓	✓	✓	camera setting, PPS	✓	experience	✓	✗	✓	✗	✗
DAG	58	Video game	✓	✓	✓	PPS, behaviour, video	✓	AV, experience	✓	✗	✓	✓	✗

Comparison in terms of **Stimulation** type, individual/interactive, spontaneous/posed, natural/induced), **Objective modalities** (type, whether contain physiology modality), **Assessment** target (type, self-reported assessment, observed assessment), and **Times-scale** (segment-based, event-based, real time).
Notations: ✓signifies satisfied, ✗signifies not satisfied, (✓) signifies partially satisfied; **PPS** signifies Peripheral physiological signals; **AV(D)** signifies Arousal, Valence, (Dominance) respectively, categ. signifies categorical representation of emotions.

database (Douglas-cowie et al. 2007), RECOLA database (Ringeval et al. 2013)) . On the contrary, in individual situations, expressive modalities are less obvious. Physiological signals overcome the deficiencies of expressive modalities in these circumstances. Examples of existing databases for this case include driver database at MIT (Healey et al. 2005), video viewing or music listening databases as DEAP, MAHNOB-HCI, DECAF (Koelstra et al. 2012; Soleymani et al. 2012; Abadi et al. 2015). The *spontaneous/posed* dimension describes whether the emotion is spontaneous or acted deliberately. DEAP, MAHNOB-HCI, DECAF, and RECOLA (Koelstra et al. 2012; Soleymani et al. 2012; Abadi et al. 2015) belong to the spontaneous category while Healey's (Healey 2000) database belongs to the posed category. The *induced/natural* dimension indicates whether the emotion is induced under controlled settings or during naturalistic interactions. Database DEAP, MAHNOB-HCI, DECAF (Koelstra et al. 2012; Soleymani et al. 2012; Abadi et al. 2015) using predefined stimuli to induce specific emotion belong to the *induced* stimulation category. While RECOLA database (Ringeval et al. 2013) created during a collaborative work meeting belongs to the *naturalistic* stimulation category. *Induced* stimulation provides a controlled experimental setting, of which all stimulations are predefined and well separated. This setting makes the emotional responses more predictable and easier to analyse. However, the induced manner can be questioned for its practicality, as emotions in real life are complex and evolve all the time. Their complexity and dynamics make it "tricky" but also interesting to analyse them.

3.1.2 *Objective modalities*

Objective modalities are the ones that are measured from objective measuring instrument. They can be used as clues to deduce emotion. Measurement of emotion has been extensively investigated in affective computing. Vision-based modalities (facial expressions, body posture and eye movement) (Krumhuber et al. 2013; Glowinski et al. 2011), speech modalities (Dellaert et al. 1996; Petrushin 2000) have been investigated to detect emotion in different context.

Despite of the unobtrusiveness and effectiveness of these modalities in the given context, they might not be applicable to digital games. Currently available vision-based system cannot operate well in real-time (Zeng et al. 2009). As it requires reliable detection and tracking of target, which is difficult to accommodate in game context due to scale and resolution, illumination changes, pose variations issues. Moreover, in the gameplay context, players tend to stay still and speechless while playing games (Yannakakis et al.

2016), so that vision-based and speech modalities cannot provide an effective measuring.

Physiological signals, on the other hand, provide objective, continuous, quantitative measure which can reflect the human emotional state in the game context. Signals from the central nervous system (CNS) such as Electro-EncephaloGram (EEG) or Magneto-EncephaloGram (MEG) have been successfully used to reflect the complex activity of the brain and evaluate emotional states under different contexts such as video music, film clips (Koelstra et al. 2012; Soleymani et al. 2012; Abadi et al. 2015). Measurement of EEG requires careful placement of electrodes and calibration of channels which may sometimes affect participants' immersion. MEG is a non-invasive technology which measures brain activity with higher spatial resolution. However, the equipment is heavy and expensive which is not practical in real-life emotion recognition. On the contrary, smart devices such as smart clothes, wristband, gamepad (Sykes et al. 2003a; Christy et al. 2013; Bonarini et al. 2011; Oliver et al. 2006) or non-contact measurements using video processing technology (Lewandowska et al. 2011; Tan et al. 2010) have been able to non-intrusively measure signals from peripheral nervous system (PNS) (ECG, HR, EDA, etc.) which proves to be a better solution in a practical context. Besides the user-based modalities, content-based modalities can also be included. Multimedia content analysis (MCA) (Koelstra et al. 2012) provides features based on content. Here, the content can be viewed as an objective supplementary evaluation of emotion.

3.1.3 *Assessment*

Assessment refers to the affective label given to the stimulation. Assessment can be self-reported or observed. The method of assessment can be ratings (Koelstra et al. 2012; Soleymani et al. 2012; Abadi et al. 2015) or ranking (Holmgård et al. 2015; Yannakakis et al. 2010; Karpouzis et al. 2015).

Concerning the assessment, even though researches on emotion have been flourishing for long, yet debate continues about the nature of emotions, their biological mechanisms, their categories (Izard 2007; Scherer 2005). Among these topics, emotion representation is one of the fundamental questions of emotion researches. There are many methods to model emotion. The two most widely accepted approaches for modelling emotions are the *categorical approach* and the *dimensional approach*.

Categorical approach claims that there exist a small set of basic emotions. A few examples of categorical approach are: Ekman's six basic emotions (Ekman 1993) (anger, disgust, fear, joy, sadness, surprise), his latter work has included more emotions such as excitement, amusement, fiero and sensory pleasure (Ekman et al. 2011), Flow theory (Csikszentmihalyi 1996) (frustration, boredom, engagement) (see (Ortony et al. 1990; Tracy et al. 2011) for a more detailed review). Categorical representation has been used in database (Soleymani et al. 2012; Healey 2000).

Concerning dimensional approach, there are Russell's original two dimension (arousal/valence) (Russell 1980b), Wastson's PANAS scales (Watson et al. 1988) (positive/negative affect), Evaluative Space Model (Norris et al. 2010) (positive/negative affect and offset). Dimensional representation has been used in database (Koelstra et al. 2012; Soleymani et al. 2012; Abadi et al. 2015).

In affective computing, the emotion representation can be model-based or model-free (Yannakakis et al. 2014). While many studies take the model-based approach (see (Kreibig 2010; Mauss et al. 2009)), in which emotional measures are mapped directly to specific emotional states given by a theoretical framework, empirical affective computing works are based on model-free approach. Model-free approaches refer to the construction of an unknown mapping (model) between input and an emotional state representation (Yannakakis et al. 2014).

In the game context, player data and emotional states assessment are collected and methods such as classification, regression and preference learning techniques adopted from machine learning or statistical approaches are used to derive the model (Martínez et al. 2011; Yannakakis et al. 2010). The emotion model for the game can be derived from emotion theory (Ravaja et al. 2006), flow theory (Karpouzis et al. 2015; Tijs et al. 2008a; Chanel et al. 2008), or the specific assessment dimension of game experience (Yannakakis et al. 2010; Toups et al. 2006; Dekker et al. 2007).

3.1.4 *Time-scale*

Time-scale refers to the time interval for evaluating the emotion. The related physiological signals and assessment are generally analysed on 3 different time-scale: real-time (Ringeval et al. 2013; Nogueira et al. 2013; Toups et al. 2006; Dekker et al. 2007; Tijs et al. 2008a), event-based (Holmgård et al. 2015; Kivikangas et al. 2011; Martínez et al. 2011; Ravaja et al. 2008; Ravaja et al. 2006), segment-based (Chanel et al. 2008; Liu et al. 2009; Tognetti et al. 2010).

In real-time scale, physiological-based indicators are calculated in real-time. Physiological signals always take more time to bring about expressive emotional response than expressive modalities such as facial expression (Ringeval et al. 2015a). In DECAF, continuous arousal/valence(AV) annotations on film clips are provided by 7 experts, and MEG and MCA features are used to learn multi-task learning (MTL) based regressors. The RECOLA database takes the idea of continuous annotations and applies it to a naturalistic interaction in a collaborative video conference. Seven observers annotate the first five minutes interaction of 46 participants in terms of AV. Obviously, annotations from observers are based on expressive modalities which neglect subjects' real internal feeling. Even though some of the real-time scale evaluation result in real-time evaluation of the emotional state (eg. arousal/valence) (Ringeval et al. 2013; Nogueira et al. 2013), many others only aim to provide a quantitative indicator (such as stress level) to realize a direct mapping in order to adjust game settings accordingly and provide a novel experience (Toups et al. 2006; Dekker et al. 2007; Tijs et al. 2008a).

In event-base scale, psychophysiological responses are evaluated based on game event. A game event is instantaneous and the effect can last several seconds. Ravaja et al. (Ravaja et al. 2006) chose the game Monkey Ball 2 and examined arousal/valence-related phasic psychophysiological responses on signals (zygo-maticus major, corrugator supercilii, and orbicularis oculi electromyographic activity (EMG), skin conductance level (EDA), and cardiac interbeat intervals (IBI)) to different video game events. Holmgard et al. (Holmgård et al. 2015) profile the stress responses on the EDA of patients diagnosed with post-traumatic stress disorder (PTSD) to individual events in the game-based virtual environment StartleMart. The analysis on event-based scale result in more targeted psychophysiological response while maintain the time sensibility in the dynamic context.

In segment-based scale, the analysis is based on a overall game segment which can sometimes last for several minutes. DEAP, MAHNOB-HCI, DECAF use one assessment for the whole sequence. The segment-based evaluations are always aiming for getting overall game experience. For example, (Chanel et al. 2008) explore the correlation between boredom, anxiety, engagement and several physiological signals (heart rate (HR), blood volume pulse (BVP) and skin conductance (EDA), respiration (RESP), temperature (TEMP)) while playing Tetris for a game length of 5 min. (Tognetti et al. 2010) investigate the correlations between preference level and BVP, ECG, EDA, RESP, TEMP on a 3D car racing game TORCS for a game length of 1 min. The segment-based scale evaluation emphasizes the overall player game experience.

Our database

Compared to existing databases, our required database using for affective game should have the following characteristics:

1. stimulation: individual, spontaneous, naturalistic;
2. objective modality: peripheral physiological signals, accelerometer, face/game recording;
3. assessment: self-reported categorical/dimensional emotion and game experience;
4. time-scale: event-based, segment-based.

In the following section, we present how we conducted an experiment to collect data for affective game research.

3.2 THE EXPERIMENT

The objective of this experiment[1] is to collect data in order to analyze the relationship between measured modalities and subjective assessments.

3.2.1 *Preliminary work*

Preliminary work includes the selection of a video game as stimulation, the recruitment of the participants, and settling the assessment dimension. Data from the pre-study is also made available.

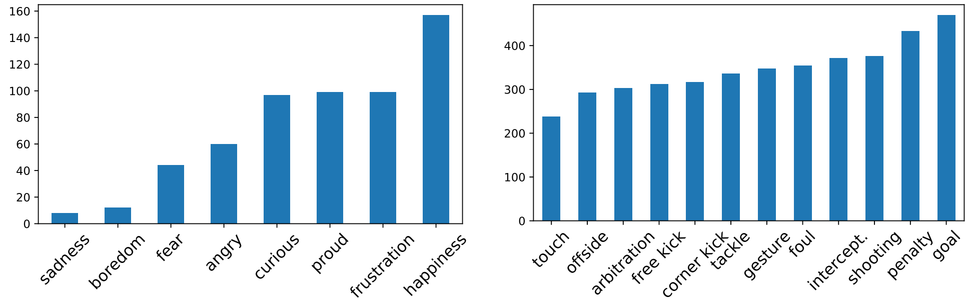
3.2.1.1 *Game selection*

The football simulation game FIFA 2016 by Electronic Arts was chosen to set up the experiment for the following reasons:

(i) it has a wide range of players so that it's easy to recruit adequate players and gather data from players with different skill levels;

(ii) short repeated event sequences may potentially generate different emotions;

[1] This experiment has been done thanks to a fund "Soutien au démarrage d'études comportementales" granted by the Centre Multidisciplinaire des Sciences Comportementales Sorbonne Universités INSEAD.



(a) Typical emotions in game

(b) Typical events in game

Figure 3.1: Typical emotions and events in game collected from preliminary study.

(iii) event types are often easily assessable from the game system which can be used as an effective reference for player's emotional state (e.g. a scored goal is mostly related with happy emotion);

(iv) level difficulty of the game can be easily changed by game setting thus being able to offer different experiences.

The FIFA game generates relatively rich emotions while disposing of limited range of experienced emotions. In dynamic game context, while the physiological-based measuring of simple emotions is still challenging, a reasonable choice is to limit emotions experienced to a small range. By setting in a sport game context, we aim to analyse various emotion components generated in a dynamic context and related to events.

3.2.1.2 *Participant recruitment*

For participant recruitment, a pre-study was conducted by gathering responses from an on-line questionnaire concerning participant's playing habit, skill level, emotional events and common emotions during the game. This pre-study was used to select qualified players. We excluded the players who had never played a football simulation game, as their emotions are more probably likely to be influenced by factors other than game event.

3.2.1.3 *Assessment*

We aim to collect assessment both on the game event and the game segment. Regarding the emotion assessment on the game event, the information we want to collect is:

1. Event type, which triggers emotion. We limit the option of the event types to the event list which has been collected from the recruiting questionnaire. The selected events reported by participants were those which have elicited the most emotions in the FIFA game. These events are goal, penalty, shooting, interception of guard, foul, gesture, tackle, corner kick, free kick, arbitration, offside, and touch. Figure 3.1b presents the frequency of each event mentioned by the people who have reply to our questionnaire. The "missing" action has often been mentioned in the formal annotation, therefore has been added later.

2. Categorical emotion associated with the event. We asked the participants to select from a predefined categorical emotion list: the 6 basic emotion theory (Ekman 1993), and the flow theory (Csikszentmihalyi 1996). These categorical emotions are happiness, frustration, proud, curious, angry, fear, boredom, sadness. Figure 3.1a presents the frequency of each categorical emotion mentioned by the people who have reply to our questionnaire.

3. Dimensional emotion associated with the event. We proposed to participants to evaluate arousal/valence dimensions, as it was commonly used in (Koelstra et al. 2012; Soleymani et al. 2012; Abadi et al. 2015; Ringeval et al. 2013).

Concerning the game experience assessment, Game Experience Questionnaire (GEQ) (IJsselsteijn et al. 2007) has been developed in the game research domain. In the GEQ, the measured experienced dimensions are competence, challenge, immersion, flow, tension, negative affect, positive affect. Among these dimensions, because of our game context, we consider that competence and challenge are both related to "difficulty"; flow and immersion are closely related; tension, negative and positive affect vary during gameplay so they are not relevant for the post-game evaluation. In addition, we consider that the "amusement" dimension is relevant to evaluate the overall appreciation of the game. The final dimensions we took were: difficulty, immersion, and amusement (DIA). Based on the work (Yannakakis et al. 2011), both rating and ranking were used to evaluate the game experience.

3.2.2 *Modalities and measuring equipments*

To analyse physiological responses during game playing, the following modalities are recorded.

Physiological signals were collected using the BioNomadix wireless sensors and physiology monitoring system Biopac MP150[2]. A sampling rate of 1000Hz has been chosen, as high sampling rate is recommended for measuring heart rate variability (HRV) from ECG. Figure 3.2 presents the placement of sensors. The sensors used are (Figure 3.2) :

- an ECG sensor with 3 pre-gel electrodes to measure electrocardiogram (Figure 3.2(a));
- an EDA sensor with 2 pre-gel electrodes to measure electrodermal activity (Figure 3.2(b)) (in order to alleviate impact of artifact caused by controlling the joystick using hand, the EDA electrodes are placed on foot);
- a respiration belt to estimate chest cavity expansion (Figure 3.2(c));
- two Electromyogram (EMG) sensors, each with 2 pre-gel electrodes to measure - Zygomaticus and corrugator muscles movement Figure 3.2 (d);
- a temperature sensor placed on back neck;
- an accelerometer sensor installed on back neck to measure body movements.

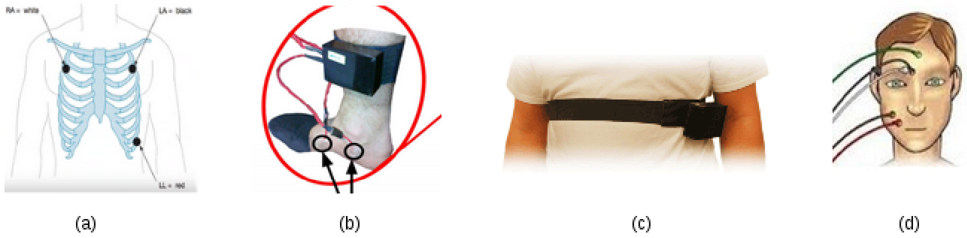


Figure 3.2: Placement of sensors (a) ECG sensor (b) EDA sensor (c)respiration belt (d) Electromyogram (EMG) sensor

[2] https://www.biopac.com

Facial recording

Facial recordings were collected using a web camera placed at the top middle edge of the screen. For privacy issues, we are not able to make publicly available this facial recordings, however the analysis result of facial coding given by FaceReader[3] is made available. The analysis result contains the presence of certain basic emotions (happy, sad, angry, surprised, scared, disgusted) and neutral, as well as the arousal and valence score.

Game screen recording

Game screen recording were collected using MediaRecorder[4]. These recordings were reviewed by participants to annotate their emotions. Researchers can use these recordings to verify a particular moment during match.

Meta-information

Meta-information such as player skill level, game difficulty level and game resulting score are also noted. They may be useful to give background knowledge for analysis.

The game screen output, webcam recording, and the screen containing the physiological data were synchronized using a software ObserverXT[5] and visualized on the same screen (Fig. 3.3) for experimenter.

3.2.3 *Experimental Protocol*

The experiment was conducted at INSEAD - Sorbonne University Multidisciplinary Centre for Behavioural Sciences[6]. In total, 58 participants (50 males, 8 females; mean age 25; all right handed) of different skill levels took part in this study. Each participant started the experiment with a physiological calibration session and a brief introduction to experiment protocol. Then the recording of different modalities of signals began. The procedure of the experiment is presented in Fig. 3.4.

Participants played the game in an isolated environment. Each experiment was composed of 4 phases: one training phase and 3 match phases. In the training phase, the participant configured the joystick to his/her customary

[3] http://www.noldus.com/facereader

[4] http://www.noldus.com/human-behavior-research/products/media-recorder-0

[5] http://www.noldus.com/the-observer-xt

[6] http://centres.insead.edu/sorbonne-behavioural-lab/eng/index.cfm

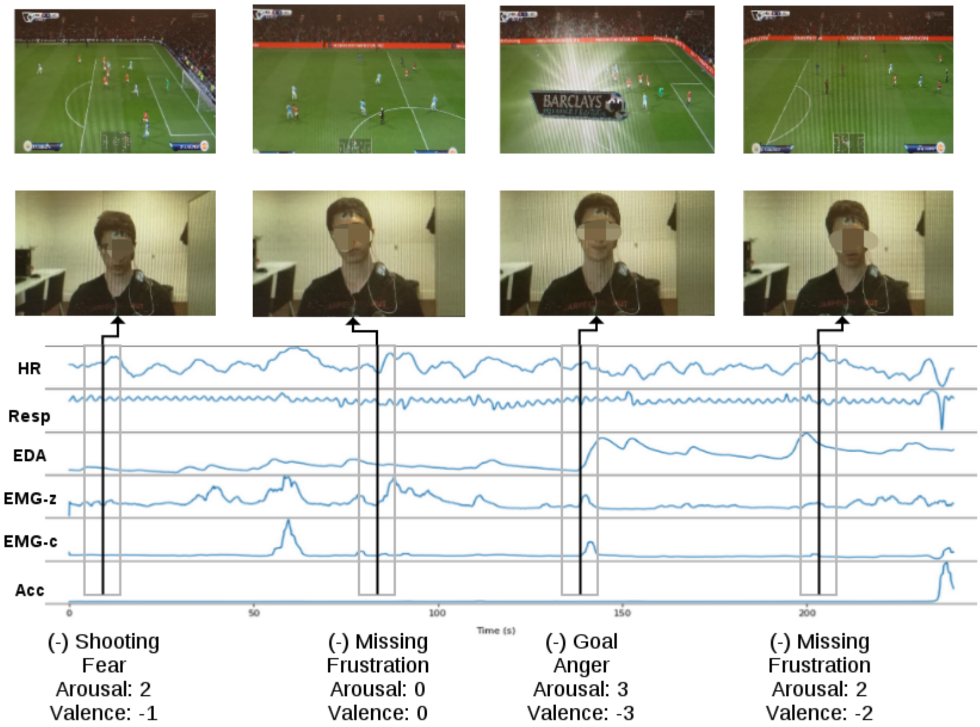


Figure 3.3: Experiment scene

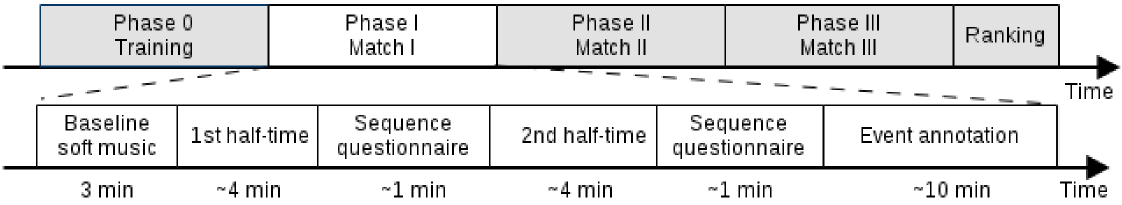


Figure 3.4: Protocol of experimentation.

settings and familiarized himself/herself with the experiment system. The skill level of the participant was evaluated by the experimenter based on his/her behaviour and performance during the training phase. Then 3 matches were presented to the participant: one of higher level, one of equal level and one of lower level. These three matches were presented in a random order to avoid sequence effect.

Each match began with 3 minutes of soft music, during which the participant got relaxed. This period allowed the participant to return to neutral state prior to each phase (Fig. 3.4 “Baseline soft music” phase). Then, the participant played two half-time (4 minutes for each) of game (Fig. 3.4 “1st half-time” and “2nd half-time” phase). The score of each half-time was also saved.

After each half-time, the participants filled out a game sequence questionnaire to evaluate his/her feeling about game in terms of difficulty, immersion, amusement (DIA) (Fig. 3.4 “Sequence questionnaire” phase). The scores of each team were also noted by the experimenter.

At the end of each match, the participant viewed the recording of the match and annotated the felt emotions (AV ratings and frustration, happiness, anger, fear or boredom) triggered by significant events (missing, goalkeeper’s interception, goal, technical gesture, shoot) during the game (Fig .3.4 “Event annotation” phase).

In the end, the participants ranked the 3 matches in terms of DIA (Fig. 3.4 “Ranking” phase).

The physiological signals are recorded during the whole sequence of experiment. Each participant was remunerated 20 euros for about 2 hours of experiment. The database has over 116 hours of logged events, game play, participant video and physiological data.

3.2.4 *Participant self-assessment*

As described above, the participant self-assessment is composed of three parts:

- **game event annotation**, during the viewing of the screen recording, contains game events and associated instant emotions;
- **game sequence questionnaire**, after each half-time, contains global evaluation of a sequence of match;
- **game ranking questionnaire**, at the end of the experiment, contains the ranking of the 3 matches in terms of DIA.

Detailed descriptions of **game event annotation** and **game sequence questionnaire** are presented below.

3.2.4.1 *Game event annotation*

At the end of each match, the screen recording of the entire match was replayed, and the participant was asked to recall and annotate the significant events and associated emotions. For each critical event, he/she gave the event name, as well as a discrete emotion and an arousal/valence note (in a range of [-3,3]) related to this event. For the convenience of annotation, we offered participants a list of emotions and events. These lists drawn from the pre-questionnaire during the recruitment process contained the most frequent emotions and events in the game. The elements included in each annotation are:

- events: missing action, goalkeeper's interception, goal, technical gesture, shoot;
- emotions: frustration, happiness, anger, fear, boredom;
- arousal/ valence score: -3, -2, -1, 0, 1, 2, 3.

To speed-up the annotation process, the experimenter helped the participant to annotate events with corresponding time-stamps using the software ObserverXT.

3.2.4.2 *Game sequence questionnaire*

Game sequence questionnaires are presented at the end of each half-time in order to evaluate the global feeling during this sequence. Evaluation items include:

- a global note for experience $(1-10)$;
- whether the participant felt emotion changes during the match (yes/no);
- whether the participant is able to indicate the critical event and emotion (yes/no);
- evaluation of the experience in terms of three dimensions DIA using *explicit* and *implicit* method.

In *explicit* method, the participants evaluate the DIA by giving a score (discrete scale $[1,5]$, with 1 representing lowest level of associated dimension and 5 representing highest level). As these ratings are based on scores, we refer them as: **r**ating from **s**core-based **d**ifficulty (*rsD*), **r**ating from **s**core-based **i**mplication (*rsI*), **r**ating from **s**core-based **a**musement (*rsA*) In *implicit* method, participant evaluates the DIA by answering the following questions also in a 5-point scale $[1,5]$ with 1 corresponding to "Strongly Disagree" and 5 corresponding to "Strongly Agree":

- difficulty
  - Q1: I felt competent in the game
  - Q2: I felt frustrated
- implication
  - Q3: I was concentrated on the game
  - Q4: I thought about other things
- amusement
  - Q5: I think the game was interesting
  - Q6: I felt happy

Based on the positive or negative relation in each dimension, the score of each question is defined as:

$$score_{Qi} = \begin{cases} r & \text{if positive relation} \\ 5-r & \text{if negative relation} \end{cases}$$

where $i = 1,2,...,6$, and r signifies the rating for the question Q_i. We define the evaluation of each DIA dimension as the mean value of the corresponding scores. 3 different ratings: **r**ating from **q**uestion-based **d**ifficulty (*rqD*), **r**ating from **q**uestion-based **i**mplication (*rqI*), **r**ating from **q**uestion-based **a**musement (*rqA*) are calculated in the following way:

- difficulty: $rqD = (score_{Q1} + score_{Q2})/2$
- implication: $rqI = (score_{Q3} + score_{Q4})/2$
- amusement: $rqA = (score_{Q5} + score_{Q6})/2$

The reason why we take both explicit and implicit measures is that multi-evaluation makes the subjective self-assessment more consistent and more reliable (IJsselsteijn et al. 2007).

All questionnaires were filled out on-line using a desktop computer to minimize the effects of interaction with experimenters.

3.3 STATISTICAL OVERVIEW

A summary of the DAG database is presented in Table 3.2. In this section, we present some statistical analysis on the collected data.

3.3.1 *Self-assessed annotations on events in game*

In this section, we give a brief statistical analysis on self-assessed annotations. First, we present the relation between the game events and the associated emotions. Then, we present the distribution of the annotated events in terms of categorical and dimensional representation. Finally, we present the distribution of categorical emotions on the dimensional plan.

Different events trigger different emotional responses and can serve as a reference for emotional state inference. We annotated 5 types of game event: missing, goalkeeper's interception, goal, technical gesture and shoot. We distinguish the event as positive or negative. Each event is beneficial either to the player (positive event) or to the adversary (negative event). For example, positive goal is a goal scored by the participant (noted as $(+)$ *goal*), while negative goal represent a scoring by the adversary (noted as $(-)$ goal). Fig. 3.5 presents the frequency of categorical emotions on different events. We notice that most positive events correspond to "happiness", while most negative events result in a negative emotion: "(-)missing" and "(-)goalkeeper's interception" are mostly associated to "frustration"; "(-)goal" is mostly related to "anger" and "(-) shooting" mostly related to "fear". These observations are consistent with our intuition. As the event type is often easily assessable from the game system, it can be used as an effective reference for player's emotional state.

Fig. 3.6 presents the distribution of the emotional responses in terms of categorical and dimensional emotions. The most frequent categorical emotions

Table 3.2: DAG Database summary

Participants and modalities	
Nb. of participants	58 (50 males and 8 females)
Recorded signals	ECG, EDA, respiration, EMG, skin temperature, accelerometer, Face video, Game screen recording
Emotional response on game events (Study I)	
Nb. of events	1730
Evaluation approach	annotations given by participant
Self-report	event, arousal/valence score, categorical emotion
Rating value	Event: missing, goalkeeper's interception, goal, technical gesture, shoot Arousal /valence: -3,-2,-1,0,1,2,3 Categorical emotion: frustration, happiness, anger, fear, boredom (one for each event)
Global game evaluation (Study II)	
Nb of matches	174 matches (348 half-time ($\sim$ 4 min))
Evaluation approach	questionnaire on DIA scores questionnaire on DIA ranking
Rating value	scores in the scale [1, 5] ranking the 3 matches

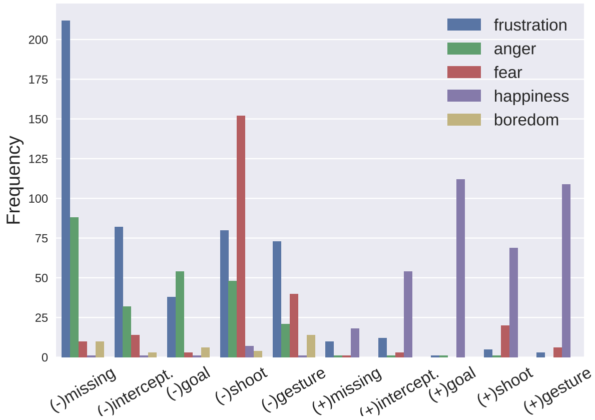


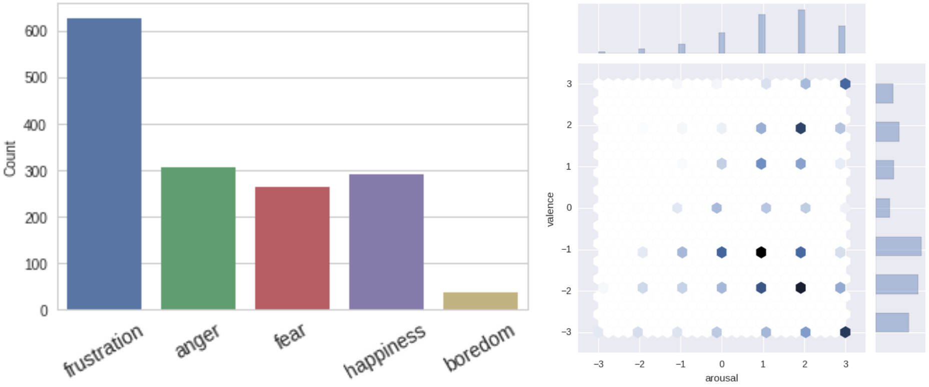
Figure 3.5: Distribution of categorical emotions for each event.

are "frustration" followed by "happiness", "anger", "fear" and "boredom". In the dimensional emotion representation, the most annotated events are events with high level of arousal and low level of valence (HALV) followed by events with high level of arousal and high level of valence (HAHV), and then the events with low level of arousal and low level of valence (LALV). Negative emotion is much more frequently annotated than the positive emotion. This can be explained by the psychological phenomenon that bad things have strong impact on us than the good ones (Baumeister et al. 2001). Events with low level of arousal and high level of valence (LAHV) are rare, as it was also the case in DEAP database (Koelstra et al. 2012). This phenomenon is also predicted by theory, as valence and arousal are not independent (Norris et al. 2010), so that they are less likely to evenly distributed on the AV plan.

The distribution of each categorical emotion on the AV dimensional plan is presented in Fig.3.7. The AV score has been jittered in order to avoid the overlapping of the points. We notice that, "frustration", "anger" and "fear" are concentrated on the HALV region. "Anger" events have slightly higher score of arousal and lower score of valence comparing with "frustration" and "fear". "Happy" is concentrated on the HAHV region. The "boredom" emotion, with few annotations, appears in the LALV region.

3.3.2 *Self-assessed experience on sequences of game*

Concerning game experience, the game sequence questionnaire assesses each sequence in terms of the DIA dimensions, while the game ranking



(a) Categorical representation.

(b) Dimensional representation

Figure 3.6: Distribution of the annotated events in terms of emotions.

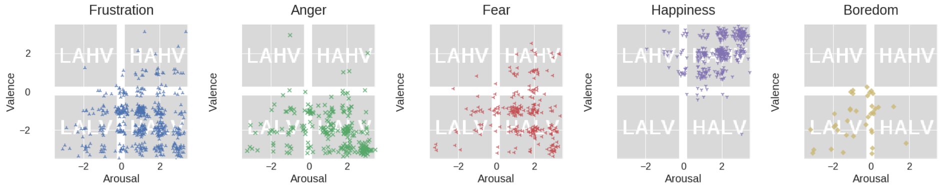


Figure 3.7: Distribution of categorical emotions on AV space

questionnaire gives a ranking of the 3 matches in terms of DIA dimensions. This section presents the statistics on game sequence questionnaire. Table 3.3 presents the correlations of the different self-assessed ratings. We notice that:

High positive correlations can be observed for the ratings of the same dimension (difficulty (0.70), immersion (0.57) and amusement (0.66)) between score-based ratings (**rs**) and question-based ratings (**rq**). Meanwhile, even though the evaluation between different dimensions is not independent, their is no high correlation between different dimensions, which shows that the evaluations on these dimensions are not redundant.

Medium positive correlations were observed within the pair (global feeling, amusement: 0.47/0.67) and the pair (immersion, amusement: 0.49/0.37/0.38/0.30), suggesting that participants have better feelings towards the sequence which amuse them, and they tend to be more implicated in these sequences. This phenomenon can be explained by the "flow theory" (Csikszentmihalyi 1996): players in the flow state are more implicated in the game and acquire a high level of enjoyment.

Medium negative correlations were observed within the pair (global feeling, difficulty: -0.36/-0.58) and the pair (amusement, difficulty), which

Table 3.3: Inter-Correlations of Self-Assessed Ratings across Individuals

	rgs	**rsD**	**rqD**	**rsI**	**rqI**	**rsA**	**rqA**	**order**
rgs	1.00	-0.36**	-0.58**	0.29**	0.28**	0.47**	0.67**	-0.06
rsD		1.00	0.70**	0.00	0.07	-0.13	-0.35**	0.03
rqD			1.00	-0.20**	-0.11	-0.37**	-0.59**	0.01
rsI				1.00	0.57**	0.49**	0.37**	0.22**
rqI					1.00	0.38**	0.30**	0.17*
rsA						1.00	0.66**	0.20**
rqA							1.00	0.08
order								1.00

Note: rating of global score (**rgs**), rating of score / question based difficulty (**rsD/rqD**), rating of score/ question based immersion (**rsI/rqI**), rating of score/question based amusement (**rsA/rqA**), ($** : p < 0.01, * : p < 0.05$).

signifies a relatively negative relationship between game difficulty level and overall satisfaction of the game.

Small and positive correlations were observed between sequence presenting order and immersion (0.17/0.22) and amusement ratings (0.08/0.2), which implies that as time went by, the participants' immersion and their evaluation of amusement slightly increase and they did not suffer from effects of fatigue or habituation.

3.3.3 *Self-reported assessment for players of different skill level*

Figure 3.8 presents the distribution of players' skill levels. We can notice that the level of players' skill level follows a normal distribution. In order to understand how the players' skill levels influence their self-reported assessment, the following analysis are conducted.

Figure 3.9 presents the normalized frequency of the arousal and valence score assessed by each level of participants. We can notice that, for arousal assessment, the score of low skill level players (amateur, semi-pro) are more dispersed than the high skill level players (professional, world-class, legendary). Most annotations concentrate on the positive arousal plan. While for valence assessment, most annotations concentrate on the negative plan.

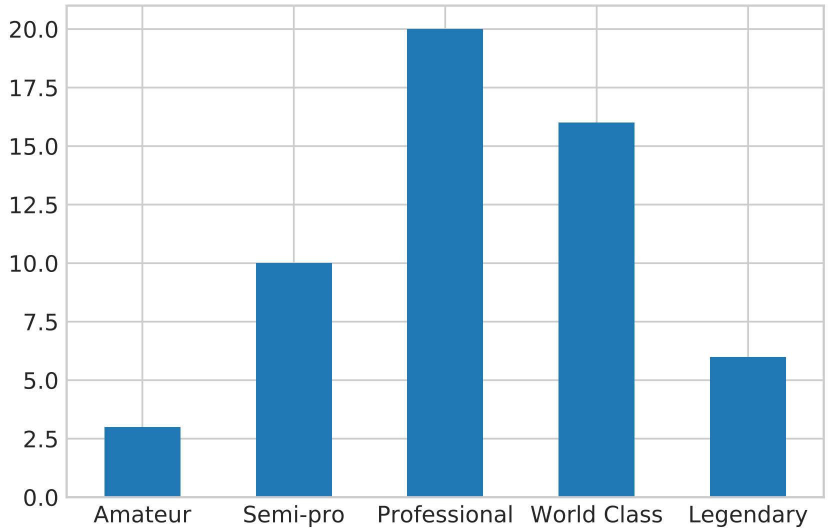
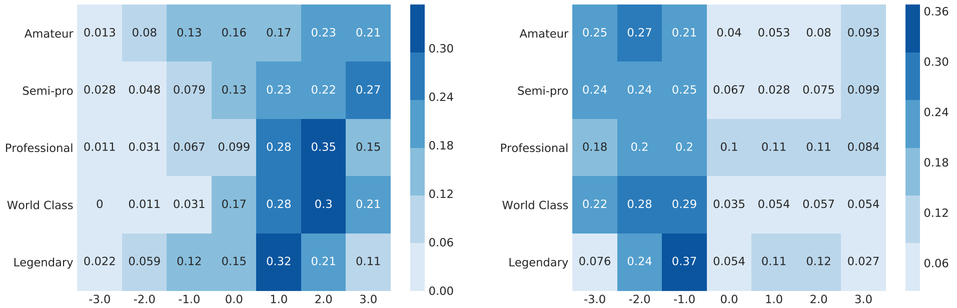


Figure 3.8: Distribution of players skill level

Players of legendary level are less likely to annotate the extreme negative emotion (-3 for the valence score).

Concerning the how the players of different skill levels evaluate the overall game experience, the result is synthesized in Figure 3.10. Players of different skill level: beginner (amateur, semi-pro), intermediate (professional), and expert (world Class, legendary) evaluate the ranking (low, medium, and high) of game experience (difficulty, immersion, and amusement) of 3 matches (easy, medium, and difficult) they played. For difficulty perception, we can notice that the beginner and intermediate players report an assessment correlated to the real game difficulty level, while for the expert player, the perception of difficulty is more blurred. This may be caused by the game settings, that higher difficulty levels don't make much differentiation. For immersion perception, beginner players report a high positive correlation between the difficulty of the match and the immersion, while for intermediate and expert players this correlation is not evident. For amusement perception, players of all levels report a perception of least amusement in the most difficult match. Intermediate players report a negative correlation between the game difficulty and amusement. This phenomenon signifies that a too difficult game ruins the game amusement and that appropriate game difficulty setting should be chosen to entertain the players.



(a) Arousal

(b) Valence

Figure 3.9: Emotion annotation for players of different skill level

3.4 CONCLUSION

In this Chapter, by reviewing the some popular public affective computing datasets, we identified key characteristics of dataset required by our affective game research. We then presented settings and protocol of the experiment for data collection and addressed the challenges concerning emotion representation, modalities measuring, subjective assessment. The proposed experimental paradigm takes a step further compared to the state of art affective game research by providing two levels of subjective assessment: minor scope evaluation on *game event* and global scope evaluation on *game sequence*. This paradigm not only make it possible to analyze the psychophysiological response in the dynamic context, but also provide a perspective to see how the game event influence overall game experience. In the following two chapters, we are going to present the feature extraction process on the measuring modality and the analysis we carried out on the collected data.

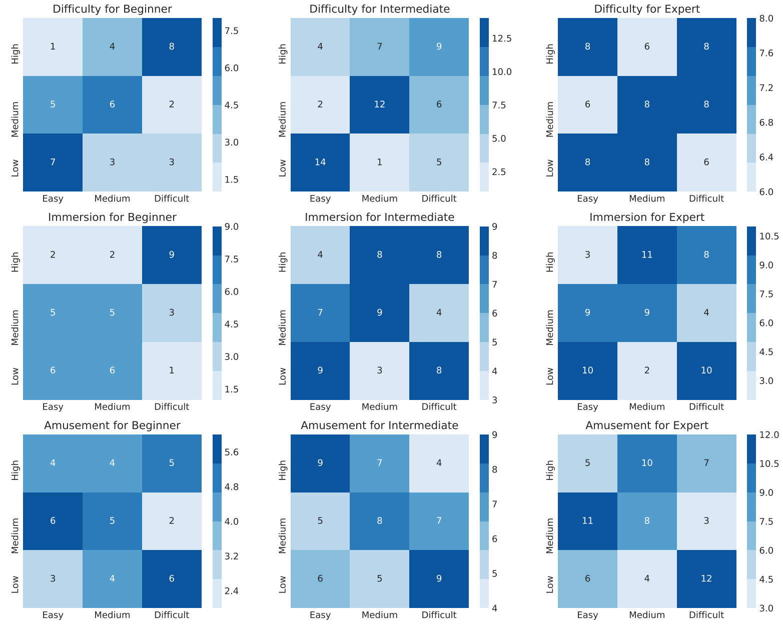


Figure 3.10: Game experience for players of different skill level

Part II

BUILDING A RECOGNITION MODEL

In this part, we present the a set of analysis and the models built on the proposed DAG database. We present the training method and evaluation of general, user-specific model and present how we improve the performance of recognition by using a group-based model.

We firstly review some feature extraction method used in the physiology-based affective computing community and present the features used in this thesis (Chapter 4). The feature extraction is an important step for creating the machine learning models and the presented features are used for all the physiology-based recognition models illustrated in the following section.

Then, we present a set of analysis to create *general models* concerning emotional moment detection, emotion recognition and game experience evaluation. We investigate the effects of segmentation length, normalization method, relevant signals and discuss the factors which influence the recognition rate of the proposed general model (Chapter 5).

One shortcoming of the general model is that all instances are taken equally so that the individual differences is neglected. Therefore, we train separate model for each subject and investigate these user-specific models. Different feature selection methods (filter and wrapper, nested LOOCV and non nested LOOCV), optimal size of feature set size were investigated. We confirm the existence of individual variability among subjects and presented the flaw of using user-specific model (Chapter 6).

Next, due to the flaw of the general model and user-specific model, the group-based model aiming to improve the recognition performance is presented. The basic idea of the group-based model is to find similar user groups, and used the model train on the similar user group to predict the emotion. For this purpose, we firstly present how to find physiologically similar user based on different views (signal, feature and model) by using clustering techniques. We show that by clustering users, the performance of recognition can be improved (Chapter 7). Then, we propose to use the group-based model that both takes into consideration the individual variability and makes the most use of existing data. We show that the proposed group-based model performs better than the general model and the user-specific model. We also investigate some characteristics of the proposed model(Chapter 8).

In the end, we evaluate the computation time for the personalized emotion recognition model. We realize a simulation on computer and an implementation on a real ARM A9 embedded system (Section 8.4).

4 SIGNAL PROCESSING AND FEATURE EXTRACTION

The collected physiological signals always contain artifacts and are of high dimension. In this chapter, we present how the signals are pre-processed to extract features for the further study.

Of all the modalities collected during the experiment, participant's face recording are eliminated from further analysis because of high rate of missing value reported by the FaceReader software. In the following section, we first report the popular physiological-based feature extraction methods used in affective game (section 4.1). Then we present the signal pre-processing and feature extraction process used in our study (Section 4.2).

4.1 FEATURE EXTRACTION REVIEW

There are basically 3 types of features used in the physiological-based affective computing: time domain features, frequency domain features and time-frequency domain features. Each feature set can be further divided into subsets based on the processing approaches:

- **time domain features**: refer to features calculated based on the time domain signals.
    1. **statistical features**: refer to features describing the statistical characteristics (such as mean, variance, kurtosis) of the signal. They offer an overall quantitative measures of the signal and are the most commonly used features (Kim et al. 2004; Kim et al. 2008; Picard et al. 2001; Rigas et al. 2007; Katsis et al. 2008). They are easy to calculate and is applicable to both short and long signal segments depending on the temporal sensitivity requirement.
    2. **morphological features**: refer to the features related to signals morphological property such as peaks. It is widely used in signals such as EDA, EMG, for which the occurrence of peaks is an effective measure of human activation (Kim et al. 2004; Kim et al. 2008; Picard et al. 2001; Katsis et al. 2008). The morphological features are less robust than the statistical features for two reasons. First, they rely on the precise detection of the morphological shape, however, the complexity of the physiological signal and the

presence of artifact make morphology detection a big challenge. For example, in order to analyse the heart rate (HR) or the heart rate variability (HRV), the detection of the ondulation pattern on the ECG is important. Methods such as Pan and Tompkins QRS detection (Pan et al. 1985), A Teager energy operator (TEO) (Maragos et al. 1993) were used to detect the R-peak. Second, they require longer segment length to acquire good morphological shape detection. For example, respiration is slow compared to heart rate, so that it should take longer to acquire several complete respiration cycles in order to deduce effective respiration features.

3. **entropy based features**: refer to a measure of disorganization or uncertainty in the random variable. (Kim et al. 2008) applied approximate entropy and sample entropy to analyze HRV. The calculation is complicated as it requires settings of multiple parameters and is less commonly used for the affective computing research.

4. **time series model features**: refer to the features derived from the time series model such as auto-regression model, or moving average model. This method takes the physiological signals as general time series, and train the time series parameters to describe the signal. For example, (Kim et al. 2004; Broersen 2000a; Broersen 2000b) choose the best time series model among auto-regressive, moving average and auto-regressive moving average models and use the learned parameter as feature for heartbeat signal. This method is less commonly used for physiological signal.

5. **chaos theory-based features**: refer to the features calculated based on chaos theory and method, such as Poincare plot or recurrent plot features. (Kim et al. 2008) applied Poincare plot and calculated the variance measures on the plot as features to the HRV. This method requires long time series and are not applicable to real time use.

- **frequency domain features**: refer to features calculated based on coefficients from frequency domain transformation such as Fourier transform. It is widely applied on physiological signals such as ECG, EDA, EMG, respiration or the transformed signals such as heart rate or the HRV. They generally require long segments, as long segments result in higher frequency sensitivity (Kim et al. 2004; Kim et al. 2008; Picard et al. 2001; Rigas et al. 2007; Katsis et al. 2008).

Table 4.1: Common feature types for physiological based affective computing

No.	Feature type	real-time	complexity	parameter
1	statistical	yes	$O(N)$	no
2	morphological	depends	depends	depends
3	entropy-based	depends	depends	required
4	time series model	depends	depends	required
5	Poincare plot	no	$O(N^2)$	required
6	frequency domain	yes	$O(NlogN)$	no
7	time-frequency	yes	$O(N)$	no

Table 4.2: Some examples of feature sets used in the physiological based affective computing

Stimulation	Feature set	Segment length	Ref
multimodal	1,2,4, 6	50 s	(Kim et al. 2004)
music	1,2,3,5,6	180 s	(Kim et al. 2008)
image	1,6	100s	(Picard et al. 2001)
image	1	10 s	(Rigas et al. 2007)
music	7	120 s	(Zong et al. 2009)
car-racing	1, 2	10 s	(Katsis et al. 2008)
game	1,2,6,7	240 s	(Rani et al. 2006)
Tetris	1,2	300 s	(Chanel et al. 2008)
TORCS	1	60 s	(Tognetti et al. 2010)

- **time-frequency domain features**: refer to features calculated based on coefficients from time-frequency transformation such as wavelet transform. One advantage of wavelet transform over the short term Fourier transform is that it provides a balance for the time and frequency sensitivity. As a result it is generally used for long segments.

Table 4.1 summarizes the common feature types used in the physiological-based affective computing. We tried to evaluate each type of feature based on whether they are able to represent short signal segments (real-time processing), calculation complexity and whether the parameter is required. For the real-time processing aspect, Poincare plot requires long signal length (∼1 min) to achieve meaningful feature which is not applicable for real-time processing. Segments requirement for morphological, entropy-based, and time series model depends on the type of signal and concerned type of feature. For calculation complexity, statistical, frequency domain and time-frequency domain features are easier to compute compared with others.

Concerning the parameter requirement, statistical, morphological, frequency, time-frequency domain features generally don't require preset parameter, which makes them more convenient to calculate. Table 4.2 presents examples of feature set used in the literature. The numbers in feature set columns correspond with the numbers in Table 4.1. Research-oriented works generally analyze the psychophysiological response using passive stimulation such as music or video, while application oriented research used stimulation in which the participants take active interaction. For both circumstances, statistical and morphological features are the most commonly used ones and also the only ones used for short segment length (10 s). Entropy-based, time series model based, and Poincare plot based model require preset parameters and are less commonly used in the literature. Time-frequency domain features such as wavelet transform has the advantage over the frequency domain feature on long varying time series segments. As we focus working on short segments and frequency domain are more commonly used, therefore, frequency domain features are chosen for our study.

In summary, we choose to use the statistical, morphological and frequency domain features.

4.2 FEATURE EXTRACTION

We focus on the peripheral signals and accelerometer data. Signal processing and feature extraction for each signal is presented below.

4.2.1 *Signal segmentation*

The collected physiological signals are continuous. Signal segmentation truncates the continuous signal into segments, on which the features are calculated. The approach of signal segmentation is related to research purpose. In our study, we apply 2 segmentation approaches: segmentation based on game events and segmentation based on overall game sequence.

Segmentation based on game events truncates the signals centered on the annotated game events with a given segmentation length. The purpose of this segmentation is to analyze the physiological response on the game events. Different segmentation lengths (10s, 14s, 20s, 30s) are applied in order to decide the effective response length of the signals or the extracted features.

As a reference to the segments based on game events, different baseline segments are also applied: the random segments from the music sequence, the random segments without annotation from the game sequence the segments just before the annotated segments. These segments are used as baseline reference and are taken to have the same length as the related event segment.

Segmentation based on the overall game sequence truncate the signals on the half-time match. The purpose of this segmentation is to assess the average physiological response along the game. We have configured the length of each half-time match to be 4 min, however, depending on matches' situation the actual lengths of different half-time varie from 4-6 mins. In order to have the same length for all the match segments, only the last 4 mins of each half-time match are kept.

4.2.2 *Feature calculation*

The general process of feature extraction on the signal segment consists of 3 stages:

1. *signal pre-processing* cleans the signals to avoid noise or artefacts (such as spiking removing, signal baseline removing, filtering),
2. *signal transformation* represents the characteristic of a signal in a different aspect (e.g. generating HR sequence from ECG signal),
3. *feature calculation* extracts common/ specific, linear/non-linear, time/frequency domain features on the pre-processed or transformed signal.

For the *feature calculation*, we present the following feature sets:

1. ***time-domain statistical feature set (time)***:
   Common time-domain statistic features applied to the sequences are: mean, median, maximum, minimum, range, variance, standard deviation, average derivative, maximum derivative, absolute deviation, kurtosis and skewness. For a given segment of physiological signal vector x, $x = (x_1, x_2, ..., x_N)$ The expression of these features are presented in Table 4.3. Time-domain feature set contains 12 features in total.
2. ***time-domain morphological feature***:
   number of peaks for the EDA signal
3. ***frequency-domain feature set (freq)***:
   power of a signal at different frequencies and their ratios using spectral analysis. The frequency band used for each signal is detailed below.

Table 4.4 summaries the feature extraction process for each signal. The detailed description of each signal is presented below:

Table 4.3: Common statistical features

Features	**Description**
mean	$\overline{x}$
median	median value
max	$max(x)$
min	$min(x)$
range	$max(x) - min(x)$
variance	$E[(x - \overline{x})^2]$
std.	$\sqrt{E[(x - \overline{x})^2]}$
avg. der	$(x_i - x_{i-1})/N$
max gra	$max(x_i - x_{i-1})$
abs dev	$abs(x_i - x_{i-1})$
kurtosis	$E[(x - \overline{x})^4]/(E[(x - \overline{x})^2])^2$
skewness	$E[(x - \overline{x})^3]/(E[(x - \overline{x})^2])^{3/2}$

ECG

ECG measures the action potentials of the heart from skin. Features calculated from ECG has been used to differentiate between positive or negative emotion or to indicate the mental effort and stress.

Figure 4.1 presents the process of ECG processing. A baseline removing is performed on the ECG raw signal. Then we apply a R-peak detection and compute the inter-beat-interval (IBI), heart rate (HR), and heart rate viability (HRV). IBI refers to the interval between consecutive heartbeats. HR refers to the number of heartbeats in a given amount of time. HRV refers to the oscillation of the interval between consecutive heartbeats. Then, on the pre-processed signal, the common time domain features for the IBI, HR and HRV are calculated.

In the frequency domain of the HR time series, three frequency bands are of general interest: the very LF band (0- 0.25 Hz), the LF band (0.25-0.5 Hz), and the HF band (0.5 - 0.75 Hz). From these subband spectra, we computed the power of each band by integrating the power spectral densities (PSDs) obtained by using the Welch's algorithm, as well as the ratio of power within the LF band to that within the HF band (LF/HF).

EDA

EDA measures the skin conductance from the activity of the eccrine sweat glands (located in the palms of the hands and soles of the feet). The EDA

Table 4.4: Feature extraction process

Sig. (nb. fea.)	Preprocess.	Trans.	Feature calcu.
ECG (44)	baseline removing, filtering	raw IBI HRV HR	raw: *freq* IBI: *time.* HRV: *time.* HR: *time, freq*.
EDA (53)	spike removing, filtering, subsampling	raw phasic dephasic tonic	raw: *time, freq* phasic: *time* dephasic: *time* tonic: *time* specific: nb. of peaks
EMG (24)	RMS smoothing, aggregating, subsampling	raw	raw: *time*
Respiration (40)	spike removing, filtering, subsampling, baseline removing	raw RR amp.	raw: *time, freq* RR: *time* amp.: *time*
ACC (12)	baseline removeing, RMS smoothing, 3-axis aggregating, subsampling	raw	raw: *time*

raw - pre-processed signals, ***time*** - calculate time domain feature set, ***freq*** - calculate frequency domain feature set

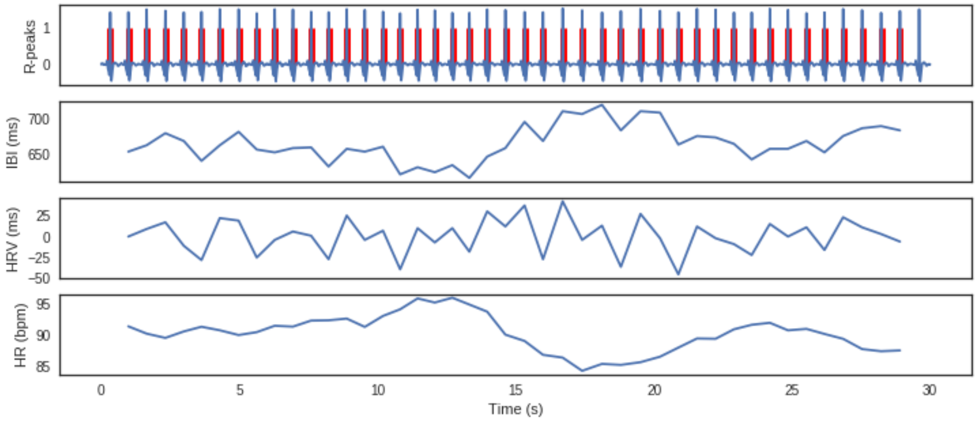


Figure 4.1: ECG signal processing

signal consists of two components: a slow moving tonic component that indicates general state of the glands and a faster phasic component that is influenced by emotions and the level of arousal. Many studies over the years have indicated that the magnitude of electrodermal change is closely associated with measures of emotion, arousal, and attention (McCurdy 1950; Lang 1995).

The pre-processing for EDA signal includes: spike removing using a median filter with a size window 10, low-pass filtering with cut-off frequency 25 Hz, and sub-sampling to 10 Hz. The transformed signals are EDA tonic part (low frequency part < 0.03 Hz) and EDA phasic part (Figure 4.2). We get also the derivative EDA phasic time series. The time domain features are calculated for the pre-processed EDA signal, the tonic and phasic part of the EDA signal and the derivative of the EDA phasic time series.

In the frequency domain of the EDA time series, three frequency bands are of general interest: the very LF band (0- 0.25 Hz), the LF band (0.25-0.5 Hz), and the HF band (0.5 - 0.75 Hz). From these subband spectra, we computed the power of each band by integrating the power spectral densities (PSDs) obtained by using the Welch's algorithm, as well as the ratio of power within the LF band to that within the HF band (LF/HF). The specific feature of EDA signal is the number of peaks.

EMG

Electromyography measures muscle activity by detecting surface voltages that occur when a muscle is contracted. EMG has been applied on the

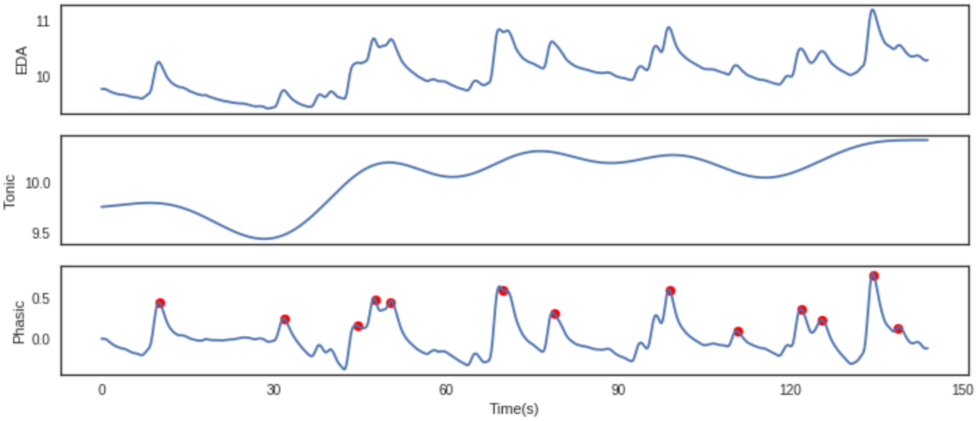


Figure 4.2: EDA signal processing

face (jaw) to distinguish "smile" and "frown" by measuring the activity of zygomatic major and corrugator activity (Sloan 2004).

The two channels of EMG are processed in the same way. EMG signals are symmetric and zero centred. In order to extract muscle movement, a RMS (root mean square) filter is applied to the signal. Then the signals pass a windowed aggregator to form the pre-processed signal. On the preprocessed signal, common time domain features are calculated.

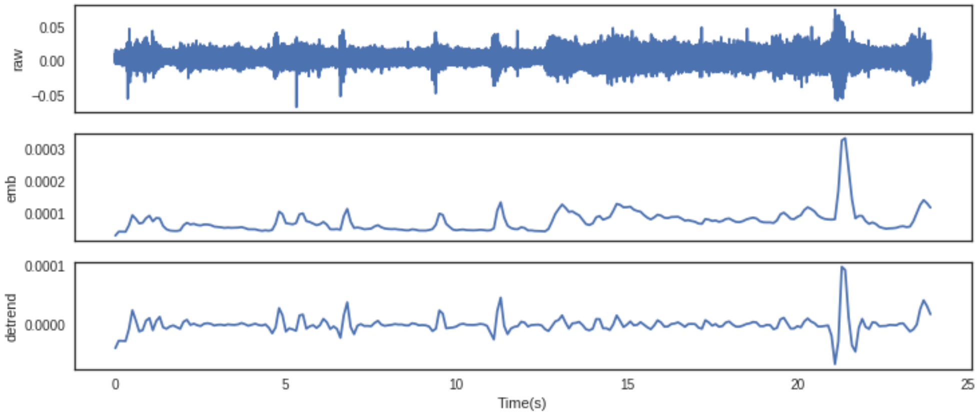


Figure 4.3: EMG signal processing

Respiration

Respiration is measured as the physical change of the thoracic expansion. Respiration rate is an indicator of stress. It generally decreases with relaxation. Startling events and tense situations may result in momentary cessation (Kim et al. 2008). Negative emotions generally cause irregularity in the respiration pattern (Kim 2007).

The pre-processing for Respiration signal includes: spike removing using a median filter, low-pass filtering with cut-off frequency 25 Hz, and sub-sampling to 10 Hz, and detrending, respiration peak detection. The transformed signals include the respiration rate (RR) and the amplitude for each respiration (Amp) (Figure 4.4). The time domain features are calculated for the pre-processed respiration signal, RR, and Amp time series. In the frequency domain of the respiration time series, three frequency bands are of general interest: the very LF band (0- 0.25 Hz), the LF band (0.25-0.5 Hz), and the HF band (0.5 - 0.75 Hz). From these subband spectra, we computed the power of each band by integrating the power spectral densities (PSDs) obtained by using the Welch's algorithm, as well as the ratio of power within the LF band to that within the HF band (LF/HF).

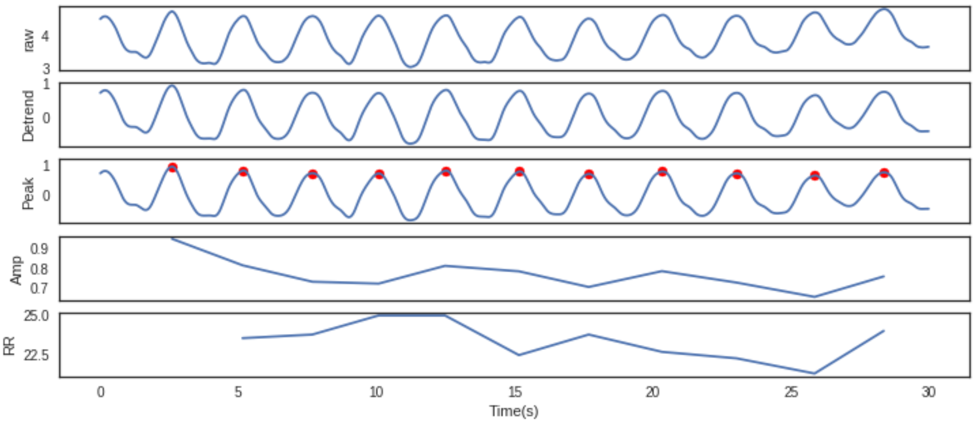


Figure 4.4: Respiration signal processing

Accelerometer

An accelerometer detects the movement of body and measures the acceleration on the 3 axis. In order to take into account the movement in all directions, after applying baseline removing and RMS smoothing, signals of 3 axis are aggregated by addition. Then the aggregated signal is sampled to 10 Hz. Time domain features are calculated on the aggregated time series.

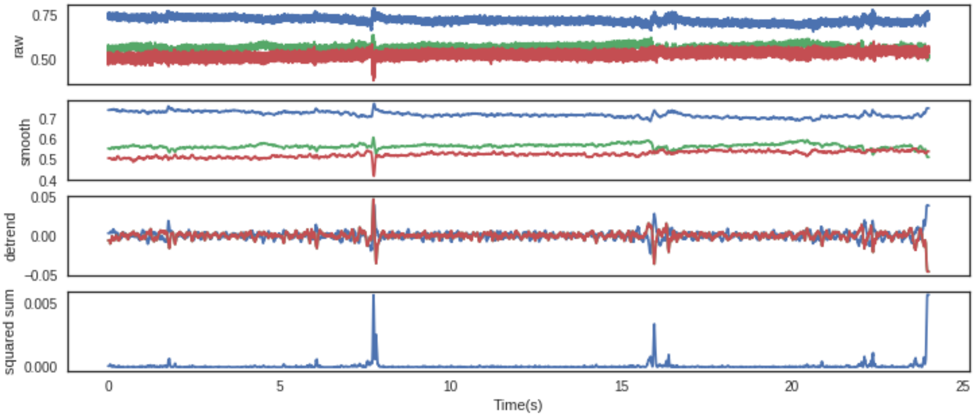


Figure 4.5: Accelerometer signal processing

4.3 CONCLUSION

In this chapter, we review the common features used in the physiological-based affective computing. We presented the features used in our study. For each segment of signals we obtain 173 features. In the following chapter, we present the classification tasks and results based on these features.

5 GENERAL EMOTION RECOGNITION MODEL

Our experimental paradigm has made it possible to make a multi-levels analysis related to game events and game sequence experience. In this chapter, we first present the analysis based on game events (Section 5.1) and then present the analysis related to global game segments (Section 5.2). The main conclusions and discussion are presented in Section 5.3.

5.1 CLASSIFICATION BASED ON GAME EVENTS

In this section, we are interested in the following questions: is it possible to find the events which have affected the emotional state of the player? To which extent can we determine the emotional state in the dynamic context? We investigate these problems with various machine learning tasks. The learning on game event is based on related segments of physiological signals. We present the methodology and the results of analysis on emotional event detection and emotion recognition of the emotional moments. These 2 problems can be further formalized into two classification tasks:

- classification of the annotated and non-annotated segments (Section 5.1.1)
- classification of low/high arousal (LA/HA) and low/high valence (LV/HV) (Section 5.1.2).

5.1.1 *Classification of the Annotated and Non-Annotated Segments*

5.1.1.1 *Description of the Classification Task*

The objective of this classification is to find out whether it is possible to distinguish the emotional segments from the others. By taking advantage of the self-assessed annotations on game events which have been reported to have triggered emotions, we obtain the 2 types of segments for this task: "segments with annotations" which can be viewed as segments with emotional responses and the "segments without annotation" which can be viewed as segments less likely to have emotional responses.

These 2 types of segments correspond to the 2 classes: *annotated* and *non-annotated*. In order to construct the learning set, we manage to get the same number of instances for each class. For each half-time, we take the segment centred around the annotation points as instances for *annotated* class, and take randomly the same number of non-annotated segments of the same length as instance for the *non-annotated* class.

Considering the variety of the emotional segments, we are interested to know how well we are able to distinguish each type of *annotated* segments from the non-annotated ones. By taking advantage of the annotation items: *dimensional emotion*, *categorical emotion* and *event types*, the *annotated* segments can be categorized based on these 3 dimensions. For example, in *dimensional emotion* dimension, *annotated* segments can be categorized into 4 groups: HAHV, HALV, LAHV, LALV. For each group of *annotated* segments (e.g. HAHV), we take the same number of *non-annotated* segments, in order to construct the learning set for the classification of *annotated* segments from the *non-annotated* segments for the given group (e.g. HAHV).

A detailed description of the learning process is presented below:

1. ***Segmentation**:* We segment physiological signals with different segmentation lengths (10s, 14s, 20s, 30s) in order to inspect the most effective signal length to detect the presence of emotions in the dynamic context.

2. ***Feature extraction:*** For each segment, we extract features presented in Chapter 4.

3. ***Normalization:*** In order to reduce the individual variability, each feature is separately normalized for each participant using standard normalization and [0,1] range normalization.

4. ***Cross validation:*** We use a 10-fold cross validation scheme.

5. ***Feature selection:*** At each step of the cross validation, we use Fisher's linear discriminant J for feature selection as has been used in (Koelstra et al. 2012):

   $$J(f) = \frac{|\mu_+ - \mu_-|}{\sigma_+^2 + \sigma_-^2}$$

   where μ_+, μ_- and σ_+, σ_- are the mean and standard deviation of feature f for the positive, negative class respectively. The J measure provides the performance score of each feature. In order to get the optimal number of features to keep, we tested it in the inner loop of the cross validation. The number of features to keep is taken ranging from 10 to 120 with a step of 10. We find that best result is obtained when the number of features equals 20, so that the best 20 features are kept for the classification.

6. ***Classification:*** Of all the possibilities of different classifiers, we present the one which we obtained the best result, the linear SVM, and report its accuracy and F1-score. The baseline is taken as the maximum F1-score from the uniform classifier and majority classifier.

The next section presents the results of classification of *annotated* and *non-annotated* segments by comparing different groups of events and different segmentation lengths.

5.1.1.2 *Result of Classification on the Annotated and Non-Annotated Segments*

Table 5.1 shows the average accuracy of classification of sequences with and without annotation. As has been explained in section 5.1.1, the *annotated* segments can be categorized based on 3 dimensions: *dimensional emotion*, *categorical emotion* and *event types*. Each dimension contains groups of annotated/non-annotated segments.

Among the *dimensional emotion* groups, event groups with high level of arousal (HAHV and HALV) obtain the best result of classification (0.641/0.645), while the event group with low level of arousal and valence (LALV) obtains the worst result (0.515) of classification. The performance of event detection on the HAHV and HALV events is significantly better ($p < 0.01$) than on the LALV events. We may conclude that it is easier to detect the HA events than the LA events and that whether the event is with high or low valence don't plays an important role in the detection efficiency.

Among the *categorical emotion* groups, by observing F1-scores on the 10 sec segmentation length, event groups with the best detection accuracies are "anger" and "frustration". These emotions are centred on the HALV region of the AV plan which corresponds with our previous observation that events with HALV have the best performance on their detection. "Boredom","fear" and "happiness" detection have a more modest result.

Among *event types* groups, no clear difference can be found among different event type groups. Neither a clear difference can be found between the positive event types and negative event types. We may conclude that the detection efficiency of different event groups are more dependent on the emotion than on the event type. This conclusion reinforce the need to exploit the self-reported feeling of the player rather than just analyse the game event log, as the same event may result in different type and intensity of emotions.

Table 5.1: Accuracy (ACC) and F1-score (F1) of the Classification on Annotated and Non-Annotated Segments For Each Event Group with Different Segmentation Length over Participants

	10 s			14 s			20 s			30 s		
Groups	**ACC**	**F1**	**Base**	**ACC**	**F1**	**Base**	**ACC**	**F1**	**Base**	**ACC**	**F1**	**Base**
Dimensional Emotion												
HAHV	0.645	0.625**	0.485	0.616	0.629**	0.476	0.630	**0.641****	0.553	0.580	0.604**	0.524
HALV	0.630	0.634**	0.512	0.633	**0.645****	0.523	0.611	0.644**	0.512	0.588	**0.645****	0.519
LALV	0.585	**0.515**	0.534	0.570	0.482	0.473	0.552	0.498	0.513	0.495	0.465	0.533
Categorical Emotion												
anger	0.700	**0.689****	0.482	0.688	0.684**	0.498	0.676	0.685**	0.495	0.588	0.597*	0.487
boredom	0.671	0.595*	0.514	0.607	0.583*	0.474	0.533	0.424	0.533	0.640	**0.607***	0.560
fear	0.623	**0.591***	0.530	0.570	0.541	0.473	0.613	0.575*	0.493	0.612	0.574*	0.462
frustration	0.663	**0.685****	0.499	0.635	0.645**	0.502	0.628	0.661**	0.492	0.583	0.629**	0.512
happiness	0.631	0.609**	0.493	0.634	0.624**	0.467	0.619	**0.628****	0.497	0.569	0.595*	0.465
Event type												
(+)goal	0.642	0.647**	0.488	0.639	**0.653****	0.498	0.614	0.642**	0.498	0.589	0.646**	0.502
(+)guard stop.	0.646	**0.653****	0.502	0.636	0.650**	0.486	0.618	0.643**	0.493	0.593	0.647**	0.495
(+)missing	0.648	**0.652****	0.511	0.635	0.651**	0.518	0.620	0.649**	0.517	0.594	0.648**	0.504
(+)shoot	0.655	**0.661****	0.485	0.637	0.654**	0.489	0.620	0.646**	0.496	0.599	0.653**	0.510
(+)tech.gesture	0.645	0.654**	0.490	0.637	0.649**	0.501	0.629	**0.656****	0.499	0.592	0.648**	0.498
(-)goal	0.651	**0.657****	0.512	0.632	0.644**	0.490	0.619	0.647**	0.497	0.589	0.646**	0.508
(-)guard stop.	0.646	**0.653****	0.505	0.637	0.652**	0.500	0.622	0.650**	0.513	0.592	0.646**	0.520
(-)missing	0.655	**0.661****	0.491	0.636	0.654**	0.505	0.615	0.641**	0.487	0.597	0.653**	0.493
(-)shoot	0.649	**0.656****	0.478	0.640	0.652**	0.505	0.624	0.653**	0.492	0.589	0.644**	0.508
(-)tech.gesture	0.653	**0.659****	0.506	0.633	0.651**	0.512	0.615	0.641**	0.525	0.593	0.650**	0.509

Stars indicate whether the F1-score on detection each type of event is significantly higher than 0.5 according to an independent-samples t-test ($** : p < 0.01, * : p < 0.05$). For comparison, baseline F1-score is given by the maximum between majority and uniform classifier and is presented in the Base columns.

When comparing the detection accuracy across different segmentation lengths, the accuracy reaches the best when taking the shorter segmentation lengths (10 s or 14 s), except for the dimensional emotion groups.

5.1.2 *Classification of LA/HA and LV/HV*

5.1.2.1 *Description of the Classification Task*

This task involves the classification of LA/HA and LV/HV on annotated events ("sequence with annotation"). The ratings of AV on each event is used as learning target. On a scale of 7 points, the AV scores are splitted into two classes: LA/HA classes for arousal classification problem and LV/HV classes for valence classification. Note that the split results in unbalanced classes (Figure 3.6). To solve this problem, we randomly sample the majority class to get a subset with balanced classes. The classification takes the similar process as in the previous task, except for the ***normalization*** step, which is detailed below.

In order to reduce the individual variability and exploit the dynamic of the physiological signals three methods of normalization are applied.

- *Standard normalization (std)* normalizes each feature for each participants so that features for each participant all have zero mean and unit variance.

- *Normalization referencing precedent segment (delta)* takes the segment just before the annotated segment as reference level. The difference is calculated between annotation segment and the segment before. Then a standard normalization is applied on this difference for each participant.

- *Normalization referencing baseline segment (base)* takes the neutral state of each participant during music session as reference level. The difference is calculated between annotation segment and music segment. Then a standard normalization is applied on the new features for each participant.

Section 5.1.2.2 presents the results of classification on LA/HA and LV/HV by comparing different segmentation lengths and normalization methods presented above.

5.1.2.2 *Result of classification on LA/HA and LV/HV*

Table 5.2 shows the average accuracies of binary classification of AV scores with different segmentation lengths and normalization methods across participants.

By comparing the accuracy and F1-scores, we notice that the classification of AV scores is more difficult than the classification of sequences with and without annotation. The classification of valence is more difficult than arousal, as the best F1-score for arousal classification is 0.602 which is better than valence classification (F1-score with 0.573).

Concerning the three normalization methods, traditional standard normalization (std), precedent sequence referencing normalization (delta) and neutral state referencing normalization (base), the best result for both arousal and valence prediction is obtained by using the precedent sequence referencing method (delta). This method takes the signal and the emotion dynamic into account by assuming that emotion recognition is more based on the relative feature change than the absolute feature value. For arousal classification, when considering both the ACC and F1 measures, the second best result is obtained by using the referencing neutral state (base) method. The improvement may be explained by the fact that referencing with neutral state reduce the individual variability. However, no clear improvement can be observed on classification of valence.

By comparing the performances of difference segmentation lengths, we notice that the best results are obtained when taking the segmentation length of 14 or 20 seconds, which is longer than the length used in the task of classify sequence with/without annotation. We may conclude that the detection of events requires short segmentation length as longer segmentation smooth the effects of events, whereas the classification of emotion requires longer sequence as physiological signal varies slowly with emotion, but too long segmentation (e.g. 30s) may cause the overlapping of the successive events. As a result, one should find a balance between reaching the necessary signal length for emotion recognition and attaining the optimal time precision to avoid overlapping.

For each classification task, the best 20 features are selected from all the 175 features calculated from different signal. In order to better understand the role of each signal on the classification results, we analyse the most relevant signals for each learning task. Figure 5.1 presents the frequency of the 6 modalities, from which the features are selected for classification of *annotated* and *non-annotated* segments (Figure 5.1 *Event*), classification of high/low arousal (Figure 5.1 *Arousal*) and classification of high/low valence (Figure 5.1 *Valence*). We notice that for classification of *annotated* and *non-*

Table 5.2: Accuracy (ACC) and F1-score (F1) of Binary Arousal and Valence Classification across Participants

		Arousal		**Valence**	
norm.	**win(s)**	**ACC**	**F1**	**ACC**	**F1**
Std	**10**	0.477	0.476	0.468	0.502
	14	0.545	0.448	0.546*	0.539
	20	0.532	**0.550***	0.524	**0.567***
	30	0.364	0.362	0.478	0.520
delta	**10**	0.505	0.509	0.457	0.489
	14	0.559	**0.602****	0.524	**0.573***
	20	0.523	0.534*	0.551*	0.557*
	30	0.477	0.479	0.511	0.526
base.	**10**	0.508	0.433	0.480	0.461
	14	0.570	**0.551***	0.531*	0.517
	20	0.502	0.473	0.554*	**0.543***
	30	0.453	0.360	0.498	0.473
majority		0,423	0,42	0.455	0,455
uniform		0,504	0,504	0.507	0,507

Stars indicate whether the F1-score on detection each type of event is significantly higher than 0.5 according to an independent-samples t-test ($** : p < 0.01, * : p < 0.05$). For comparison, baseline F1-scores of classification by majority classifier, uniform classifier is presented below.

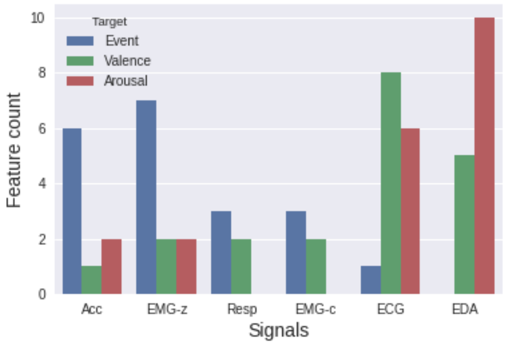


Figure 5.1: Importance of modalities for learning tasks: classification of with/without annotation event (*event*), classification of binary arousal score (*arousal*), classification of binary valence score (*valence*)

.

annotated segments, the most relevant features belongs to the accelerometer (Acc) and Zygomaticus muscles signals (EMG-z). This further explains why shorter segmentation length is demanded for this task, as the reactions on Acc and EMG-z are instant, longer segmentation length smooths the response effects. For valence classification the most relevant modalities are ECG and EDA, while for arousal classification EDA is more important than ECG. This observation is consistent with the work (Ringeval et al. 2015b) where ECG is more accurate for classifying valence and EDA for classifying arousal.

5.2 PREFERENCE LEARNING OF EVALUATION ON GAME SEQUENCE

In this section, we present the methodology and the results of learning the evaluation of the game sequences. Common game experience questionnaires evaluate the overall feeling as a whole which neglect the detailed emotion changes triggered by events. In our experimental paradigm, emotion refers to the instant affective sensation related to the in-game events and experience refers to the afterward sensation related to the overall game. One should distinguish the difference between the emotion related with in-game event and the evaluation towards the overall experience. For example, a player may suffer from consecutive frustrating events but finally achieve the final goal. Concerning the in-game events and their related emotions, most of them are negative. Meanwhile the player's game experience may still be good, as overcoming varies difficulties and achieving the final goal bring strong sense of accomplishment. By taking emotional details of events into account, we can get a better understanding of the player's overall game experience.

5.2.1 *Description of the classification task*

As has been presented in Chapter 3, each participant went through 3 matches and the game experiences are evaluated based on 3 dimensions: difficulty, immersion and amusement (DIA). In order to reduce the influence of individual variability on subjective evaluation, we carried out a preference learning on each pair of the 3 matches on each participants.

More precisely, for each dimension of DIA, the subjective ranking of the 3 matches (M1, M2, M3) is taken as ground truth, and is transformed into match couples ((M1, M2), (M2, M1), (M1, M3), (M3, M1), (M2,M3), (M3, M2)). The ranking problem is then transformed into a binary classification question: whether the one match is ranked higher than the other. If it is true, the current instance belongs to the positive class, otherwise, it belongs to the negative class. For example, if the difficulty ranking of the 3 matches is $M2 > M1 > M3$, the preference judgment of (M1, M2) is false (i.e. M1 is not more difficult than M2), so that this instance belongs to the negative class.

Among the 58 participants, 3 of them failed to finish all the three matches, as a result 55 rankings are available in the database, that is 330 (55*3) instances for the learning set. The following learning process is detailed below:

1. **Feature extraction:** In this study, three feature sets based on *game level*, *game outcome*, and *in-game events* are calculated. Table 5.3 summaries the features extracted for the preference learning.
   - **game level**: A proper setting of game difficulty level helps players to enter a "flow" state where they immerses totally in the game as their ability matches with the challenge. Features extracted are player's skill level (*pLevel*), game difficulty level (*gLevel*), and game relative difficulty level with respect to the player's skill level(*rLevel*);
   - **game outcome**: A better game outcome can satisfy players, as it helps to improve their sense of accomplishment. Features extracted are relative score of the match (*goal*), difference of the score between the first and the second half-time (*diffGoal*);
   - **in-game events**: Different events have different affective effects, by analyzing on event scale we can get better insight on how the instance emotions triggered by in-event influence the final game

Table 5.3: Features for Global Evaluation Preference Learning

	Extracted Features
Game level	player's skill level (**pLevel**), game difficulty level (**gLevel**), relative skill level (**rLevel**)
Game outcome	scored goals (**goal**), difference of goals of 1st, 2nd half (**diffGoal**)
In-game event	# **all**, # **anger**, # frustration (**fru**), # **happy**, # (+.)valence (**posVa**), # (-.)valence (**negVa**), # (+.)arousal (**posAr**), # (-.)arousal (**negAr**), # valence>1 (**bigVa**), #arousal>1 (**bigAr**)

# signifies the number of corresponding events

experience. Features extracted are the number of annotated events during the sequence (*all*), the number of anger events (*anger*), the number of frustration events (*fru*), the number of happy events (*happy*), the number of positive valence events (*posVa*), the number of negative valence events (*negVa*), the number of positive arousal events (*posAr*), the number of negative arousal events (*negAr*), the number of high score valence events (*bigVa*), the number of high score arousal events (*bigAr*).

2. **Normalization:** The features are normalized for each participant using the min-max normalization in the range [0,1]. The difference of features on each of the match couples are calculated as features.

3. **Cross validation:** The performance is evaluated using the 10-fold cross validation schema.

4. **Classification:** Decision tree (DT) is used for this classification task as it naturally provides relevant features and offers interpretable model. For each of the DIA dimension, a DT is created using the game level, game outcome and in-game experience feature sets.

Table 5.4: Accuracy (ACC) and F1-score (F1) of Preference Learning on Game Sequence Evaluation over Participants

	Difficulty		Immersion		Amusement	
	ACC	**F1**	**ACC**	**F1**	**ACC**	**F1**
svm	0,774	0,768**	0,661	0,569**	0,651	0,671**
majority	0,468	0.495	0.478	0,435	0.478	0.495
uniform	0,499	0,503	0.504	0,503	0.504	0,503

Stars indicate whether the F1-score on detection each type of event is significantly higher than 0.5 according to an independent-samples t-test $(\ast\ast = p < 0.01, \ast = p < 0.05)$. For comparison, baseline F1-scores of classification by majority classifier, uniform classifier is presented on the bottom.

5.2.2 *Result of Classification on the Game Experience*

Table 5.4 presents the result of classification of preference on the pairwise matches in terms of DIA. The performance for classifying *difficulty* is better than *amusement* and *immersion*. As *difficulty* is a relatively objective measure, it can be easier perceived and evaluated by subject. The most relevant feature for classifying difficulty are: the total number of annotated events, the relative scores and the number of negative arousal events. *Amusement* describes the subjective liking, the most relevant features include the difference between score of first and second half and the number of positive arousal events. The classification on *immersion* is the most difficult task which is not surprising as it is also difficult to evaluated by participant.

In order to investigate the most the relevant features, we present the DT models created for each of the DIA dimension. The model for *difficulty* dimension presented in Figure 5.2. In the figure, each box represents a node of the DT. Nodes with no output vertex going outside of it are leaves of the DT. In each node of the DT, we present the selected attributes (first line in the node), the proportion of positive and negative instances in this node (second line) and the majority class of the node (third line). Orange nodes are associated with the positive class and blue nodes are associated with the negative class. Similarly the DT created for immersion and amusement is presented in Figure 5.3a and Figure 5.3b.

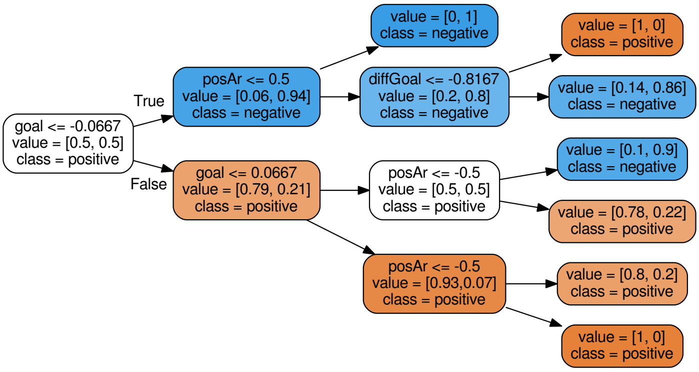
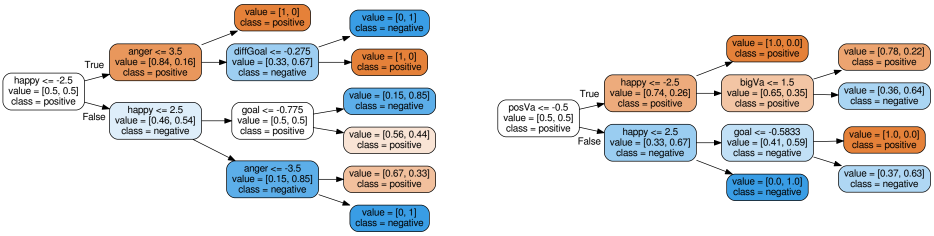


Figure 5.2: DT for difficulty



(b) DT for amusement

(a) DT for immersion

Figure 5.3: Distribution of the annotated events in terms of emotions.

By investigating the DT model created for each DIA dimension, and selected feature by each model, we may get an insight on the most effective factors that influence the DIA game experience dimensions.

For the *difficulty* dimension, the most relevant attributes selected by the DT model are: the number of scored goals (*goal*) and the number of positive arousal events (*posAr*). Even though the difficulty levels (*pLevel, gLevel, rLevel*) are proposed as features, they are not selected by the DT model, it reflects the fact that the proposed objective difficulty level is different than the perceived subjective difficulty level. Player's perception of the match are closely related with the game outcome and high arousal moments.

For the *immersion* dimension, the most relevant attributes selected by the DT model are: the number of happy events (*happy*), the number of anger events (*anger*), the number of scored goals (*goal*), and the difference scored goals of the first and the second half-time (*diffGoal*). By reviewing the related in-games events, we notice that the sense of immersion are both triggered by discriminant positive (such as *happy* and *goal*) and negative (such as presence of *anger*).

For the *amusement* dimension, the most relevant attributes selected by the DT model are: the number of happy events (*happy*), the number of positive (*posVa*) and big valence events (*bigVa*) and the number of scored goals (*goal*). The sense of amusement is all related with positive events, this can be explained by amusement raise with the sense of accomplishment.

Of all the three DIA dimensions, features from *in-game experience* feature set are the most selected, which indicates the importance of taking into account the in-game experience. Furthermore, with the interpretable DT model, it's easy and clear to locate in-game issues which influence the player's experience.

5.3 DISCUSSION AND CONCLUSION

5.3.1 *Discussion*

As far as we know, the proposed database is the first available one which allows analysing human emotional reactions from physiological signals under naturalistic interactive game context. Emotion recognition from game playing data using physiological signals is not a trivial task. In this section, we discuss different aspects which may influence the recognition performances. Figure 5.4 illustrates the process of constructing a machine learning model from this database. In the figure, black points (point 1, 2, 3) present the factors related to the database construction and white points (point 4, 5, 6) present the factors related to the data analysis method.

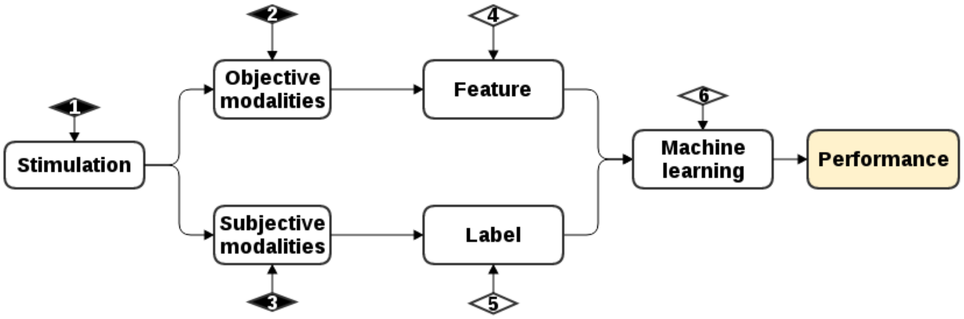


Figure 5.4: Factors which influence recognition performances

1. **Effectiveness of stimulation**
   One major concern when creating an affective database is the effectiveness of the stimuli. To address this issue, the construction of the databases (Koelstra et al. 2012; Soleymani et al. 2012; Abadi et al. 2015) underwent a strict stimuli selection process in which only the most effective stimuli is kept for emotion induction. In our context of affective gaming, game scenario varies for each subject, the stimulation cannot be pre-determined, thus making the emotion induction difficult to be controlled. We tried to settle this problem in 3 aspects: firstly, an interactive game context was chosen in order to improve the participant's engagement; secondly, emotions were annotated with the event type which can be used as an objective reference of emotion; thirdly, subjects were asked to only annotate the critical events which had influenced their emotional state. Moreover, we carried out a study on the classification of annotated and non-annotated sequences. A significant difference can be observed between these two classes which further confirmed the effectiveness of using game events as stimuli. Still, in our experiment, the effectiveness of stimulation mainly depends on participant's posterior self-assessment. This is an unavoidable choice in analysing self-assessed emotion in dynamic game context, so that the performance may suffer from drawbacks of cognitive error and imprecision of the recalled memory related to game stimulation. Another limitation is the uncontrollability of the stimulation in our experiment due to the naturalistic game interaction. In perspective, efforts can be made to design interactive game stimulation which can be more controllable in terms of emotion type, time, and intensity. This can be a collaborative work with a game design company, psychologist, and machine learning practitioner. Moreover, the physiological signals are taken all along the experiment and the timetable of each phase during the experiment (music, gameplay, questionnaire and annotation) is also available. The presented analysis have not yet exploited data from this aspect, but it could be interesting for the researchers who want to perform context detection using physiological signals (pervasive computing (Ye et al. 2012)).

2. **Capacity of objective modalities**
   The main modality we use in this study is the peripheral physiological signals. Despite of controversy of the relationship between physiology and emotion, physiological signals have been successfully used in affective computing especially in the specific context. Our objective is to investigate the usability of physiological signals in the specific affective

gaming context. In the literature, peripheral physiological signals are mostly used for recognition activation level using long time window (almost 1 min or more) (Kreibig 2010). Psychophysiological responses within short time-window on game event have rarely been addressed (Ravaja et al. 2008; Ravaja et al. 2006). In the presented classification result, only a modest performance has been achieved to distinguish HA/LA, HV/LV emotional state using physiological responses, which may signify that the physiological signals are not an ideal choice to recognize emotion in dynamic context. However, the effectiveness of PPS still needs to be verified by a more detailed study. In perspective, two measures can be taken. Firstly, in order to validate the effectiveness of using peripheral physiological signal for emotion recognition in game, a more strictly designed experiment can be conducted to investigate: whether there exists a consistent physiological response to different emotional state; whether is it possible to reduce the time window and what are the most appropriate segment lengths for each signal; and furthermore, when changing onto dynamic context, how the physiological responses change. Secondly, taking into account other modalities such as EEG, as it is viewed to have faster response and richer information. Recent developments of the EEG headset EMOTIVE[1] has made it less intrusive to use in the game context.

3. **Effectiveness of subjective modalities**
   Subjective evaluation is notorious for the noise it may produce due to the cognitive bias. In individual gameplay context, expressive modalities are less presented, so that evaluation by an observer is less evident. Hence, we applied the self-reported annotation in the recall manner. The proposed annotations reflect the participants' emotional state, but suffer from the problems such as cognitive bias, or memory issues. Besides the self-reported assessment, game event log, the observer/expert's annotation can still serve as a reference. In perspective, the experiment can also take into account the evaluation from the observer by viewing the expressive modalities such as facial expression and body gesture or from the expert who has the competence of assessing physiological signals response.

4. **Signal representation - feature**
   Given the hypotheses that stimulation is effective and that the physiological signals are able to reflect the emotion change triggered by stimulation, the classification performance can be largely influenced by

[1] www.emotiv.com

the signal representation process, also referred to as feature extraction. The objective of the feature extraction is to try to reduce the dimension of the input. Different features can be extracted, such as statistical features on the time and frequency domain, entropy-based features, morphological features, or a deep-learning framework can be used to learn a multi-level representation. The choice of the representation should be based on the validated segmentation lengths, as different features of different signals may have different effective length. Besides, the alignment of multi-variant input may also influence the final result, especially in the real-time dynamic context. The present work covers some of the most common features used in the peripheral physiological based affective computing. In perspective, more work can be dedicated to find new features extracted from different feature extraction methods or different signal segment lengths.

5. **Label processing**
   Given the cognitive bias which may happen during the subjective evaluation, the obtained labels should be processed in order to reduce this bias. In this paper, techniques such as the discretization of the dimensional evaluation to form a binary classification problem or preference learning to learn the game experience rankings were applied. In perspective, other discretization options, emotion recognition on specific categorical emotions or working directly on dimensional label can be tested.

6. **Prediction**
   In this paper, machine learning methods were used to train a model to differentiate different emotional states or game experiences. Considering the complex characteristics of the physiological signals, the subjectiveness of assessment achieving good performance is difficult. Moreover, there also exists variability among individuals. The individual variability can occur on several levels. On the stimulation level, according to participants' preference towards game, effectiveness of different stimulation varies; on the signal level, different participants may have different signals sensible to emotion change; on the subjective evaluation level, different subjects may have different habit of auto-evaluation, on the feature level, different participants may have different valid features or feature variation patterns. All these effects are unavoidable in affective research and can influence the final recognition result. In our database, even though the overall length of physiological signals recording is relatively long, and the total number of emotion

annotation on the game events is large, due to short play times (about 8 min for each of the three matches and in total 24 min) and small and varying number of events ($avg = 30.82, std = 11.7$) mean that the individual variability cannot be countered by intra-subjects repetitions. In the presented analysis, the emotional event detection and emotion recognition model tackle the individual variability problem by normalizing the calculated features, then a general model neglecting other sources of individual difference is created. In perspective, more attention should be paid to individual differences from the model view, for example by investigating the similar subjects and creating models for them.

5.3.2 *Conclusion*

In this Chapter, we carried out a series of analysis on the proposed multimodal database constructed under naturalistic interactive game context. Multi-levels analysis were made on both local game events and global game experience.

Concerning the local game events, two classification tasks were conducted in order to answer the following questions:
First, is it possible to detect the events which have affect the player's emotional state? To answer this question, we realized a classification of annotated and non-annotated segments. We found that 1) events with high arousal level are more detectable than that with low arousal level, 2) different event types don't make significant difference on event detection efficiency. Detection efficiency of different events is more dependent on the emotion than on the event type. 3) the detection accuracy reaches the best when taking the shorter segmentation lengths (10 s or 14 s).
Second, for all the annotated segments, to which extent can we determine the emotional state in this dynamic context? To answer this question, we realized a classification of f low/high arousal (LA/HA) and low/high valence (LV/HV) based on participants' self-assessment. We found that 1) low/high arousal (LA/HA) recognition performs better than the low/high valence (LV/HV) recognition, 2) using precedent sequence as reference outperform the the standard normalization and the neutral state referencing normalization, 3) segment length for emotional state recognition is longer than the detection task (14 s or 20 s), 4) performance of emotional state recognition is modest which illustrates the difficulty of using physiological-based method to recognize emotional states in a dynamic context. We discussed the limitation of our work and future work in the Section 5.3.1.

Concerning the global game experience, in order to inspect the key factors that influence the player's game experience, we realized a preference learning on the participant's game experience evaluation in terms of the perceived difficulty, immersion and amusement. We noticed that: 1) difficulty is easier to be recognized than immersion and amusement 2) player's perceived difficulty, immersion and amusement can be better understood by taking into account the game level, game outcome and in-game events factors: (a) Regarding difficulty: A player's perception of game difficulty is best understood as a function of how much they score and/or are aroused, also if he/she scores poorly, how much they improve within a match. (b) Regarding immersion: A player's immersion is tied to the experience of happiness, anger, and performance (goals and improvement). (c) Regarding amusement, the key factors are valence, happiness, and scored goals.

In the end, we discussed the possible aspects which can influence the emotion recognition performance and indicated the perspective work. In the following section, we try to optimize the prediction step by taking into account the individual differences and constructing the personalized model. We detail the approach and the result in the following chapters.

6 USER SPECIFIC EMOTION RECOGNITION MODEL

In the previous Chapter, a general emotion recognition model is created for all the subjects. Without taking into account the individual variability, we showed that the performance was not very satisfactory. In order to obtain a better performance on a given subject, many studies take the data from the given subject and train a *user-specific* model (Koelstra et al. 2012). The user-specific models are based on data from a specific given subject, which results in a small dataset and a large feature space.

In this chapter, we present the process of creating a user-specific model. We firstly present in more detail the feature selection techniques (Section 6.1) then we present the training process of the user-specific models (Section 6.2), in the end we present the performances of the user specific model under different methods of feature selection and present the optimal feature set (Section 6.3).

6.1 FEATURE SELECTION AND RANKING

The feature extraction process results in a large number of features. In order to analyze the individual differences we begin by inspecting the most discriminant features among the participants. Feature selection is a common process in machine learning to select the most relevant features. By reducing feature space, a faster and more cost-effective model can be built. It can also help to improve the prediction performance and to provide a better interpretability of the underlying pattern of the data (Guyon et al. 2003). Generally, there are two methods of feature selection: *filters* and *wrappers*. In this section, we apply some of the typical methods of feature selection and ranking (filter and wrapper), then investigate the most relevant features for each participant by evaluating the models' performances.

Filter

In the *filter* method, features are selected based on the measures which determine their relevance or discriminant power with regard to the target class. Common measures include mutual information, statistical tests (t-test, F-test), or more complex scores such as in mRMR (maximum relevant minimum redundancy) method. Features with high ranking of these measures are retained. Filter methods can be implemented easily and efficiently, as the measures are defined explicitly. The selected features are independent of the learning method, so that they have better generalization property, but usually filter methods give lower prediction performance than a wrapper.

For the filter method, the mRMR (maximum relevance minimum redundancy) is applied. It both maximizes relevance of the features with the target class and reduces the redundancy among different features. It uses an incremental search method by adding one optimal feature to the set at a time. Suppose we have already $m-1$ selected features in S_{m-1} and we want to select the mth feature. The new added feature should meet the following condition:

$$max\left[I(x_j;c) - \frac{1}{m-1}\sum_{x_i \in S_{m-1}} J(x_j;x_i)\right]$$

where x_j denotes the current new feature, $I(x_j;c)$ measures the relevancy of the current feature vector x_j to the target class vector c and $J(x_j;x_i)$ measures the correlation of the current feature with all the previously selected features. In our implementation, the $I(x_j;c)$ is evaluated by the F-test. And $J(x_j;x_i)$ is evaluated by the absolute value of the correlation of the variable x_j and x_i.

Regarding the ranking of the features, the first selected feature is the most important. The features are ranked in the same order as they have been selected.

Wrapper

The *wrapper* method associates the feature selection with a learning model. A new model is trained for each feature subset, and the effectiveness of the feature set is directly judged by the performance of the learning model. Common methods for constructing the feature set include, SFS (sequential forward selection) that starts from null set and recursively adds the most relevant features or SBS (sequential backward selection) that starts from the full set and recursively removing the least useful features, RFE (recursive feature elimination) that starts from a initial feature set and eliminates the features with the smallest score. The wrapper method results in the effective features subset adapted to the given learning method. As wrapper methods train a new model for each subset, they are very computationally intensive. The filter method can be used as a preprocessing step to reduce feature dimension and allowing the wrapper method to treat larger problem.

For the wrapper method the RFE method is applied. As it not only used the performance of learning model but also take advantage of the models' intrinsic feature importance ranking scheme. In this work, we apply recursive feature elimination (RFE) (Guyon et al. 2002) using Python *scikit learn*[1] toolkit. The RFE method first trains an estimator on the initial set of features and assigns weights to each one of them. Then, features whose absolute weights are the smallest are eliminated from the current set. The features are selected by recursively considering sets of features with decreasing sizes. We have tested 2 learning algorithms: linear SVM and Random Forest under the RFE feature selection scheme. Both algorithm are able to evaluate the feature importance besides reporting the prediction result. The feature importance of Linear SVM is associated with the absolute value of the coefficient, while the feature importance for Random Forest is associated with the feature presented in the trees.

Regarding the ranking of the features, the last remained features from the elimination process is the most important. So that the features are ranked in the inverse order of the elimination.

6.2 DESCRIPTION OF THE CLASSIFICATION TASK

In this section, the learning is implemented on each participant. We keep the subjects who have annotated more than 20 events, thus resulting in 45 persons in total. For each participant, the obtained data is processed as below:

1. ***Segmentation**:* We segment physiological signals with a segmentation length 14 s (best segmentation length obtained in Section 5.1.2.2).
2. ***Feature extraction:*** For each segment, we extract the same features presented in Chapter 4.
3. ***Normalization:*** In order to have the features in the same scale, each feature is separately normalized by using standard normalization (zero-mean and unit standard deviation).

The remaining procedures are feature selection, cross-validation and evaluation the performance of models on each participant. The feature selection methods used are as it has been presented above: for the **filter** method, mRMR was used, while for the **wrapper** method, RFE was used. Then the performances are evaluated using Linear SVM, Random Forest. For the cross-validation, as the number of the examples for each participant is limited ($avg = 30.82, std = 11.7$) and the feature size is relatively large (175), we use

[1] http://scikit-learn.org/

a Leave One Out cross-validation (LOOCV) scheme, in which each sample is used once as a test set while the remaining samples form the training set.

We distinguish 2 types of feature selection and cross-validation procedures:

- **Non-nested feature selection for LOOCV**
  In this procedure, the feature selection step is not embedded in the cross-validation scheme. Features are selected using all samples together, then the data with the reduce feature space is feed to the LOOCV step to evaluate the performance of learning algorithm. This method selects globally optimal feature set based on the knowledge from all examples. As it has also been used to evaluate feature selection on gene expression data (Ding et al. 2005). However, strictly speaking, this method suffers from the problem of data leakage, as the feature selection process has already seen the test data. So that the selected feature may be hard to generalize to new examples and may yield over-optimist classification accuracy. As a result, one should be careful when interpreting the feature selection and classification result.

- **Nested feature selection for LOOCV**
  In this procedure, the feature selection step is embedded in the each round of cross-validation. The features are selected by using only the training data, so that the testing samples have never been met in the whole learning process and is a data leakage free procedure. Meanwhile, in the case of small dataset and big feature space, the learning algorithm may not be able to achieve meaningful classification result, so that the features selected within this scheme is not discriminant enough to make the prediction.

We applied both feature selection and cross-validation procedure in our test to make it a more comprehensive study. In every setting of learning algorithm, feature selection, and LOOCV, we change the number of retained features, and evaluate the classification result using accuracy and F1-score.

6.3 CLASSIFICATION RESULT

In this section, we first present the performance of user-specific model using different feature selection methods (Section 6.3.1). Then we investigate the optimal feature set for the classfication task (Section 6.3.2).

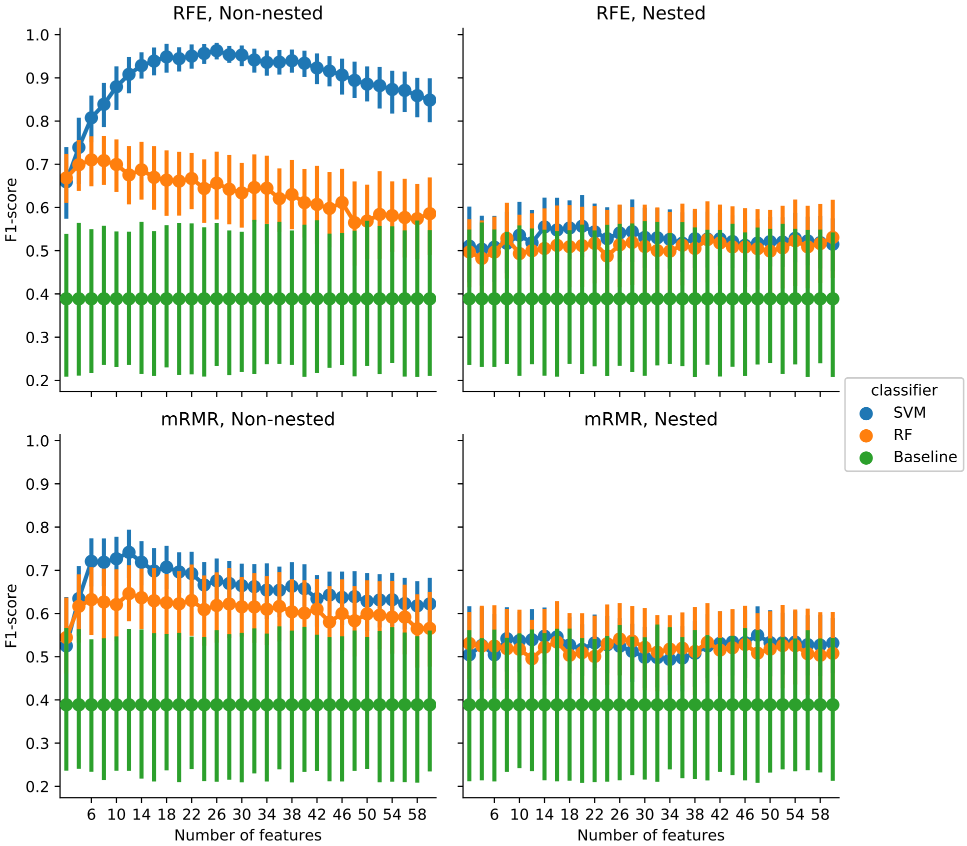


Figure 6.1: Feature selection result for each subject

6.3.1 *Performance of Features Selection Methods*

Figure 6.1 presents the performances of user-specific models using different methods of features selection by changing the size of retained feature sets. We aggregate the result of all participants by presenting a point plot on the F1-score. A point plot represents an estimate of central tendency for a numeric variable by the position of scatter plot points and provides some indication of the uncertainty around that estimate using error bars. For comparison, we present the performance of the a majority vote classifier as a baseline.

We can notice that:

1. Comparison between non-nested LOOCV and the nested LOOCV.
   As expected, the performance of non-nested LOOCV is significantly better than the nested LOOCV, as the model fully fit to the examples of specific given user. The good performance of non-nested LOOCV method signifies the existence of the discriminant features for each user. However, performance of nested LOOCV are low, which signifies the lack of ability of generalization of these features. Concerning the nested LOOCV, with proper size of feature set, the performance is better than the baseline. Meanwhile, the performance is modest. This may be caused by the great noise in the data and scarcity of the examples.

2. Comparison of feature selection using filter (mRMR) and wrapper (RFE).
   For the non-nested methods, the feature selection using wrapper method achieves significantly better performance than the filter method. While in the case of nested methods, both method achieves similar results.

3. Comparison of feature selection using Linear SVM and Random Forest.
   For the non-nested methods, the feature selection using Linear SVM method achieves significantly better performance than the Random Forest method. While in the case of nested methods, the SVM method is still slightly better when retaining an appropriate size of feature set.

4. Optimal size of feature sets.
   Based on the feature selection using RFE with non-nested LOOCV, we notice that the performances of model reach an optimal score with a size of feature set about 20. For the other methods, the optimal size of feature set is about 14.

6.3.2 *Optimal feature sets*

Based on the feature selection tests, we set the optimal size of feature set to 14. Figure 6.2 presents the results of feature selection on each subject. Each column presents a subject and each row presents a feature. The elements in the heatmap present the coefficients provided by the Linear SVM model, in which, the red color signifies a positive relationship with the positive class, while the blue color signifies a negative relationship. The darkness of color presents the absolute value of the coefficient, which reflects the importance of the feature in the model. Features that have never been chosen by any subject have been eliminated from the plot. We notice that the selection of features is very dispersed, that different set of features have been chosen for the subjects in order to achieve the optimal result.

The efficiency of a feature can be evaluated by two values, 1) the number of times it has been selected by the subjects (the frequency) and 2) its importance represented by the absolute value of the coefficient (the weight). The most efficient feature should have both high selection frequency and high weight. Figure 6.3 presents the performances of features in terms of selection frequency and weight and the list of best features. We notice that the selection of feature among subjects dispersed a lot. Most features have been selected only a few times. The most relevant signals are Respiration and EDA. Only two features have been chosen by 7 subjects. The majority of features have only been chosen less than 10% of subjects. Subjects can hardly reach a consensus in their feature selection and thus reflect the difficulty in generalizing the classification among different subjects.

6.4 CONCLUSION

In this chapter, we present the learning process of a user-specific model. More specifically, two typical features selection method: mRMR of the filter method, and RFE of the wrapper method under both nested and non-nested LOOCV schemes are applied on the individual dataset to select the most relevant feature set. The performance of the user-specific model are evaluated. We may draw the following conclusions:

- Confirmation of the great individual variability. The presented results show that there exists a great individual variability in feature selection. So that a personalized model should be created instead of the user-independent general model in order to improve the performance.

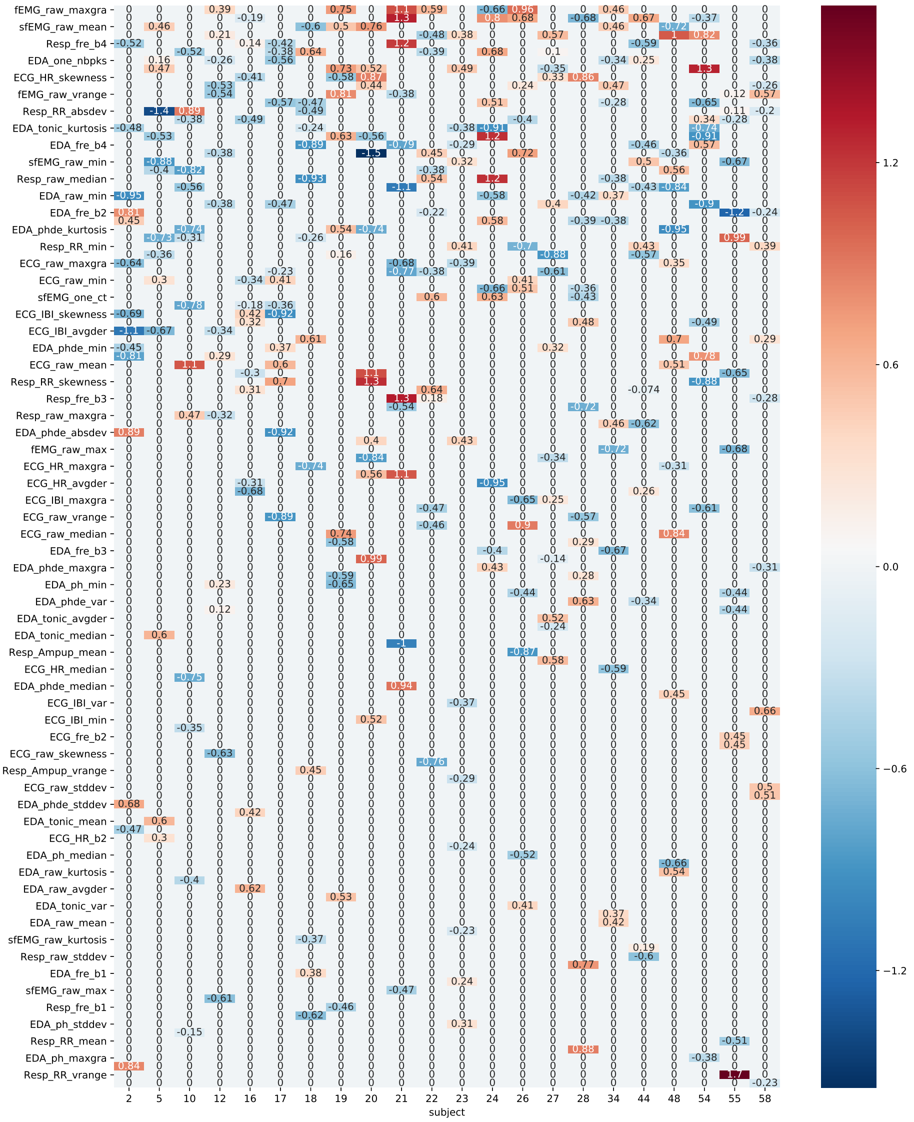


Figure 6.2: Classification results of user-specific model using different feature selection strategies

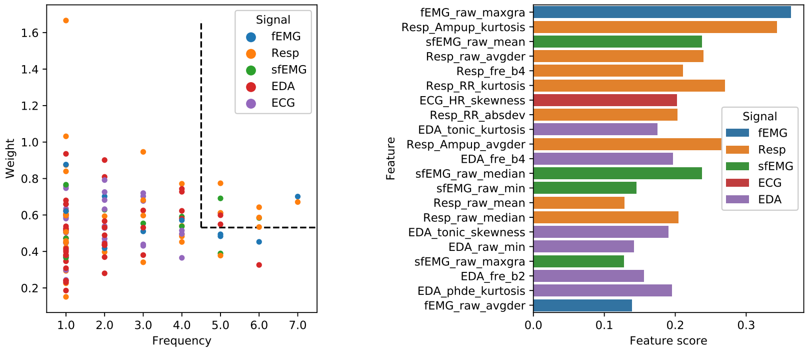


Figure 6.3: Best features for arousal classification

- Lack of practicability of the user-specific model. Even though the user-specific model beats the majority vote baseline model, the performance is still modest. Feature selection using non-nested RFE confirms the existence of relevant features for each subject, however, with the existing data, even by using the LOOCV scheme, the feature selection can hardly generalize to new examples. The poor performance may be explained by the lack of high quality labeled data, which is also a common problem in the practical application.

Acquisition of adequate high-quality, well-labeled data is costly, thus making the application of *user-specific model* unrealistic. In order to improve the performance, one intuition is to create model for similar users, so that the learned knowledge can be transferred. In the following chapter, we present how we find the groups of physiologically similar users using the clustering techniques.

7 INDIVIDUAL CLUSTERING

In the previous chapters, we have presented the performance of classification of emotions using user independent general model (Chapter 5) and user dependent user-specific model (Chapter 6). We have presented flaws by using these two method: general model neglects the individual difference among users and cannot achieve satisfactory result, while user-specific method requires enough high quality labeled training data which is difficult in real application. In order to improve the performance, the individual variability and the lack of data concerns should both been taken into consideration. One intuition is to create model for physiologically similar users, so that knowledge learned can be transferred within the group of similar users.

In this chapter, we first investigate how we find the physiologically similar users using the clustering techniques, three views (signal, feature and model) of individual clustering are presented (Section 7.1). Then we present the learning process on the similar user groups based on these three views. In the end, we present the evaluation of the models.

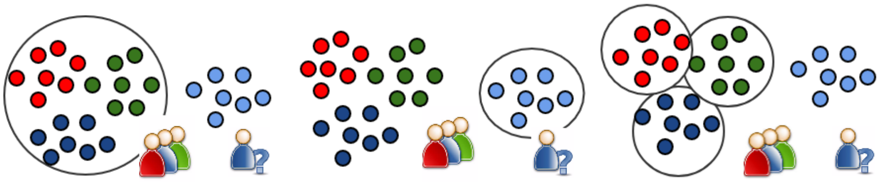
7.1 INDIVIDUAL CLUSTERING

Figure 7.1 summarizes and illustrate the three strategies to create the emotion recognition models:

- **general model** (Figure 7.1a) based on all the existing data, and applied to new user without taking into account the characteristics of the new user. Due to the large individual variability, the performance is not satisfactory.

- **user-specific model** (Figure 7.1b). In this case, one specific model is created for a specific user. An advantage of user-specific model is that it is fully adapted to the given user. However, in practical case, we can not always get enough annotated data for everyone to train a specific model. The lack of data may result in great variance in the model's performance and make the model sensible to noise.

- **group-based model** (Figure 8.4). In this case, instead of creating specific model for each user, we try to personalize the model by using



(a) General model (b) User-specific model (c) Group-based model

Figure 7.1: Three strategies to train recognition models

knowledge learned from the similar users. So that the similar users form groups, that contain more data and can provide more robust model.

Clustering methods group a set of objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in the other groups (clusters). Individuals can be grouped by different criteria. There are mainly two approaches for analyzing the underlying individual difference patterns: one based on social traits, and one based on physiological signals themselves. The former approach relies on the social attribute such as personality traits (Canli et al. 2002; Canli et al. 2001; Amin et al. 2004), gender (Bradley et al. 2001; Bailenson et al. 2008) or culture background (Eid et al. 2001; Ekman et al. 1979). It is claimed that these factors can substantially modulate the neural bases of emotion processing in brain regions such as prefrontal and limbic region and can influence emotional reactions, emotional memory, emotion perception (Hamann et al. 2004) and further influence the physiological response. However, the evaluation and acquisition of the social traits is sometimes subjective and lacks validation (Niven et al. 2011). Physiological mechanism as well as other unknown factors may also influence physiological response besides the social traits, which reinforce the need of a more objective analysis. From this perspective, the second approach focuses on the patterns of the physiological signals themselves and reveals the individual variability based on the signals' characteristics. Physiological-based IRS (Individual Response Specificity) models (Li et al. 2014) were created using data from subjects base. K-means clustering method was applied to cluster subjects into different groups. The result shows that the created IRS model performs better than the general model. Individual differences stem from factors such as personality, culture background, or gender. Such differences widely exist and indeed impact the physiological response. We try to investigate the possibility of clustering individuals by revealing the characteristics associated with the physiological signals themselves.

The common learning process constitutes the steps of transforming the *signals* to *features*, and then to *models*, and thus offering 3 views of analyzing individual differences. In this section, we first present these 3 views of clustering the individuals.

7.1.1 *Three views of clustering*

7.1.1.1 *Signal View*

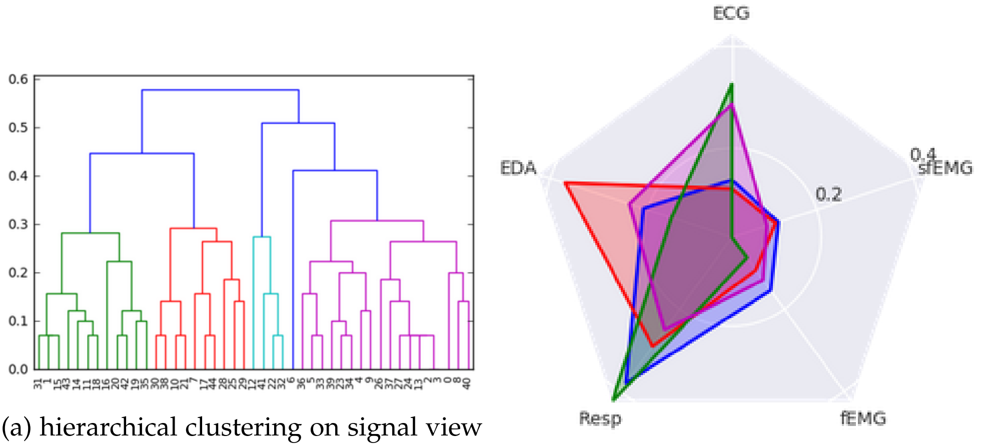
Signal View supposes that the individual differences rely on the signals. Different individuals dispose different discriminant signals to reflect the emotional state. For example, subjects who are easy to sweat may have strong feedback on EDA signal; subjects who are sensitive in muscle activity may have strong response on EMG signal, etc. Based on this assumption, the frequencies of signals appear in the feature sets are counted for each participant. Then we use hierarchical clustering to cluster the subject based on the signal appearance frequency.

Formally, the dataset for clustering D has N samples and M attributes, where N signifies the number of subjects in test (in our case, $N = 45$) and M denotes the number of signals from which we have extracted features (in our case M = 5, that are ECG, EDA, Resp, fEMG, sfEMG). For each subject i, sample $X_i = \{x_{ij}, j = 1, 2..., M\}$, in which x_{ij} signifies the frequency of the corresponding signal used in the best features set. As the size of the feature set is 20 (best user-specific model feature set size obtained in Section 6.3), we have:

$$\sum_{j=1,2,..,5} x_{ij} = 20$$

Based on dataset D, we carry out a hierarchic clustering using euclidean distance and complete linkage method. The hierarchical clustering result is presented in Figure 7.2a. We notice that the most evident cut in the hierarchy highlights 4 distinct groups (we remove the subject 6, as he himself form one specific group).

By aggregating the features using mean value in each group, Figure 7.2b presents the detailed characteristics of the 4 groups: green group is more associated with ECG and Respiration, red group is more associated with EDA and Respiration, blue group is more associated with Respiration and purple group is more associated with ECG. This observation confirm the existence of huge individual difference among individuals even in signal level.



(a) hierarchical clustering on signal view

(b) 4 cluster of individuals based on signal view

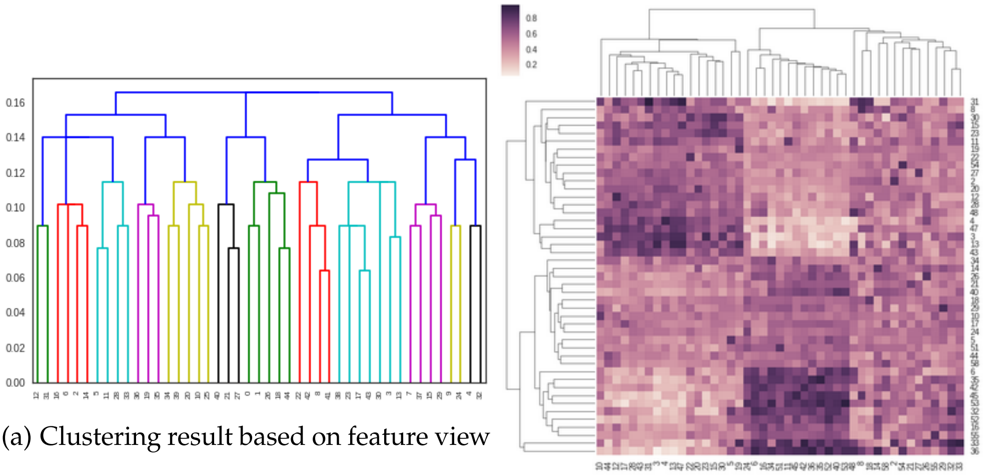
Figure 7.2: Clustering result based on signal view.

7.1.1.2 *Feature View*

Feature View supposes that the individual difference are related with selected features. Individuals who have more resemblance on relevant features may have similar characteristics. By investigating the presence of 20 features, we use hierarchical clustering to cluster the subjects. Figure 7.3a presents the result of clustering based on feature view. Due to the large number of possible combinations of feature sets, valued as $C_k^n = \frac{n!}{k!(n-k)!}$ where n denotes the feature set base size and k denotes the number of features selected for each subject (20 in our case), individuals select different set of features, so that no clear similarities can be found on this level and no clear cluster group can be found according to the feature view.

7.1.1.3 *Model View*

Model View supposes that individual differences can be investigated by using model. A user-specific model can be generalized to other users if they are similar. Based on this idea, we evaluate the performance of each user-specific model on the other individuals, and group the individuals based on this performance. The process is as follows: first, user-specific models are created for each subject using the linear SVM algorithm thus resulting in 45 models, one for each person; then, each user-specific model is evaluated on other subjects using F1-score metric, thus resulting a matrix of 45*45 (number of



(a) Clustering result based on feature view

(b) Clustering result based on model view.

Figure 7.3: Clustering result based on feature and model view.

user-specific model * number of users); in the end, hierarchical clustering is used to group the individuals based on this matrix of F1-scores.

Figure 7.3b presents the clustering result. The color map shows the F1-scores, with dark color representing a high accuracy and light color representing a low accuracy. Each column represents a user-specific model's prediction accuracy on different subjects. From the vertical axis, we notice that the subjects are generally clustered into 4 groups. Moreover, the dark color blocs indicate the effectiveness of applying user-specific model on the others group members and reflects high similarity between the group members.

7.1.1.4 *Remark*

Analyzing the individual differences from the view of *signal*, *feature* and *model* result in difference clustering groups. Among them, *signal* and *model* view result in 4 clusters. In order to form comparable number of clusters for the *feature* view, we roughly split the individuals into 4 groups based on hierarchical clustering result. Therefore, we obtain clustering results based on *signal*, *feature* and *model* options.

Table 7.1: Clustering results based on signal, feature and model view

	Signal
Group 1	3, 18, 19, 20, 22, 23, 24, 36, 44, 54, 55
Group 2	11, 14, 21, 26, 30, 33, 34, 35, 48, 58
Group 3	16, 27, 40, 53
Group 4	2, 4, 5, 6, 8, 10, 12, 13, 17, 28, 29, 31, 32, 42, 43, 45, 47, 51, 52
	Feature
Group 1	4, 8, 10, 14, 15, 16, 18, 20, 23, 24, 30, 33, 40, 42, 43, 44, 45, 51
Group 2	2, 3, 22, 26, 31, 32, 52, 58
Group 3	5, 12, 17, 21, 27, 28, 35, 48, 53, 54, 55
Group 4	6, 11, 13, 19, 29, 34, 40, 47
	Model
Group 1	2, 8, 11, 12, 15, 19, 20, 22, 23, 27, 28, 30, 31, 48, 54
Group 2	3, 4, 13, 43, 47
Group 3	5, 10, 14, 17, 18, 21, 24, 26, 29, 34, 40, 44, 51, 58
Group 4	6, 16, 32, 33, 35, 36, 42, 45, 52, 53, 55

7.1.2 *Clustering result*

The clustering based on signal, feature and model provides different groups. Table 7.1 presents the index of the subjects distributed in each group. In this section, we try to investigate whether the clustering can help to group the subjects and improve the classification performance. We first present the learning process and then the results of classification performance of different groups under different clustering views. No clear relationship can be found between these groups and the meta information such as player's gender, game skill level. We suppose that these clustering are more closely related to the the physiological characteristics. In the future, if more subjects and more detailed information on the subjects such as their personality, physical condition are given, a more detailed analysis can be carried out to find the relevance and the meaning of the clustering result. At the present, we only focus on these 3 views of clustering, and investigate whether a better performance can be achieved by grouping the similar users together.

7.2 LEARNING PROCESS

The learning process is detailed as follows:

- **Dataset:** For each of the signal, feature and model view, we test the classification accuracy on each group.
- **Feature selection:** RFE based on linear SVM is used to select the best 20 features.
- **Cross-validation:** 5-fold cross-validation and the average F1-score is reported
- **Classifier:** We tested the performance of decision tree, linear SVM, naive bayes, nearest neighbors ($K = 3$), RBF SVM and random forest ($N = 50$). We also present the performance of general model by taking into account all the 45 subjects and the baseline is given by using a majority classifier.

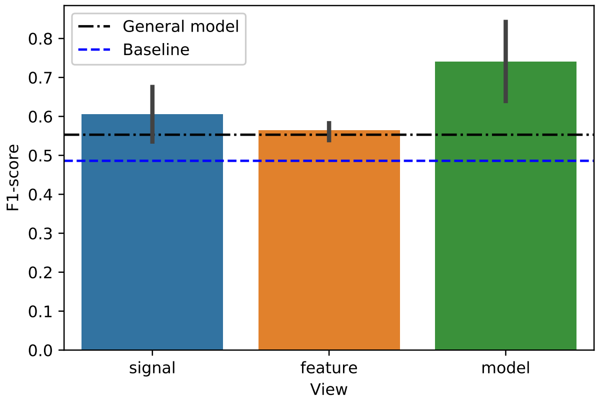


Figure 7.4: Classification result based on clustering groups

7.3 CLASSIFICATION RESULT

Table 7.2 presents the performance of classification for each group under different views. For comparison, we present also the accuracy of a general model and the accuracy of a baseline (a naive algorithm: the majority classifier). We notice that among all the learning algorithms, linear SVM, neural network and random forest achieve better performance than others. The general model obtains performance equivalent to the baseline which is not satisfying. The proposed group-based models achieve better performance

Table 7.2: Average Accuracy of the Group-Based Model with Difference Clustering Settings

	Decision Tree	linear SVM	Naive Bayes	Nearest Neighbors	Neural Net	RBF SVM	Random Forest
Signal							
Group 1	0.416	**0.530**	0.510	0.425	0.504	0.439	0.430
Group 2	0.536*	0.582*	0.616**	0.493	**0.639****	0.539*	0.533
Group 3	0.554*	0.692**	0.658**	0.575*	**0.712****	0.637**	0.665**
Group 4	0.483	0.525	0.495	0.495	**0.541***	0.522	0.520
Feature							
Group 1	0.483	**0.525**	0.499	0.501	0.513	0.473	0.480
Group 2	0.502	0.537	0.549*	0.513	**0.584***	0.541	0.502
Group 3	0.533	0.548	0.520	0.542	**0.581***	0.568*	0.542
Group 4	0.464	0.535	0.439	0.491	**0.569***	0.531	0.444
Model							
Group 1	0.587*	0.670**	0.569*	0.599**	0.670**	0.655**	**0.674****
Group 2	0.742**	**0.856****	0.736**	0.792**	0.856**	**0.856****	**0.856****
Group 3	0.573*	**0.605****	0.510	0.563*	0.605**	**0.605****	0.578*
Group 4	0.727**	**0.829****	0.750**	0.770**	0.829**	**0.829****	0.817**
General	0.466	0.507	0.500	0.472	0.553	0.487	0.461
Baseline	0.486	0.486	0.486	0.486	0.486	0.486	0.486

Stars indicate the accuracy is significantly higher than the baseline according to t-test ($** : p < 0.01, * : p < 0.05$).

than the general model. Figure 7.4 aggregates the results under different views. Among the 3 clustering options, *model* level clustering achieves the best result, followed by *signal* level clustering. The *feature* level based clustering achieves only modest result compared to the general model.

By applying clustering, we achieve our goal of improving the performance of general model. Compared with general model, the clustering groups result in smaller dataset with more relevant examples in the same group. Therefore relevant feature space is reduced and easier to achieve a good model. It should be mentioned that the increasing performance doesn't just come from the reduced size of dataset but more importantly from the increase of similarity in the data. For example, in the feature based clustering view, the Group 2 and 4 also have relatively small number of examples but still can't achieve equivalent performance as in the model based clustering view. This may be explained by the fact that model-based clustering is more capable of discovering the underlining relevance among individuals than the feature based clustering.

7.4 CONCLUSION

In this chapter, we applied clustering based on 3 views: signal, feature and model and by implementing classification on the clustered groups we showed that model-based clustering groups provide the best performance as it is more capable of discovering the underlying relevance among individuals than the feature-based and signal-based clustering. In the following chapter, we present formally how we realized the personalized emotion recognition by training the group-based model.

8 PERSONALIZED EMOTION RECOGNITION MODEL

In the previous chapter, we reveal that models trained on similar user groups can achieve better performances in emotion prediction. The advantages of group-based model are 3 folds: 1) by investigating similarity, the model has been better tailored to the given new user; 2) by clustering similar users together, group-based models are trained with more relevant examples thus resulting in more robust performance; 3) it retains several group-based models based on clustering result instead of creating a separate model for each individual. The memory space can be saved for model storing.

In this chapter, we firstly present the approach to construct group-based model (Section 8.1), then we present how the model is trained on our dataset (Section 8.2), in the end, we evaluate the performance of the proposed model (Section 8.3) and its implementation (Section 8.4).

8.1 CONSTRUCTING THE GROUP-BASED MODEL

The procedure of constructing the group-based model can be summarized as follows: existing individuals are clustered based on certain criteria. In the previous chapter, among the 3 views of clustering, the model-based clustering achieves the best performance. Hence, we applied it in the group-based model to group the users based on the user-specific model performance matrix. Based on the result of clustering, a separate model is created for each clustering group. For a given user, a small amount of user-specific data is used to decide which group-based model is the most representative. The group-based model which achieves the best performance on the given user is used to predict his/her emotional state. The resulting model, even though not created from user-specific data, is customized to the given user to some extent. Formally, we divide them as *training process* and *prediction process*, detailed as follows:

8.1.1 *Getting individual cluster*

We have a dataset D containing training data for N subjects S_i, with $1 \leq i \leq N$:

$$D = (S_1, S_2, ..., S_N)^\intercal$$

Training set for each subject:

$$S_i = (x_1, x_2, ..., x_{n_i})^\intercal$$

with $(x_1, x_2, ..., x_{n_i})^\intercal \in \mathbb{R}^{n_i \times m}$, where the n_i denotes the number of examples for the ith subject and m denotes the number of features (features refers to the physiological features described in Chapter 4 such as average heart rate, etc.).

We apply model-based view (Chapter 7.1) to cluster subjects. In order to obtain similar user groups, the $k - means$ clustering is used on the user-specific model performance matrix $P \subseteq \mathbb{R}^{N \times N}$. The matrix P is obtained as follows: The user-specific model for S_i is denoted as M_{S_i}, and the performance (such as F1-Score) of model M_{S_i} on subject S_j is denoted as $Perf(M_{S_i}, S_j) \in \mathbb{R}$ so that each row P_i of the performance matrix P is:

$$P_i = (Perf(M_{S_1}, S_i), Perf(M_{S_2}, S_i), Perf(M_{S_3}, S_i), ..., Perf(M_{S_N}, S_i))$$

The performance matrix P which represents the models performances of N subjects can be clustered into K groups $(C_1, C_2, ..C_K)$, with each C_g containing the list of subjects belong to this cluster. The sum of distances of samples to their closest cluster centroid is denoted as *inertia* and it is a measure of the clustering quality.

8.1.2 *Training process*

8.1.2.1 *Training model for each cluster*

Using the K clusters $C_1, C_2, ..C_K$, the original dataset D can be now divided into K groups, with each group D_g containing all the examples of the subject within the same group:

$$D_g = \bigcup_{i \in C_g} S_i$$

The group-based model M_{D_g} is trained on each D_g. Its training performance of the model M_{D_g} is called predicting ability (PA) of the model,

$$PA = Perf(M_{D_g}, D_g)$$

It also gives the upper limit of the performance of the group-based model.

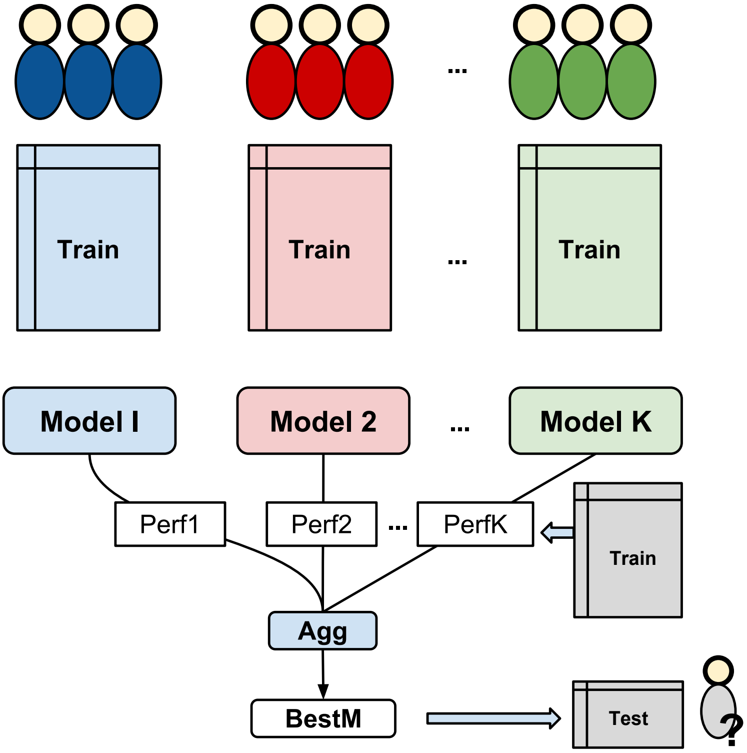


Figure 8.1: Training process

8.1.2.2 *Assigning the new user to a relevant group-based model*

Given the examples from a new user S_{new}, a relevant group-based model $BestM$ is the model on which the training data of the new user $S_{newTrain}$ achieves the best performance.

$$BestM = \underset{M_{D_{g,(g=1,2...,K)}}}{\operatorname{argmax}} \; Perf(M_{D_g}, S_{newTrain})$$

The performance of the $BestM$ on the $S_{newTrain}$ represents the generalization ability (GA) of the group-based model:

$$GA = Perf(BestM, S_{newTrain})$$

The training process of the group-based model is illustrated in Figure 8.1.

8.1.3 *Prediction process given a new user*

Once the *BestM* is determined it can be used to predict new examples of the new user. The performance of the group-based model (PerfGM) can be defined as:

$$PerfGM = Perf(BestM, S_{newTest})$$

We assume that the performance of *BestM* model *PerfGM* is closely related to the predicting ability (*PA*) and the model's generalization ability (*GA*) .

8.2 LEARNING PROCESS

In order to evaluate the performance of group-based model and comparing with the general model and user-specific model on the DAG dataset, we applied the following learning procedure for the constructing and evaluating the general model, user-specific model, and group-based model. All of the three models are trained and compared using different learning algorithms: decision tree, linear SVM, RBF SVM, naive Bayes, neural network and random forest. We report the performances in terms of the F1-score.

8.2.1 *Evaluating general model*

For general models, all collected instances from every subject are taken and *leave-one-subject-out cross-validation* approach is applied. In each iteration, samples from only one subject are retained for testing, while samples of all the other subjects are used to construct the general model. This process is iterated over all subjects to evaluate the performance of the general model on each subject.

8.2.2 *Evaluating user-specific model*

For user-specific model, data concerning the specific user is used and *leave-one-out cross-validation* approach is applied. For each given subject, only one sample is retained for testing, while all the others are used to train a user-specific model. This process is iterated over all samples of the given subject to evaluate the performance of a user-specific model on each given subject and then iterated over all subjects to evaluate the performance of this model on different subjects.

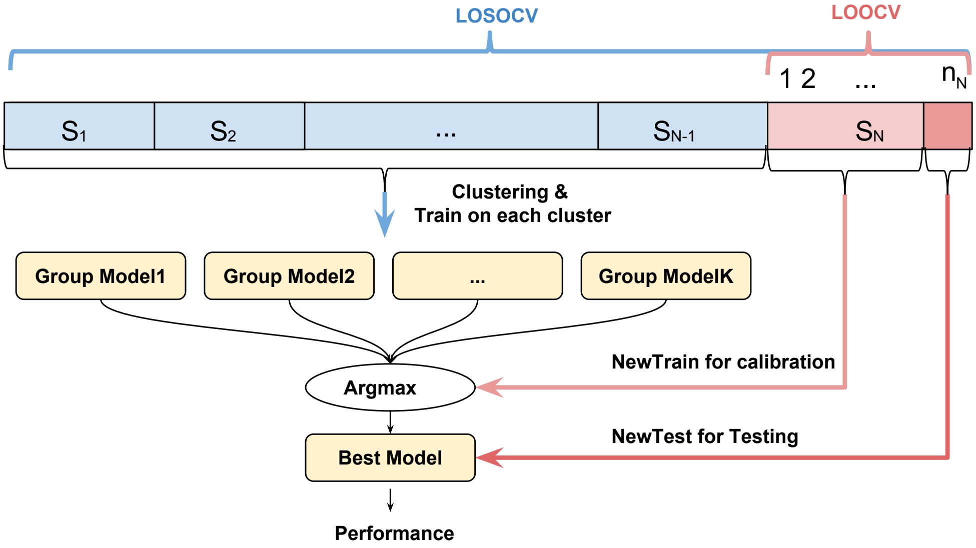


Figure 8.2: Evaluation process of the group-based model

8.2.3 *Evaluating group-based model*

The evaluation process of the group-based model is illustrated in Figure 8.2. Firstly, a *leave-one-subject-out cross-validation (LOSOCV)* approach is applied so that data from one user are used for test (we refer to it as "one user data"), while the rest of the data are used for generating the group-based models (we refer to it as the "rest data"). On the "rest data", the *Getting individual cluster* (Section 8.1.1) and *Training model for each cluster* (Section 8.1.2.1) is applied, and resulting in K models. The method used to get the cluster is K-means. The number of clusters K is decided using the elbow method (Thorndike 1953), and all the K resulting in "elbow" are tested. Next, a *leave-one-out cross-validation (LOOCV)* approach is applied on the "one user data", so that all data except one is used to determined the relevant model $BestM$ among the K models, and the last is used for test. Then, tests are iterated on the *LOOCV* and LOSOCV. In each LOSOCV iteration, we retain the values: the number of resulting clusters (K), the associated *inertia*, the model predicting ability (PA), the model generalization ability (GA), and the final performance of the group-based model (PerfGM) on the test examples.

Table 8.1: Average Accuracies of Personalized Model Using Different Learning Algorithms

	Decision Tree	**Linear SVM**	**Naive Bayes**	**Neural Network**	**RBF SVM**	**Random Forest**
group-based	0.497	**0.631****	0.506*	0.626**	0.617**	0.607**
user-specific	0.458	**0.506***	0.503*	0.504*	0.492*	0.481*
general	0.449	0.304	0.356	0.409	0.323	**0.450***
baseline	0.385	0.385	0.385	0.385	0.385	0.385

Stars indicate that the accuracy is significantly higher than the baseline according to t-test ($** : p < 0.01, * : p < 0.05$).

8.3 PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed group-based models. We first present the results of the group-based model by comparing it with the general and user-specific model and then investigate some characteristics of this model.

8.3.1 *Performance of the group-based model*

Table 8.1 presents the average accuracies of 4 models (general, user-specific, group-based and baseline) using different machine learning algorithms. Each line gives one learning strategy and each column gives one machine learning algorithm. The best accuracy under each model is presented in bold font. More intuitively, Figure 8.3 highlights the performance differences of the 4 types of models using different machine learning algorithms.

We notice that most of *general models* achieve similar performance or even worse performance than the *baseline*. Only *general models* based on decision tree or random forest achieve slightly better than the *baseline*. Most of the *user-specific models* perform better than the *baseline* and *general models* in terms of the mean of the F1-score, however, they all have a large variance compared with *general models*. The *group-based models* obtain consistently better performances than the *general models* and *user-specific models* both in terms of mean and variance of the F1-score, except by using decision tree or naive bayes classifiers.

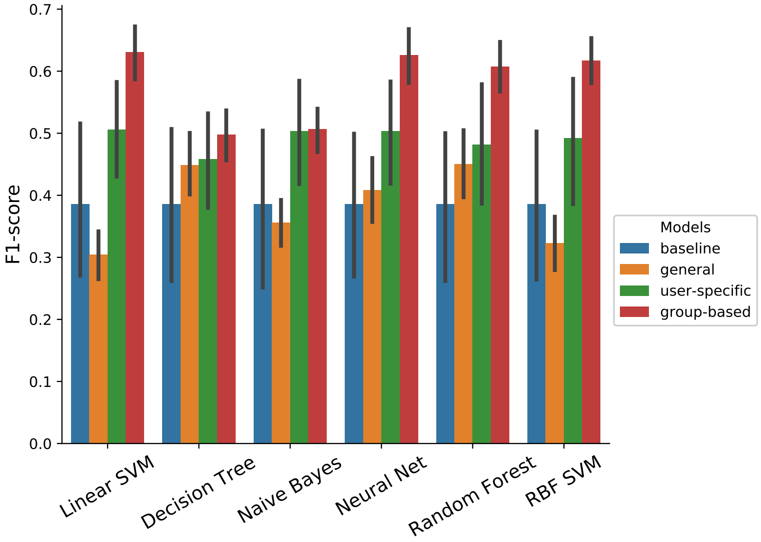


Figure 8.3: performance between the group-based model, general model, user-specific model, and baseline

When comparing horizontally, the most effective learning algorithms are neural network, linear SVM, RBF SVM and random forest, as they always achieve the better performances in all group-based models.

The performance of *general, user-specific, group-based* model generally follow the patterns that:

- The *general model* achieves the a performance equivalent to the *baseline*. This can be explained by the fact that the *general model* doesn't take into account the individual variability;

- The *user-specific model* performs better than the *general model* but has a large variance. Therefore, even though user-specific model is tailored by training on specific user data, it is not robust. The problem may issue from the lack of high quality labeled data has more impacted this model;

- The *group-based model* performs better than the *general model* and *user-specific model* both in terms of average value and variance. This result validates the efficiency of the proposed *group-based* method, as it provides a personalized and a robust model.

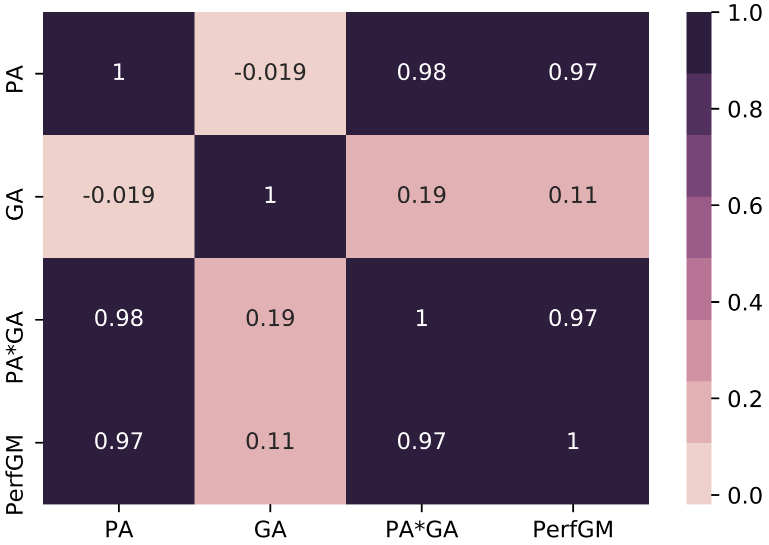


Figure 8.4: Correlation Matrix PerfGM, PA, GA

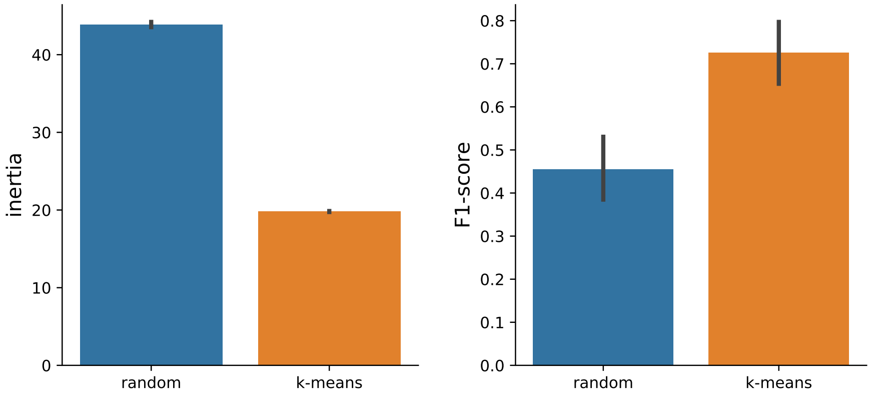
8.3.2 *Relation between the PerfGM and PA, GA*

The performance of the group-based model ($PerfGM$) is closely related with the model's predictions ability (PA) and the model's generalization ability (GA). Figure 8.4 presents the relationships between the PA,$PA * GA$ and $PerfGM$. We can notice that there is a strong positive correlation between PA and $PerfGM$, $PA * GA$ and $PerfGM$.

This phenomenon can help evaluating the performance of *group-based models* even before applying the model on the new user. High $PA * GA$ score means that there is a great chance that the *group-based models* can perform well on the user, while a low $PA * GA$ score signifies that they are not applicable to the given user. In this case, new data should be collected and a new model should be trained to achieve a meaningful prediction.

8.3.3 *Relation between quality of cluster and performance of group-based model*

One important process of the *group-based models* is to find the groups of similar users. In this section, we are interesting to see whether the construction of such groups really helps to improve the *group-based models*. In order



(a) Inertia

(b) F1-score

Figure 8.5: Relationship between the quality of cluster and performance of group-based model

to inspect this problem, we compare *group-based model* trained on groups formed by the K-means clustering and the random groups.

For both settings, the tests are done by setting the number of groups (K) as 3, as 3 is the optimal number of groups in the K-means clustering found in our previous tests. We evaluate both the clustering quality *inertia* and $F1 - score$ based on this two settings: K-means and random. Figure 8.5 shows that using K-means to find groups of similar users improves the quality of the group-based model, both by reducing inertia(intra-cluster distance) and F1-score.

8.3.4 *Conclusion*

In this section, we proposed a new *group-based model*. By comparing with the *general model*, *user-specific model*, the *group-based model* performs more accurate and more robust. The key innovation point of the *group-based model* is that it lies on a new learning strategy to train the model in the context of highly noisy data with high inidividual variabilities. In the next chapter, we evaluate the computation resource required in the whole learning process, and the process of implementation on an embedded system.

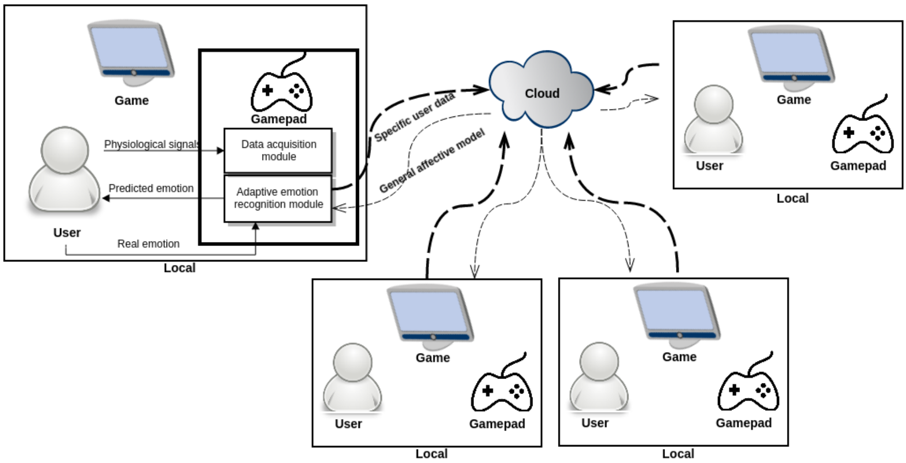


Figure 8.6: Physiology-based emotion recognition used for affective gaming

8.4 IMPLEMENTATION OF THE PERSONALIZED MODEL

In this section, we present whether the proposed personalized emotion recognition model can meet the implementation requirements. We first present a use case of physiological-based emotion recognition system for affective gaming and highlight the implementation issues (Section 8.4.1). Then we make a computation resource profiling of the proposed emotion recognition model by using software simulation (Section 8.4.2). In the end, we implement the one of the solutions on a real embedded system (Section 8.4.3).

8.4.1 *Framework of physiological-based emotion recognition system*

In real application, the personalized recognition model should be improved as more users are coming and more data is accumulated. One way to take advantage of this growing data is to connect the terminal to a cloud (Figure 8.6). This framework is composed of two parts: the cloud and the terminal.

The functionality of the cloud is:

- storage of the emerging data
- training the personalized group-based model and send to the terminal
- updating the model with the new collected data

The functionality of the terminal is:

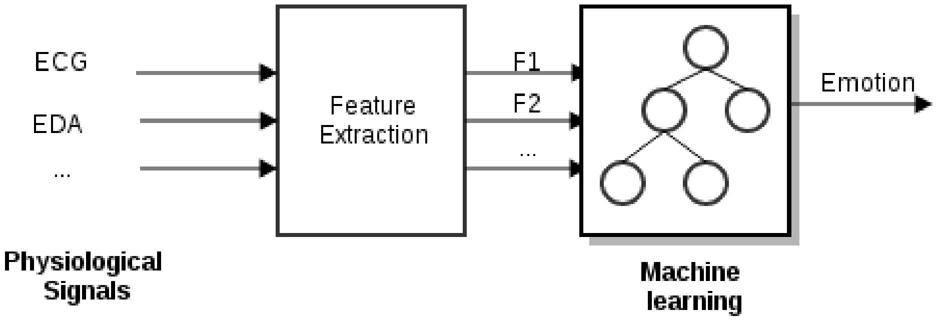


Figure 8.7: Computation process on the terminal

Table 8.2: Simulation environment

Processor	Intel(R) Core(TM) i7 CPU 950 @ 3.07GHz
CPU (MHz)	1600.000
CPU max (MHz)	3068.0000
CPU min (MHz)	1600.0000

- collecting the physiological signals and extracting the features
- using the model trained on the cloud to recognize the emotions

Our focus in on the terminal part. More specifically, as our objective is to realize a near-real-time emotion recognition during game playing, one critical requirement is that, in the prediction phase, the computation speed should be faster than the physiological signals inputting speed. In the following section, we evaluate if the proposed method can meet this requirement.

8.4.2 *Computation time simulation*

The computation process on the terminal is composed of 2 steps (Figure 8.7), *feature calculation* and *model prediction*. We evaluate the computation time of these 2 steps using a PC. The code of feature calculation and model prediction have been written in Python and ran on a computer under Linux system with the settings shown in Table 8.2

The feature calculation step is evaluated on a physiological signal segment of 20 seconds (appropriate segmentation size for emotion recognition task presented in Section 5.1.2). In order to reduce the influence of other processes of the machine, the computation is iterated 1000 times, and the average time is reported.

Table 8.3: Computation time

	Linear SVM	RBF SVM	Decision Tree	RF	Neural Net	Naive Bayes
Feature extraction (ms)			2293			
Prediction (ms)	0.122	0.140	0.065	**5.957**	0.135	0.199

The model prediction step is evaluated by constructing the model on 600 examples (data from one third of subjects), with 20 features (best user-specific model feature set size obtained in Section 6.3). We test different machine learning models: decision tree, linear SVM, RBF SVM, naive bayes, neural network and random forest. The computation has been ran for 10 000 times, and the average time is reported.

Table 8.3 presents the time required to make one prediction from a physiological signal segment of 20 seconds. The feature extraction step takes 2293 ms, which takes great part of time in the whole process. The prediction takes much less time, about 0.1 ms for most of the cases. Among them, model decision tree is the fastest one, followed by linear SVM and neural network. random forest is the worst. Compared with signal arriving time interval of 20 s, all learning algorithms are able to meet the time requirement.

8.4.3 *Implementation on an embedded system*

The simulation validates the feasibility of the proposed method on a computer. We implement the terminal side computation process on an embedded system. The test environment is a ZedBoard [1], with Dual A9 ARM processor which has up to 667 MHz operation. The feature extraction and model prediction processes are implemented on C. Finally, the feature extraction took 7600 ms and prediction took 1.02 ms, which still meet the time requirement.

8.4.4 *Conclusion*

In this section, we come over implementation issue of the proposed model. By evaluating the time required for terminal side computation process, we validate the feasibility of the proposed method in terms of time requirement.

[1] http://zedboard.org/

9 CONCLUSION AND PERSPECTIVE

In this chapter, we summarize the thesis achievements and discuss the perspectives.

9.1 ACHIEVEMENTS AND CONTRIBUTIONS

Our objective is to realize an automatic emotion recognition system and validate its implementation on an embedded system. We achieve our goal by dividing the problem in three aspects: affective gaming, emotion recognition model, and implementation.

9.1.1 *Affective gaming*

Concerning affective gaming, the contributions are composed of two parts: a literature review involving the critical affective gaming issues (Chapter 2) and an experiment for creating physiological-based affective gaming database (Chapter 3).

Literature review involving critical affective gaming issues

Regarding the literature review on affective gaming, the following questions are examined: how the emotions can be represented (emotion representation, Section 2.1)), how the emotions can be measured (measuring modalities, Section 2.2) and how the emotions can be assessed (subjective assessment 2.3). A literature review concerning the critical elements of physiological-based affective game research (Section 2.4) is presented.

1. **Emotion representation**. We review the different emotion representation theories and highlight the advantage and disadvantages of the two major models: categorical and dimensional. Categorical representation is intuitive and concise, however, there are issues such as vagueness of language, cognitive bias and it cannot handle the subtle differences or changes of emotions. On the other hand, dimensional representation is compatible with the categorical representation and, at the same time, provides continuous and quantitative measures. It also solves the

problem of vagueness or limitation of language. However, it may sometimes suffer from a loss of information, due to an inappropriate setting of the emotion dimension. By carrying out a review on the emotion representation in the game research, we showed that the choice of the representation should be adapted to the game context.

2. **Measuring modalities**. We review common measuring modalities in affective computing, *behavioral modalities* such as vision-based facial expression, gesture/posture, audio-based such as voice modulation, Human-machine-interaction modalities such as event log, pressure on button, and *physiological modalities*, such as central nervous system signal EEG and peripheral nervous system signals ECG, EDA. We conclude that physiological modalities are adapted to affective gaming as it provides an objective, continuous, quantitative, real-time way to assess the user's inner state, and cost fewer computation resources (time series analysis cost less than video analysis).

3. **Subjective assessment**. We review common methods for collecting self-reported assessment in affective computing: questionnaire, interview, focus group, or think aloud. We found that the last 3 approaches contain rich information on the player's game experience but are time-consuming to process, so that method such as interview can be applied in the preliminary study in order to explore all the questions related to the design of the future experiment, and then contribute to make a relevant questionnaire for the effective study.

Construction of physiological-based affective gaming database

Existing open databases concerning physiological signals for emotion recognition task use music/image/video as affective stimuli which cannot fit the video game context. Therefore, we have designed and carried out an experiment to collect data. We introduce the DAG database[1] - a multi-modal **D**atabase for **A**ffective **G**aming. The database contains 1730 annotated emotional events, 174 game-experience ratings, 58 match rankings by 58 participants. Each participant played 3 matches of soccer simulation game[2] at various difficulty levels. In total, about 116 hours of physiological signals have been recorded. We focus on peripheral physiological signals (ECG, EDA, Respiration, EMG).

[1] http://erag.lip6.fr

[2] FIFA 2016 soccer video game by Electronics Arts.

Two kinds of self-assessed evaluations are available in the database: minor scope evaluation on *game event* and global scope evaluation on *game sequence*. In terms of evaluation on *game events*, players annotated the events based on two most widely accepted approaches for modelling emotions: categorical representation and the dimensional representation. For categorical representation, happiness, frustration, anger, fear, and boredom have been used. For dimensional representation, the arousal/valence model has been used, with arousal ranging from inactive (e.g. bored) to active (e.g. excited), and valence ranging from negative (e.g. sad) to positive (e.g. happy), both within a scale of 7 values (ranging in [-3,3]). Regarding the evaluation on *game sequence* experience, GEQ has been used to evaluate each game sequence in terms of **D**ifficulty, **I**mmersion and **A**musement (DIA).

9.1.2 *Emotion recognition model*

Based on the data from DAG database, we presented different emotion recognition models: general emotion recognition model (Chapter 5), user-specific emotion recognition model (Chapter 6), and personalized emotion recognition model (Chapter 8).

General emotion recognition model

Concerning general model, models are trained without making the differences among different subjects. We present a set of analyses concerning:

1. *emotional moment detection*: realized by classification of the sequences with and without annotations. Effects of segmentation lengths and relevant features are discussed. We found that 1) events with high arousal level are more detectable than those with low arousal level, 2) different event types don't make a significant difference on event detection efficiency. Detection efficiency of different events is more dependent on the emotion than on the event type. 3) the detection accuracy reaches the best when taking the short segmentation lengths (10 s or 14 s).

2. *emotion recognition*: realized by classification of emotions on game events. Effects of segmentation lengths and relevant features, as well as three normalization methods (standard normalization, neutral baseline referencing normalization, precedent moment referencing normalization) aiming to reduce individual variability are discussed. We found that 1) low/high arousal (LA/HA) recognition performs better than the

low/high valence (LV/HV) recognition, 2) using precedent sequence as reference outperforms the the standard normalization and the neutral state referencing normalization, 3) the best segment length for emotional state recognition is longer than the detection task (14 s or 20 s), 4) performance of emotional state recognition is modest which illustrates the difficulty of using physiological-based method to recognize emotional state in a dynamic context.

3. *game sequence evaluation*: realized by preference learning on match rankings. We noticed that: 1) difficulty is easier perceived than immersion and amusement 2) player's perceived difficulty, immersion and amusement can be better understood by taking into account the game level, game outcome and in-game events factors: (1) Regarding difficulty: A player's perception of game difficulty is best understood as a function of how much they score and/or are aroused, also if they score poorly, how much they improve within a match. (2) Regarding immersion: A player's immersion is tied to their experience of happiness, anger, and performance (goals and improvement). (3) Regarding amusement, the key factors are valence, happiness, and goals scored.

User-specific emotion recognition model

Concerning user-specific model, models are trained on each of the subjects. Different feature selection methods (filter and wrapper, nested LOOCV and non nested LOOCV), the optimal size of features were investigated and user-specific models were trained on each subject. We found that 1) wrapper method works better than filter method to find the optimal feature set. 2) Models trained on non-nested LOOCV feature selection is much better than the nested LOOCV one, which signifies the optimal feature set can hardly generalize to the new examples. 3) Relevant features are quite different regarding different subjects.

Based on these observations, we confirmed the 1) great individual variability exists among people. So that a personalized model should be created instead of the user-independent general model in order to improve the performance.2) Lack of practicability of the user-specific model. Acquisition of adequate high-quality, well-labeled data is costly, thus making the application of user-specific model unrealistic. In order to improve the performance, the ideal model should both take into consideration the individual variability and make the most use of existing data.

Personalized emotion recognition model

Based on the results of the general and user-specific models, we proposed a new group-based model that both take into account the individual variability and make the most use of existing data. The idea is that creating clusters of individuals (Chapter 7), so that similar users are grouped together; and training group-based model for similar users (Chapter 8), so that learned knowledge within the group can be transferred to the new user.

1. **Clustering of individuals**. Different criteria can be used to find similar users such as social trait, gender, personality. We focus on the physiological characteristics and group the subjects based on 3 views: signals, features, and model performance, and evaluate the recognition on the clusters based on these 3 views. We found that 1) based on signal sensitivity view, subjects can be grouped into the 4 groups: the group that is sensible with ECG and respiration, the group that is sensible with EDA and respiration, the group that is sensible with Respiration and the group that is sensible with ECG. This confirms the existence of individual differences on the signal level 2) based on the model view, the user-specific model of one person sometimes performs well on the other persons, which confirms the similarity among the subjects. 3) by evaluating the performance of emotion recognition on each group based on the 3 views, the model-based view achieves the best performance which signifies that it performs well in finding the similar users.

2. **Group-based model**. We proposed an approach to train the group-based model, evaluate the performance by comparing with the general model, user-specific model, and present some property of the proposed method: we found that 1) the proposed group-based model performs better than the user-specific and general model both in terms of average and variance of F1-score. That means the group-based model is better and more robust. 2) the performance of final prediction is highly and positively correlated with predicting the ability of the model trained on the user group. 3) the step of clustering does influence the performance of the group-based model.

9.1.3 *Embedded system implementation*

In Section 8.4, we evaluate the computation time for the personalized emotion recognition model. We realized a simulation on a computer and an implementation on a real ARM A9 embedded system. On the embedded system implementation, the feature extraction process is much longer than the model prediction time. The whole process has proven that the proposed solution can meet the time requirement.

9.2 PERSPECTIVES

The major contributions of this thesis concern affective gaming, emotion recognition model, and implementation on an embedded system. In this section, we give some future works that can be promising in these 3 fields.

9.2.1 *Affective gaming*

Concerning the affective gaming, the main contribution is the construction of a new multi-modal and multi-scale assessment affective database DAG. Future work can be dedicated to both the improvement of the current dataset and construction a new dataset.

Improvement of DAG

Future work can be done to improve the current DAG database in the following aspects.

Taking advantage of the entire time table. The current work has been dedicated to the game-play events and evaluate the relationship between physiology and emotion. Meanwhile, the DAG database contains richer information such as the complete physiology recording during the whole process of experiment. The time table for each experiment phase (music, game, questionnaire, annotation) can also be made available and be used to analyze the physiological responses under different contexts.

Realize a more robust event annotation. In the current study, the decision to annotate one event or not is totally given by the subject, which may suffer from cognitive error or memory issues. In order to have a more robust method of evaluation of game events, one possible way is to extract a game event log or realize an objective game event annotation by the observers. Based on these controlled annotations, the relationships between no-event segments, events without self-reported annotations, and events with self-reported annotations can be different from those that we found. Besides, global game experience can be investigated base on the game event sequence.

Refer to observer assessments. Current assessments are produced by subjects' self-report. These assessments can be enriched by observers to provide a more complete and objective evaluation of the subjects' felt emotions. Observers can be field expert, for example, psychologists who know how to interpret the micro-expressions on face or physiologists who know how to read physiological signals, or crowd-sourcing workers.

Expand the database

More experiments can be carried out to expand the database. The affective gaming database oriented for emotion recognition can be characterized by 3 dimensions: stimulation, objective modalities, and subjective modalities (Section 3.1).

Stimulation/game selection. The current game has been chosen because it provides an immerse interactive experience with various types of emotions. At the same time, the emotion types are relatively limited and related to the game events, which make them easier to analyze. Meanwhile, with the selected game, the limitation is the uncontrollability of the stimulation in our experiment due to the naturalistic game interaction. In perspective, efforts can be made to select or design interactive game stimulation which can be more controllable in terms of emotion type, time, and intensity. This can be a collaborative work of a game design company, a psychologist, and a machine learning practitioner. Besides, an ideal game should also provide different levels of event log, which can serve as an objective reference of player's experience.

New objective modalities. The major objective modalities used in the current study are physiological signals. The choice has been made by finding a balance between unobtrusiveness, processing ease and effectiveness. In the recent years, with the development of intelligent unobtrusive hardware and new processing techniques, measuring of new modalities can be much easier. For example, measuring of EEG had long been requiring uncomfortable EEG cap and long process of electrode calibration. Even though it is viewed to have faster response and richer information, we don't apply it in our experiment due to the fact that it may influence the players' game experience. Recent developments of the EEG headset EMOTIVE[3] has made it less intrusive to use in the game context. It has been showed that, emotion recognition from multi-modal sources can greatly improve the recognition performance (Sebe et al. 2005; Zhalehpour et al. 2016). Without affecting

[3] www.emotiv.com

the player experience, the introduction of more modalities can provide more material for emotional analysis.

New subjective modalities. Both the self-reported and observed assessments should be included in the database. The current assessments are realized on overall game segments and events. Self-reported assessments of emotions on multiple dimensions in real time is not realistic. In perspective, by using observer assessment, real-time assessment can also be included (Cowie et al. 2000; Cowie et al. 2013; Lopes et al. 2017).

Collection of meta-information. We have studied similar users groups based on their physiological characteristics. With the meta-information on subjects (such as personality, demographic information, game habit, game skill, etc.) or on the game (such as game type, game system, etc.), the information can be combined with physiological characteristics and be used to create a better interpretable, and personalized emotion recognition model. Though DAG is a large physiological-based affective database, it is still too small to construct the player model or draw a significant conclusion based on meta-information such as gender or game skill. In perspective, we hope to include these types of information, so that with the accumulated number of databases emerging in the affective gaming domain, it can help us to get a better understanding of players' emotions or experiences.

9.2.2 *Emotion recognition model*

Concerning the model construction, the main contributions are the study of general, user-specific and group-based models based on the proposed database. Future work can be dedicated to both the improvement of the models' performance and implementation of other related tasks.

Improve the performance of current model

Following works can be done to improve the performance of the current model.

Implementing new features. In the current study, time domain statistical and frequency domain features have been applied, because they are robust and don't require extra parameters. Besides, different features can be extracted, such as entropy-based features, morphological features, etc. Also, a deep-learning framework can be used to learn a multi-level signal representation. The choice of the representation should be based on the validated segmentation lengths, as different features of different signals may have different effective lengths. Besides, the alignment of multi-variant input may also influence the final result, especially in the real-time dynamic context. The present work covers some of the most common features used in the peripheral physiological based affective computing. In perspective, more work can be dedicated to find new features extracted from different feature extraction methods or different signal segment lengths.

Using new label processing method. In the current study, techniques such as the discretization of the dimensional evaluation to form a binary classification problem or preference learning to learn the game experience rankings were applied. In perspective, other discretization options, emotion recognition on specific categorical emotions or working directly on dimensional labels can be tested.

Improving the machine learning method. Our study proposed a new method by training group-based model. It has been shown that it can improve the classification performance by finding explicitly similar user groups using clustering techniques. We've tested several classical machine learning algorithms to validate the effectiveness of the proposed training process. Meanwhile, the proposed method is not limited to use a particular machine learning algorithm. In perspective, the machine learning model can be improved in the following aspects:

- Using more advanced model. Advanced models can have a better fitting ability, but are more likely to get over-fitting. In the future, the same learning process can be applied to more complex model such as ensemble models (other than Random forest) and deep learning models. With fine-tuned hyper parameters, a better performance may be achieved.

- Constructing robust model. Real-word data is always full of noises. In the current study, the noise may come from the ineffectiveness of the stimulation, the measuring artifacts, and self-reported assessment bias. This problem can be settled from 2 aspects, data and model. From the data aspect, low-quality data or suspicious data can be removed or given less weight during training. From the model aspect, more robust algorithm such as fuzzy decision tree (Adamo 1980; Damez et al. 2005), new regularization term can be added.

- Considering the scalability of the model. Current method relies on a batch mode to train the model. In implementation, with the accumulation of new users and new data, the online incremental method (Shalev-Shwartz et al. 2012; Schlimmer et al. 1986) should be more appropriate to this case. Based on this consideration, and the fact that proposed method is not model specific, online machine learning methods such as VFDT (Domingos et al. 2000) can be tested. Balance should be found between the updating frequency and models' robustness. Available data in the DAG database may still be too small to run such a test. However, the model updating mechanism is important in real-world application, and should be taken into account in the early stage.

Make new relevant affective models

Besides the emotion recognition model presented in this thesis, the proposed dataset provides the possibility to investigate other relevant affective models.

Relationship between physiology and categorical emotion. Current study has used the binarized dimensional emotion representation of high/low arousal and valence. Meanwhile, in gameplay context, detection of certain categorical emotions are critical to understand and improve the game experience. We have shown in Chapter 3 that there is a close relationship between the event type and categorical emotion. Fused with physiological signal features, the categorical emotions may be better detected.

Relationship between physiology and event. Detection of game events which have triggered emotional response is of vital importance to engage players. A simple event log cannot accomplish this objective, as in the long list of high-level or low-level game event provided by the event log, few are directly relevant to the player's emotional state. One of the short-term perspectives on the database improvement is to realize objective event annotations. Combine with these annotations, one can get a better understanding of why certain events have been annotated and what are the corresponding physiological responses.

Relationship between physiology and overall game experience. Current work focused on the physiological responses on the game events which have a relatively higher time sensibility compared to a lot of existing affective gaming research that only focus on evaluating physiological responses related to overall game experience. Even though, we think that the overall game experience can be better understood with game events, game outcome and other meta-information. It should be mentioned that as the physiological segment on the overall game segment is much longer than on the game events, the processing techniques such as signal processing, features extraction, are also different.

Relationship between physiology and context. Besides the gameplay, the proposed database DAG also contains the physiological signal recording under different contextss (music, game, questionnaire, and annotation). Detection of certain context can help to create the context-related emotion recognition model, an important step of realizing pervasive computing.

Apply the findings to practical application

Game adaptation. The main objective of realizing emotion detection in the game context is to provide sufficient information to the biocybernetic loop, so that the game experience can be improved. To address this objective, the game can be adapted or be set to react to the player in the following manners:

- real-time adaptation. The application of real-time adaptation is closely related to event-based and real-time emotion recognition. The advantage of this adaptation is that it reacts more quicker to the players' emotional states. However, it may take more computation resources, be hard to train and be less robust.

- critical point adaptation. The application of critical point adaptation is closely related to the task such as emotional event detection. For example, constant failures may cause serious frustration and result in the abandon of playing. This problem can be settled by the detection of this breaking point and assist the player to accomplish the goal.

- game-level based adaptation. The application of the game-level based adaptation is closely related to the analysis of overall game experience. By acquiring a better understanding of the players' game experience, a better decision can be made to change the settings (such as game plot, difficulty level, etc) in the following game sessions.

Pervasive computing. This application is closely related to tasks such as context detection, and emotion recognition on the game events or game segment. Understanding emotions in game also improves our knowledge of emotions in real life, as games provide rich emotion experience and interactive experience. The insights got from the affective gaming research can help to improve the development of pervasive computing.

9.2.3 *Implementation on an embedded system*

Concerning the implementation on the embedded system, the presented work was just in the first stage. It only shows that the presented model can meet the time requirement on an A9 ARM processor. Future work can be dedicated to complete the implementation related evaluations, investigate the requirement of hardware design under different use case and software hardware co-design to optimize the general performance of the system.

ANNEXES

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ACC	Accuracy
AG	Affective Gaming
AI	Artificial Intelligence
ARM	Advanced RISC Machine
AUC	Area Under Cover
AV	Arousal Valence
AVD	Arousal Valence Dominance
BNT	Bayesian Networks
BVP	Blood Volume Pulse
CNS	Central Nervous System
DAG	Database for Affective Gaming
DDA	Dynamic Difficulty Adaptation
DET	Determinism
DIA	Difficulty, Immersion, Amusement
DT	Decision Trees
ECG	Electrocardiography
EDA	Electrodermal Activity
EEG	Electroencephalography
EMG	Electromyography
ENTR	Entropy
EOG	Electrooculography
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FPS	First Person Shooter
GA	Generalization Ability
GEQ	Game Experience Questionnaire
GM	Group-based Model
GSR	Galvanic Skin Response
HA	High Arousal
HAHV	High Arousal High Valence
HALV	High Arousal Low Valence
HCI	Human Computer Interaction
HF	High Frequency
HR	Heart Rate

HRV	Heart Rate Variability
HV	High Valence
IBI	Interbeat Intervals
IRS	Individual Response Specificity
KNN	K Nearest Neighbour
LA	Low Valence
LAHV	Low Arousal High Valence
LALV	Low Arousal Low Valence
LBP	Local Binary Patterns
LDA	Linear Discriminant Analysis
LF	Low Frequency
LOOCV	Leave-one-out cross-validation
LOSOCV	Leave-one-subject-out cross-validation
LV	Low Valence
MC	Mean-centering
MCA	Multimedia Content Analysis
MEG	Magneto-EncephaloGram
MRI	Magnetic Resonance Imaging
MTL	Multi-Task Learning
NB	Naive Bayes Classifier
fNIR	functional Near-Infra-Red Spectroscopy
NN	Neural Network
NPC	Non-Player Character
PA	Predicting Ability
PAD	Pleasure Arousal Dominance
PCA	Principle Component Analysis
PNS	Peripheral Nervous System
PPG	Photoplethysmogram
PPS	Peripheral Physiological Signals
PSD	Power Spectral Density
PTSD	Post-Traumatic Stress Disorder
QRS	Refers to graphical deflections seen on a ECG
RESP	Respiration
RF	Random Forest
RFE	Recursive Feature Elimination
mRMR	maximum Relevance Minimum Redundancy
RMS	Root Mean Square
RQA	Recurrent Quantification Analysis
RR	Recurrent Rate
RT	Regression Tree

SBS	Sequential Forward Selection
SFS	Sequential Backward Selection
SKT	Skin Temperature
ST	Standardization
SVM	Support Vector Machines
TEMP	Temperature
TEO	Teager Energy Operator
TN	True Negative
TP	True Positive
TPR	True Positive Rate
VFDT	Very Fast Decision Tree

BIBLIOGRAPHY

Abadi, M. K., R. Subramanian, S. M. Kia, P. Avesani, I. Patras, and N. Sebe (2015). "DECAF: MEG-Based Multimodal Database for Decoding Affective Physiological Responses." In: *IEEE Tran. on Affective Computing* 6, pp. 209–222. DOI: 10.1109/TAFFC.2015.2392932 (cit. on pp. 5, 40–43, 47, 92).

Adamo, JM (1980). "Fuzzy decision trees." In: *Fuzzy sets and systems* 4.3, pp. 207–219 (cit. on p. 137).

Aigrain, Jonathan, Séverine Dubuisson, Marcin Detyniecki, and Mohamed Chetouani (2015). "Person-specific behavioural features for automatic stress detection." In: *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. Vol. 3. IEEE, pp. 1–6 (cit. on p. 24).

Amin, Zenab, R Todd Constable, and Turhan Canli (2004). "Attentional bias for valenced stimuli as a function of personality in the dot-probe task." In: *Journal of Research in Personality* 38.1, pp. 15–23 (cit. on p. 108).

Arnold, Magda B (1960). "Emotion and personality." In: (cit. on p. 17).

Ayaz, Hasan, Patricia A Shewokis, Scott Bunce, Maria Schultheis, and Banu Onaral (2009). "Assessment of cognitive neural correlates for a functional near infrared-based brain computer interface system." In: *International Conference on Foundations of Augmented Cognition*. Springer, pp. 699–708 (cit. on p. 33).

Bailenson, Jeremy N, Emmanuel D Pontikakis, Iris B Mauss, James J Gross, Maria E Jabon, Cendri AC Hutcherson, Clifford Nass, and Oliver John (2008). "Real-time classification of evoked emotions using facial feature tracking and physiological responses." In: *International journal of human-computer studies* 66.5, pp. 303–317 (cit. on p. 108).

Bateman, Chris, Rebecca Lowenhaupt, and Lennart E Nacke (2011). "Player typology in theory and practice." In: *Proceedings of DiGRA: Think Design Play* (cit. on p. 22).

Baumeister, Roy F, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D Vohs (2001). "Bad is stronger than good." In: *Review of general psychology* 5.4, p. 323 (cit. on p. 56).

Bonarini, Andrea, Fiammetta Costa, Maurizio Garbarino, Matteo Matteucci, Maximiliano Romero, and Simone Tognetti (2011). "Affective videogames:

the problem of wearability and comfort." In: *Human-Computer Interaction. Users and Applications*. Springer, pp. 649–658 (cit. on pp. 27, 32, 42).

Bontchev, Boyan (2016). "Adaptation in Affective Video Games: A Literature Review." In: *Cybernetics and Information Technologies* 16.3, pp. 3–34 (cit. on pp. 5, 31, 32).

Bos, Danny Oude (2006). "EEG-based emotion recognition." In: *The Influence of Visual and Auditory Stimuli*, pp. 1–17 (cit. on p. 27).

Bradley, Margaret M, Maurizio Codispoti, Dean Sabatinelli, and Peter J Lang (2001). "Emotion and motivation II: sex differences in picture processing." In: *Emotion* 1.3, p. 300 (cit. on p. 108).

Broersen, Piet MT (2000a). "Facts and fiction in spectral analysis." In: *IEEE Transactions on instrumentation and measurement* 49.4, pp. 766–772 (cit. on p. 68).

Broersen, Piet MT (2000b). "Finite sample criteria for autoregressive order selection." In: *IEEE Transactions on Signal Processing* 48.12, pp. 3550–3558 (cit. on p. 68).

Canli, Turhan, Zuo Zhao, John E Desmond, Eunjoo Kang, James Gross, and John DE Gabrieli (2001). "An fMRI study of personality influences on brain reactivity to emotional stimuli." In: *Behavioral neuroscience* 115.1, p. 33 (cit. on p. 108).

Canli, Turhan, Heidi Sivers, Susan L Whitfield, Ian H Gotlib, and John DE Gabrieli (2002). "Amygdala response to happy faces as a function of extraversion." In: *Science* 296.5576, pp. 2191–2191 (cit. on p. 108).

Chanel, Guillaume, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun (2008). "Boredom, engagement and anxiety as indicators for adaptation to difficulty in games." In: *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*. ACM, pp. 13–17 (cit. on pp. 22, 33, 36, 43, 44, 69).

Chanel, Guillaume, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun (2011). "Emotion assessment from physiological signals for adaptation of game difficulty." In: *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 41.6, pp. 1052–1063 (cit. on pp. 6, 31).

Christy, Thomas and Ludmila I Kuncheva (2013). "Amber shark-fin: An unobtrusive affective mouse." In: *ACHI 2013, The Sixth Intl. Conf. on Advances in Computer-Human Interactions*, pp. 488–495 (cit. on pp. 27, 32, 42).

Christy, Thomas and Ludmila I Kuncheva (2014). "Technological Advancements in Affective Gaming: A Historical Survey." In: *GSTF Journal on Computing* 3.4 (cit. on pp. 15, 25).

Coan, James A and John JB Allen (2007). *Handbook of emotion elicitation and assessment*. Oxford university press (cit. on pp. 27, 39).

Cowie, Roddy, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder (2000). "'FEELTRACE': An instrument for recording perceived emotion in real time." In: *ISCA tutorial and research workshop (ITRW) on speech and emotion* (cit. on p. 136).

Cowie, Roddy, Martin Sawey, Cian Doherty, Javier Jaimovich, Cavan Fyans, and Paul Stapleton (2013). "Gtrace: General trace program compatible with emotionml." In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, pp. 709–710 (cit. on p. 136).

Csikszentmihalyi, Mihaly (1985). "Das Flow-Erlebnis." In: *Jenseits von Angst und Langeweile: im Tun aufgehen. Stuttgart* (cit. on pp. 34, 35).

Csikszentmihalyi, Mihaly (1990). *Flow: The psychology of happiness*. Harper Perennial (cit. on p. 22).

Csikszentmihalyi, Mihaly (1996). "Flow and the psychology of discovery and invention." In: *New York: Harper Collins* (cit. on pp. 43, 47, 57).

Damez, Marc, Thanh Ha Dang, Christophe Marsala, and Bernadette Bouchon-Meunier (2005). "Fuzzy decision tree for user modeling from human-computer interactions." In: *Proceedings of the 5th International Conference on Human System Learning, ICHSL*. Vol. 5, pp. 287–302 (cit. on p. 137).

Dekker, Andrew and Erik Champion (2007). "Please Biofeed the Zombies: Enhancing the Gameplay and Display of a Horror Game Using Biofeedback." In: *DiGRA Conference* (cit. on pp. 3, 23, 33, 34, 43, 44).

Dellaert, Frank, Thomas Polzin, and Alex Waibel (1996). "Recognizing emotion in speech." In: *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*. Vol. 3. IEEE, pp. 1970–1973 (cit. on pp. 25, 41).

Ding, Chris and Hanchuan Peng (2005). "Minimum redundancy feature selection from microarray gene expression data." In: *Journal of bioinformatics and computational biology* 3.02, pp. 185–205 (cit. on p. 100).

Domingos, Pedro and Geoff Hulten (2000). "Mining high-speed data streams." In: *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 71–80 (cit. on p. 138).

Douglas-cowie, Ellen, Roddy Cowie, Ian Sneddon, Cate Cox, Orla Lowry, Margaret Mcrorie, Jean claude Martin, Laurence Devillers, Sarkis Abrilian, Anton Batliner, Noam Amir, and Kostas Karpouzis (2007). "The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and." In: *Induced Emotional Data," Affective Computing and Intelligent Interaction*, pp. 488–500 (cit. on p. 41).

Edwards, Jane, Henry J Jackson, and Philippa E Pattison (2002). "Emotion recognition via facial expression and affective prosody in schizophrenia:

a methodological review." In: *Clinical psychology review* 22.6, pp. 789–832 (cit. on p. 24).

Eid, Michael and Ed Diener (2001). "Norms for experiencing emotions in different cultures: inter-and intranational differences." In: *Journal of personality and social psychology* 81.5, p. 869 (cit. on p. 108).

Ekman, Paul (1993). "Facial expression and emotion." In: *American psychologist* 48, p. 384 (cit. on pp. 42, 47).

Ekman, Paul and Harriet Oster (1979). "Facial expressions of emotion." In: *Annual review of psychology* 30.1, pp. 527–554 (cit. on p. 108).

Ekman, Paul, Friesen Wallace V., Ellsworth Phoebe, Goldstein Arnold P., and Leonard Krasner (1982). *Emotion in the human face (2nd ed.)* Cambridge University Press (cit. on pp. 16, 17).

Ekman, Paul and Erika L Rosenberg (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA (cit. on p. 24).

Ekman, Paul and Daniel Cordaro (2011). "What is meant by calling emotions basic." In: *Emotion Review* 3.4, pp. 364–370 (cit. on pp. 16, 43).

Emmen, Dirrik HG and Georgios Lampropoulos (2014). "BioPong: Adaptive Gaming Using Biofeedback." In: *Creating the Difference*, p. 100 (cit. on p. 33).

Fairclough, Stephen and Kiel Gilleade (2012). "Construction of the biocybernetic loop: a case study." In: *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, pp. 571–578 (cit. on pp. 23, 33–36).

Fairclough, Stephen H (2008). "Fundamentals of physiological computing." In: *Interacting with computers* 21.1-2, pp. 133–145 (cit. on pp. 2–4, 27, 32).

Frijda, Nico H. (1986). *The emotions*. Cambridge University Press (cit. on p. 17).

Gilleade, Kiel, Alan Dix, and Jen Allanson (2005). "Affective videogames and modes of affective gaming: assist me, challenge me, emote me." In: *DiGRA 2005: Changing Views–Worlds in Play.* (Cit. on p. 2).

Glowinski, Donald, Nele Dael, Antonio Camurri, Gualtiero Volpe, Marcello Mortillaro, and Klaus Scherer (2011). "Toward a minimal representation of affective gestures." In: *IEEE Transactions on Affective Computing* 2.2, pp. 106–118 (cit. on pp. 24, 41).

Gow, Jeremy, Paul Cairns, Simon Colton, Paul Miller, and Robin Baumgarten (2010). "Capturing player experience with post-game commentaries." In: *Proceedings of the International Conference on Computer Games, Multimedia, and Allied Technologies (CGAT), Singapore*. Citeseer (cit. on p. 28).

Gray, Jeffrey A (1982). *The neuropsychology of anxiety*. Oxford University Press (cit. on p. 17).

Grimm, Michael, Kristian Kroschel, and Shrikanth S. Narayanan (2008). "The Vera am Mittag German audio-visual emotional speech database." In: *Proc. of the IEEE Intl. Conf. on Multimedia and Expo (ICME)*. Hannover, Germany, pp. 865–868 (cit. on p. 39).

Guyon, Isabelle, Jason Weston, Stephen Barnhill, and Vladimir Vapnik (2002). "Gene Selection for Cancer Classification using Support Vector Machines." In: *Machine Learning* 46.1, pp. 389–422. doi: 10.1023/A:1012487302797 (cit. on p. 99).

Guyon, Isabelle and André Elisseeff (2003). "An introduction to variable and feature selection." In: *Journal of machine learning research* 3.Mar, pp. 1157–1182 (cit. on p. 97).

Hamann, Stephan and Turhan Canli (2004). "Individual differences in emotion processing." In: *Current opinion in neurobiology* 14.2, pp. 233–238 (cit. on p. 108).

Harms, Madeline B, Alex Martin, and Gregory L Wallace (2010). "Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies." In: *Neuropsychology review* 20.3, pp. 290–322 (cit. on p. 24).

Healey, J. A. and R. W. Picard (2005). "Detecting stress during real-world driving tasks using physiological sensors." In: *IEEE Tran. on Intelligent Transportation Systems* 6, pp. 156–166. doi: 10.1109/TITS.2005.848368 (cit. on p. 41).

Healey, Jennifer Anne (2000). "Wearable and automotive systems for affect recognition from physiology." PhD thesis. Massachusetts Institute of Technology (cit. on pp. 40, 41, 43).

Holbrook, Morris B, Robert W Chestnut, Terence A Oliva, and Eric A Greenleaf (1984). "Play as a consumption experience: The roles of emotions, performance, and personality in the enjoyment of games." In: *Journal of consumer research* 11.2, pp. 728–739 (cit. on p. 1).

Holmgård, Christoffer, Georgios N Yannakakis, Hector P Martinez, and Karen-Inge Karstoft (2015). "To rank or to classify? Annotating stress for reliable PTSD profiling." In: *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, pp. 719–725 (cit. on pp. 42–44).

IJsselsteijn, Wijnand, Yvonne De Kort, Karolien Poels, Audrius Jurgelionis, and Francesco Bellotti (2007). "Characterising and measuring user experiences in digital games." In: *Intl. conf on advances in computer entertainment technology*. Vol. 2, p. 27 (cit. on pp. 47, 54).

Isbister, Katherine and Noah Schaffer (2008). *Game usability: Advancing the player experience*. CRC Press (cit. on pp. 28, 29).

Izard, Carroll E. (1971). *The face of emotion*. Appleton-Century-Crofts (cit. on p. 17).

Izard, Carroll E (2007). "Basic emotions, natural kinds, emotion schemas, and a new paradigm." In: *Perspectives on psychological science* 2.3, pp. 260–280 (cit. on pp. 15, 42).

James, William (1884). "What is an emotion?" In: *Mind* 34, pp. 188–205 (cit. on p. 17).

Kaplan, S, RS Dalal, and JN Luchman (2013). *Measurement of Emotions. Research Methods in Occupational Health Psychology* (cit. on p. 25).

Karpouzis, Kostas, Georgios N Yannakakis, Noor Shaker, and Stylianos Asteriadis (2015). "The platformer experience dataset." In: *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, pp. 712–718 (cit. on pp. 40, 42, 43).

Katsis, Christos D, Nikolaos Katertsidis, George Ganiatsas, and Dimitrios I Fotiadis (2008). "Toward emotion recognition in car-racing drivers: A biosignal processing approach." In: *IEEE Tran. on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38, pp. 502–512 (cit. on pp. 67–69).

Kim, Jonghwa (2007). "Bimodal emotion recognition using speech and physiological changes." In: *Robust speech recognition and understanding*. InTech (cit. on p. 76).

Kim, Jonghwa and Elisabeth André (2008). "Emotion recognition based on physiological changes in music listening." In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30.12, pp. 2067–2083 (cit. on pp. 67–69, 76).

Kim, K. H., S. W. Bang, and S. R. Kim (2004). "Emotion recognition system using short-term monitoring of physiological signals." In: *Medical and Biological Engineering and Computing* 42, pp. 419–427. DOI: 10.1007/BF02344719 (cit. on pp. 67–69).

Kivikangas, J Matias, Guillaume Chanel, Ben Cowley, Inger Ekman, Mikko Salminen, Simo Järvelä, and Niklas Ravaja (2011). "A review of the use of psychophysiological methods in game research." In: *Journal of Gaming & Virtual Worlds* 3.3, pp. 181–199 (cit. on pp. 26, 43).

Kleinsmith, Andrea and Nadia Bianchi-Berthouze (2013). "Affective body expression perception and recognition: A survey." In: *IEEE Transactions on Affective Computing* 4.1, pp. 15–33 (cit. on p. 24).

Koelstra, S., C. Muhl, M. Soleymani, J. S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras (2012). "DEAP: A Database for Emotion Analysis ;Using Physiological Signals." In: *IEEE Tran. on Affective Computing* 3, pp. 18–31. DOI: 10.1109/T-AFFC.2011.15 (cit. on pp. 5, 6, 40–43, 47, 56, 80, 92, 97).

Kotsia, Irene, Stefanos Zafeiriou, and Spiros Fotopoulos (2013). "Affective gaming: A comprehensive survey." In: *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*. IEEE, pp. 663–670 (cit. on pp. 29, 30).

Krech, David, Richard S Crutchfield, and Norman Livson (1974). *Elements of psychology.*. Alfred a. knopf (cit. on p. 20).

Kreibig, Sylvia D (2010). "Autonomic nervous system activity in emotion: A review." In: *Biological psychology* 84.3, pp. 394–421 (cit. on pp. 1, 2, 4, 6, 27, 35, 43, 93).

Krumhuber, Eva G, Arvid Kappas, and Antony SR Manstead (2013). "Effects of dynamic aspects of facial expressions: a review." In: *Emotion Review* 5.1, pp. 41–46 (cit. on pp. 24, 41).

Lang, Peter J (1995). "The emotion probe: Studies of motivation and attention." In: *American psychologist* 50.5, p. 372 (cit. on p. 74).

Lazzaro, Nicole (2004). "Why we play games: Four keys to more emotion without story." In: (cit. on p. 22).

Lewandowska, Magdalena, Jacek Rumiński, Tomasz Kocejko, and Jedrzej Nowak (2011). "Measuring pulse rate with a webcam—a non-contact method for evaluating cardiac activity." In: *Computer Science and Information Systems (FedCSIS), 2011 Federated Conf. on*. IEEE, pp. 405–410 (cit. on pp. 27, 42).

Li, Chao, Zhiyong Feng, and Chao Xu (2014). "Physiological-based emotion recognition with IRS model." In: *Smart Computing (SMARTCOMP), 2014 International Conference on*. IEEE, pp. 208–215 (cit. on p. 108).

Li, Ma, Quek Chai, Teo Kaixiang, Abdul Wahab, and Hüseyin Abut (2009). "EEG emotion recognition system." In: *In-vehicle corpus and signal processing for driver behavior*. Springer, pp. 125–135 (cit. on p. 27).

Liao, Lun-De, Chi-Yu Chen, I-Jan Wang, Sheng-Fu Chen, Shih-Yu Li, Bo-Wei Chen, Jyh-Yeong Chang, and Chin-Teng Lin (2012). "Gaming control using a wearable and wireless EEG-based brain-computer interface device with novel dry foam-based sensors." In: *Journal of neuroengineering and rehabilitation* 9.1, p. 5 (cit. on pp. 33, 34).

Lindley, Craig A and Charlotte C Sennersten (2006). "Game play schemas: from player analysis to adaptive game mechanics." In: *Proceedings of the 2006 international conference on Game research and development*. Murdoch University, pp. 47–53 (cit. on p. 2).

Liu, Changchun, Pramila Agrawal, Nilanjan Sarkar, and Shuo Chen (2009). "Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback." In: *International Journal of Human-Computer Interaction* 25.6, pp. 506–529 (cit. on pp. 6, 31, 33–36, 43).

Lopes, Phil, Georgios N Yannakakis, and Antonios Liapis (2017). "RankTrace: Relative and Unbounded Affect Annotation." In: *Proc. of ACII* (cit. on p. 136).

Mandryk, Regan Lee (2005). "Modeling user emotion in interactive play environments: A fuzzy physiological approach." PhD thesis. School of Computing Science-Simon Fraser University (cit. on pp. 23, 26, 28, 29).

Maragos, Petros, James F Kaiser, and Thomas F Quatieri (1993). "On amplitude and frequency demodulation using energy operators." In: *IEEE Transactions on signal processing* 41.4, pp. 1532–1550 (cit. on p. 68).

Martínez, Héctor P and Georgios N Yannakakis (2011). "Mining multimodal sequential patterns: a case study on affect detection." In: *Proceedings of the 13th international conference on multimodal interfaces*. ACM, pp. 3–10 (cit. on p. 43).

Massaro, Dominic W, Jonas Beskow, Michael M Cohen, Christopher L Fry, and Tony Rodgriguez (1999). "Picture my voice: Audio to visual speech synthesis using artificial neural networks." In: *AVSP'99-International Conference on Auditory-Visual Speech Processing* (cit. on p. 25).

Mauss, Iris B and Michael D Robinson (2009). "Measures of emotion: A review." In: *Cognition and emotion* 23.2, pp. 209–237 (cit. on pp. 25, 43).

McCurdy, Harold Grier (1950). "Consciousness and the galvanometer." In: *Psychological Review* 57.6, p. 322 (cit. on p. 74).

McDougall, William (1926). *An introduction to social psychology*. Courier Corporation (cit. on p. 17).

Mehrabian, Albert (1996). "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament." In: *Current Psychology* 14.4, pp. 261–292 (cit. on pp. 19, 21).

Mowrer, Orval (1960). "Learning theory and behavior." In: (cit. on p. 17).

Nacke, Lennart, Craig Lindley, and Sophie Stellmach (2008). "Log who's playing: psychophysiological game analysis made easy through event logging." In: *Fun and games*. Springer, pp. 150–157 (cit. on p. 26).

Nacke, Lennart Erik, Michael Kalyn, Calvin Lough, and Regan Lee Mandryk (2011). "Biofeedback game design: using direct and indirect physiological control to enhance game interaction." In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM, pp. 103–112 (cit. on p. 33).

Negini, Faham, Regan L Mandryk, and Kevin G Stanley (2014). "Using affective state to adapt characters, NPCs, and the environment in a first-person shooter game." In: *Games Media Entertainment (GEM), 2014 IEEE*. IEEE, pp. 1–8 (cit. on pp. 6, 31, 33).

Nicolaou, Mihalis A, Hatice Gunes, and Maja Pantic (2011). "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space." In: *Affective Computing, IEEE Transactions on* 2.2, pp. 92–105 (cit. on p. 21).

Nielsen, Jakob (1994). *Usability engineering*. Elsevier (cit. on pp. 28, 29).

Niven, Karen, Peter Totterdell, Christopher B. Stride, and David Holman (2011). "Emotion Regulation of Others and Self (EROS): The Development and Validation of a New Individual Difference Measure." In: *Current Psychology* 30.1, pp. 53–73. DOI: 10.1007/s12144-011-9099-9 (cit. on p. 108).

Nogueira, Pedro A, Rui Rodrigues, and Eugénio Oliveira (2013). "Real-time psychophysiological emotional state estimation in digital gameplay scenarios." In: *International Conference on Engineering Applications of Neural Networks*. Springer, pp. 243–252 (cit. on pp. 33, 34, 36, 43, 44).

Norris, Catherine J, Jackie Gollan, Gary G Berntson, and John T Cacioppo (2010). "The current status of research on the structure of evaluative space." In: *Biological psychology* 84.3, pp. 422–436 (cit. on pp. 43, 56).

Novak, Domen, Matjaž Mihelj, and Marko Munih (2012). "A survey of methods for data fusion and system adaptation using autonomic nervous system responses in physiological computing." In: *Interacting with computers* 24.3, pp. 154–172 (cit. on p. 6).

Oatley, Keith and Philip N Johnson-Laird (1987). "Towards a cognitive theory of emotions." In: *Cognition and emotion* 1.1, pp. 29–50 (cit. on p. 17).

Oliver, Nuria and Fernando Flores-Mangas (2006). "HealthGear: a real-time wearable system for monitoring and analyzing physiological signals." In: *Wearable and Implantable Body Sensor Networks, 2006. BSN 2006. Intl. Workshop on*. IEEE, 4–pp (cit. on pp. 27, 42).

Oppenheim, Abraham Naftali (1992). *Questionnaire design, interviewing and attitude measurement*. Bloomsbury Publishing (cit. on pp. 28, 29).

Ortony, Andrew (1990). *The cognitive structure of emotions*. Cambridge university press (cit. on p. 16).

Ortony, Andrew and Terence J Turner (1990). "What's basic about basic emotions?" In: *Psychological review* 97.3, p. 315 (cit. on pp. 16, 17, 43).

Pan, Jiapu and Willis J Tompkins (1985). "A real-time QRS detection algorithm." In: *IEEE transactions on biomedical engineering* 3, pp. 230–236 (cit. on p. 68).

Panksepp, Jaak (1982). "Toward a general psychobiological theory of emotions." In: *Behavioral and Brain sciences* 5.03, pp. 407–422 (cit. on p. 17).

Parnandi, Avinash and Ricardo Gutierrez-Osuna (2015). “A comparative study of game mechanics and control laws for an adaptive physiological game.” In: *Journal on Multimodal User Interfaces* 9.1, pp. 31–42 (cit. on pp. 2, 33, 34).

Petrushin, Valery A (2000). “Emotion recognition in speech signal: experimental study, development, and application.” In: *studies* 3.4 (cit. on pp. 25, 41).

Picard, Rosalind W (1995). “Affective Computing-MIT Media Laboratory Perceptual Computing Section Technical Report No. 321.” In: *Cambridge, MA* 2139 (cit. on pp. 3, 22).

Picard, Rosalind W (2003). “Affective computing: challenges.” In: *International Journal of Human-Computer Studies* 59.1, pp. 55–64 (cit. on pp. 4, 5).

Picard, Rosalind W., Elias Vyzas, and Jennifer Healey (2001). “Toward machine emotional intelligence: Analysis of affective physiological state.” In: *IEEE transactions on pattern analysis and machine intelligence* 23.10, pp. 1175–1191 (cit. on pp. 67–69).

Plutchik, Robert (1980). “A general psychoevolutionary theory of emotion.” In: *Theories of emotion* 1 (cit. on p. 17).

Plutchik, Robert Ed and Hope R Conte (1997). *Circumplex models of personality and emotions.* American Psychological Association (cit. on p. 19).

Pope, Alan T, Edward H Bogart, and Debbie S Bartolome (1995). “Biocybernetic system evaluates indices of operator engagement in automated task.” In: *Biological psychology* 40.1, pp. 187–195 (cit. on p. 2).

Rani, Pramila, Changchun Liu, Nilanjan Sarkar, and Eric Vanman (2006). “An empirical study of machine learning techniques for affect recognition in human–robot interaction.” In: *Pattern Analysis and Applications* 9.1, pp. 58–69 (cit. on p. 69).

Ravaja, Niklas, Timo Saari, Mikko Salminen, Jari Laarni, and Kari Kallinen (2006). “Phasic emotional reactions to video game events: A psychophysiological investigation.” In: *Media Psychology* 8.4, pp. 343–367 (cit. on pp. 43, 44, 93).

Ravaja, Niklas, Marko Turpeinen, Timo Saari, Sampsa Puttonen, and Liisa Keltikangas-Järvinen (2008). “The psychophysiology of James Bond: Phasic emotional responses to violent video game events.” In: *Emotion* 8.1, p. 114 (cit. on pp. 43, 93).

Reading, Mind (2004). “The Interactive Guide to Emotions.” In: *Lire l’Esprit: Guide interactif des émotions), Baron-Cohen* 142.179,253, p. 258 (cit. on p. 21).

Rigas, Georgios, Christos D Katsis, George Ganiatsas, and Dimitrios I Fotiadis (2007). “A user independent, biosignal based, emotion recognition method." In: *User Modeling*. Vol. 4511. Springer, pp. 314–318 (cit. on pp. 67–69).

Ringeval, F., A. Sonderegger, J. Sauer, and D. Lalanne (2013). "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions." In: *2013 10th IEEE Intl. Conf. and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8. DOI: 10.1109/FG.2013.6553805 (cit. on pp. 35, 40, 41, 43, 44, 47).

Ringeval, Fabien, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller (2015a). "Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data." In: *Pattern Recognition Letters* 66, pp. 22–30 (cit. on p. 43).

Ringeval, Fabien, Florian Eyben, Eleni Kroupi, Anil Yuce, Jean-Philippe Thiran, Touradj Ebrahimi, Denis Lalanne, and Björn Schuller (2015b). "Prediction of Asynchronous Dimensional Emotion Ratings from Audiovisual and Physiological Data." In: *Pattern Recogn. Lett.* 66, pp. 22–30 (cit. on p. 86).

Rugg, Michael D and Michael GH Coles (1995). *Electrophysiology of mind: Event-related brain potentials and cognition.* Oxford University Press (cit. on p. 22).

Russell, James A (1980a). "A circumplex model of affect." In: *Journal of personality and social psychology* 39.6, p. 1161 (cit. on p. 18).

Russell, James A (1980b). "A circumplex model of affect." In: *Journal of personality and social psychology* 39, p. 1161 (cit. on p. 43).

Russell, James A (1994). "Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies." In: *Psychological bulletin* 115.1, p. 102 (cit. on pp. 24, 25).

Scherer, Klaus R (2005). "What are emotions? And how can they be measured?" In: *Social science information* 44.4, pp. 695–729 (cit. on p. 42).

Schlimmer, Jeffrey C and Richard H Granger (1986). "Incremental learning from noisy data." In: *Machine learning* 1.3, pp. 317–354 (cit. on p. 138).

Schlosberg, Harold (1954). "Three dimensions of emotion." In: *Psychological review* 61.2, p. 81 (cit. on p. 19).

Schröder, Marc, E. Douglas-Cowie, and R. Cowie (2000). "A New Emotion Database: Considerations, Sources and Scope." In: *Proc. of the ISCA Workshop on Speech and Emotion: A Conceptual Framework for Research*. Belfast: Textflow, pp. 39–44 (cit. on p. 39).

Schwartz, Mark S and Frank Andrasik (2017). *Biofeedback: A practitioner's guide*. Guilford Publications (cit. on p. 22).

Sebe, Nicu, Ira Cohen, Thomas S Huang, et al. (2005). "Multimodal emotion recognition." In: *Handbook of Pattern Recognition and Computer Vision* 4, pp. 387–419 (cit. on p. 135).

Shalev-Shwartz, Shai et al. (2012). "Online learning and online convex optimization." In: *Foundations and Trends in Machine Learning* 4.2, pp. 107–194 (cit. on p. 138).

Sloan, Denise M (2004). "Emotion regulation in action: Emotional reactivity in experiential avoidance." In: *Behaviour Research and Therapy* 42.11, pp. 1257–1270 (cit. on p. 75).

Soleymani, M., J. Lichtenauer, T. Pun, and M. Pantic (2012). "A Multimodal Database for Affect Recognition and Implicit Tagging." In: *IEEE Tran. on Affective Computing* 3, pp. 42–55. doi: 10.1109/T-AFFC.2011.25 (cit. on pp. 6, 40–43, 47, 92).

Sweeney, Marian, Martin Maguire, and Brian Shackel (1993). "Evaluating user-computer interaction: a framework." In: *International journal of man-machine studies* 38.4, pp. 689–711 (cit. on p. 26).

Sykes, Jonathan and Simon Brown (2003a). "Affective gaming." In: *CHI '03 extended abstracts on Human factors in computing systems - CHI '03*, p. 732. doi: 10.1145/765891.765957 (cit. on pp. 27, 32, 42).

Sykes, Jonathan and Simon Brown (2003b). "Affective gaming: measuring emotion through the gamepad." In: *CHI'03 extended abstracts on Human factors in computing systems*. ACM, pp. 732–733 (cit. on p. 26).

Tan, K Song, Reza Saatchi, Heather Elphick, and Derek Burke (2010). "Real-time vision based respiration monitoring system." In: *Communication Systems Networks and Digital Signal Processing (CSNDSP), 2010 7th Intl. Symposium on*. IEEE, pp. 770–774 (cit. on pp. 27, 42).

Tarasenko, Sergey (2010). "Emotionally colorful reflexive games." In: *arXiv preprint arXiv:1101.0820* (cit. on p. 20).

Thorndike, Robert L (1953). "Who belongs in the family?" In: *Psychometrika* 18.4, pp. 267–276 (cit. on p. 121).

Tijs, Tim, Dirk Brokken, and Wijnand IJsselsteijn (2008a). "Creating an emotionally adaptive game." In: *International Conference on Entertainment Computing*. Springer, pp. 122–133 (cit. on pp. 3, 33–35, 43, 44).

Tijs, Tim JW, Dirk Brokken, and Wijnand A IJsselsteijn (2008b). "Dynamic game balancing by recognizing affect." In: *Fun and Games*. Springer, pp. 88–93 (cit. on pp. 6, 31, 33).

Tognetti, Simone, Maurizio Garbarino, Andrea Bonarini, and Matteo Matteucci (2010). "Modeling enjoyment preference from physiological responses in a car racing game." In: *Computational Intelligence and Games*

(CIG), 2010 IEEE Symposium on. IEEE, pp. 321–328 (cit. on pp. 33, 34, 36, 43, 44, 69).

Tomkins, Silvan (1962). *Affect imagery consciousness: Volume I: The positive affects*. Springer publishing company (cit. on p. 16).

Tomkins, Silvan (1963). *Affect Imagery Consciousness: Volume II: The Negative Affects*. Springer Publishing Company (cit. on p. 16).

Tomkins, Silvan S (1984). "Affect theory." In: *Approaches to emotion* 163, p. 195 (cit. on p. 17).

Toups, Zachary O, Ross Graeber, Andruid Kerne, Louis Tassinary, Sarah Berry, Kyle Overby, and Maritza Johnson (2006). "A design for using physiological signals to affect team game play." In: *Foundations of Augmented Cognition*, pp. 134–139 (cit. on pp. 33, 36, 43, 44).

Tracy, Jessica L and Daniel Randles (2011). "Four models of basic emotions: a review of Ekman and Cordaro, Izard, Levenson, and Panksepp and Watt." In: *Emotion Review* 3.4, pp. 397–405 (cit. on p. 43).

Watson, David, Lee A Clark, and Auke Tellegen (1988). "Development and validation of brief measures of positive and negative affect: the PANAS scales." In: *Journal of personality and social psychology* 54.6, p. 1063 (cit. on pp. 18, 23, 43).

Watson, John Broadus et al. (1925). *Behaviorism*. Transaction Publishers (cit. on p. 17).

Weiner, Bernard and Sarah Graham (1984). "An attributional approach to emotional development." In: *Emotions, cognition, and behavior*, pp. 167–191 (cit. on p. 17).

Yannakakis, Georgios N and John Hallam (2008). "Entertainment modeling through physiology in physical play." In: *International Journal of Human-Computer Studies* 66.10, pp. 741–755 (cit. on p. 2).

Yannakakis, Georgios N, Héctor P Martínez, and Arnav Jhala (2010). "Towards affective camera control in games." In: *User Modeling and User-Adapted Interaction* 20.4, pp. 313–340 (cit. on pp. 40, 42, 43).

Yannakakis, Georgios N and John Hallam (2011). "Ranking vs. preference: a comparative study of self-reporting." In: *International Conference on Affective Computing and Intelligent Interaction*. Springer, pp. 437–446 (cit. on p. 47).

Yannakakis, Georgios N and Ana Paiva (2014). "Emotion in games." In: *Handbook on affective computing*, pp. 459–471 (cit. on p. 43).

Yannakakis, Georgios N, Hector P Martinez, and Maurizio Garbarino (2016). "Psychophysiology in games." In: *Emotion in Games*. Springer, pp. 119–137 (cit. on p. 41).

Ye, Juan, Simon Dobson, and Susan McKeever (2012). “Situation identification techniques in pervasive computing: A review.” In: *Pervasive and mobile computing* 8.1, pp. 36–66 (cit. on p. 92).

Zavaschi, Thiago HH, Alceu S Britto, Luiz ES Oliveira, and Alessandro L Koerich (2013). “Fusion of feature sets and classifiers for facial expression recognition.” In: *Expert Systems with Applications* 40.2, pp. 646–655 (cit. on p. 24).

Zeng, Zhihong, Maja Pantic, Glenn I Roisman, and Thomas S Huang (2009). “A survey of affect recognition methods: Audio, visual, and spontaneous expressions.” In: *IEEE transactions on pattern analysis and machine intelligence* 31.1, pp. 39–58 (cit. on p. 41).

Zhalehpour, Sara, Zahid Akhtar, and Cigdem Eroglu Erdem (2016). “Multimodal emotion recognition based on peak frame selection from video.” In: *Signal, Image and Video Processing* 10.5, pp. 827–834 (cit. on p. 135).

Zong, Cong and Mohamed Chetouani (2009). “Hilbert-Huang transform based physiological signals analysis for emotion recognition.” In: *Signal processing and information technology (isspit), 2009 ieee international symposium on*. IEEE, pp. 334–339 (cit. on p. 69).