An empirical study of players' emotions in VR racing games based on a dataset of physiological data

Marco Granato[1] · Davide Gadia[1] · Dario Maggiorini[1] · Laura A. Ripamonti[1]

Received: 4 April 2019 / Revised: 29 October 2019 / Accepted: 11 December 2019 /
Published online: 02 March 2020

Abstract

A video game is an interactive software able to arouse intense emotions in players. Consequentially, different theories have been proposed to understand which game aspects are able to affect the players' emotional state. However, only few works have tried to use empirical evidence to investigate the effects of different game aspects of the players' emotions. In this paper, we present the results of a set of experiments aimed at predicting the players' emotions during video games sessions using their physiological data. We have created a physiological dataset from the data acquired by 33 participants during video game fruition using a standard monitor and a Virtual Reality headset. The dataset contains information about electrocardiogram, 5 facials electromyographies, electrodermal activity, and respiration. Furthermore, we have asked the players to self-assess their emotional state on the Arousal and Valence space. We have then analyzed the contribution of each physiological signal to the overall definition of the players' mental state. Finally, we have applied Machine Learning techniques to predict the emotional state of players during game sessions at a precision of one second. The obtained results can contribute to define game devices and engines able to detect physiological data, as well to improve the game design process.

Keywords Affective computing · Video games · Emotions recognition · ECG · EMG · GSR · EDA · Respiration · Physiological dataset · Valence · Arousal · Machine learning · Virtual reality · Players' emotions

✉ Marco Granato
marco.granato@unimi.it

Davide Gadia
gadia@di.unimi.it

Dario Maggiorini
dario@di.unimi.it

Laura A. Ripamonti
ripamonti@di.unimi.it

[1] Department of Computer Science, University of Milan, Via Celoria 18, 20133 Milan, Italy

1 Introduction

Huizinga [51] defines the action to play a game as "a voluntary activity or occupation executed within certain fixed limits of time and place, according to rules freely accepted but absolutely binding, having its aim in itself and accompanied by a feeling of tension, joy and the consciousness that it is 'different' from 'ordinary life". Thus, generally, the development of a game is focused on the players, with the main goal of inducing entertainment and engagement. A video game uses audio/visual information presented through electronic devices in order to communicate the game structure. It considers all the peculiar features of a game, including the ability to solicit players' emotions [42, 43, 52]. *Spatial Presence* [24] and *Flow* [18] are two theories of positive psychology that are commonly used to identify how a video game, or a generic entertainment product (e.g., a movie), interacts with the human emotions. *Spatial Presence* is a psychological condition describing how much a player has the illusion to be transported in a virtual environment. This condition can be better elicited by immersive technologies (e.g., Virtual Reality). Researchers have suggested that a high sense of *Spatial Presence* can improve players' entertainment and it may also facilitate the players' performance [68]. The theory of *Flow* tries, instead, to define a mental state where a user is completely absorbed in a task. This theory describes a balanced channel between challenge and ability: when in the *Flow* state, a person can benefit of an experience of achievement and happiness. In a video game session, the sensation of *Flow* seems to be connected with an increase of dopamine level, a neurotransmitter that increases human attention [56].

In this paper, we present a set of experiments, where a number of participants have played at 2 racing games in two different contexts, with and without Virtual Reality (VR). During the experiments, we have recorded physiological signals of the players: electromyography (EMG) on the players' face, in order to include the facial expressions in the analysis, electrocardiogram (ECG), galvanic skin response (GSR), and respiration intensity/rate (RESP). We have also collected, after game session, the players' self-assessment information about their emotional state during the game fruition. The self-assessment data and the physiological dataset have been used in order to infer the emotional state of the players in each moment of the game sessions, and to build a prediction model of players' emotions using Machine Learning (ML) techniques. Therefore, the three main contributions of this paper are: to create an affective dataset using video games as stimuli, to understand which physiological conditions are the most relevant in order to determine the players' emotions, and to propose a method for the real-time prediction of a player's mental state during a video game session.

The remainder of this paper is organized as follows: in Section 2, we present a brief overview of related works; in Section 3, we describe the affective database and the methodology used to acquire the data. Then, in Section 4, we explain the experimental setup, providing information about the study sample, the procedure, and the algorithm used for emotions assessment. The paper proceeds with the signals analysis in Section 5, focusing on filtering, features creation and selection, and with the application of ML techniques. In Section 6, we present the results and we also briefly discuss their connection with the *Spatial Presence* and *Flow* theories. Lastly, we provide conclusions and final considerations for future works in Section 8.

2 Background and related works

In the game design field, some studies have been proposed to understand which game components could be used to raise the players' engagement. Koster has stated [56] "the destiny of games is to become boring, not to be fun", because the enjoyment in video games persists until the player has the feeling to have learned something new that helps her to master the game mechanics. Lazzaro [58], through the observation of subjects playing their favorite video games, has identified 4 elements able to arouse emotions without an explicit narrative: *Hard Fun*, *Easy Fun*, *Altered States*, and *The People Factor*. The author chose to not consider the explicit narrative, since it may modify the emotions aroused in player in a similar way of a movie or a book. Instead, her main goal was to investigate the relationship between emotions and game mechanics. Freeman [34] has listed 33 game design techniques useful to elicit players' emotions during the games' fruition.

Since video games are media able to arouse emotions, several studies have been proposed to understand the sensations elicited by these entertainment products. These researches are based mainly on two approaches: the definition of a set of parameters which describes the players' behavior, and the direct analysis of users during a game session. Modeling the players' characteristics allows to understand their behavior and, consequentially, to suppose the emotions elicited by a specific video game feature through simulations. In [21, 71] the researchers have studied the players' satisfaction regarding the game rewards (Looting System) in Massive Multiplayer Online (MMO) and Multiplayer Online Battle Arena (MOBA) games, through a set of simulations where different agents have peculiar features (like, e.g., Bartle type [3], time usually spent playing video games, etc.). The analysis of a player during a game session can be performed in two ways: studying the player's behavior in a video game, or using her physiological information to infer her emotions. The former is usually considered in several moments during the development of the game, and sometimes also after the game release to the market. This procedure is called *playtest*, and it has the main purpose to have an overview of "the entire design process to gain an insight into whether the game is achieving your player experience goals" [35]. Playtesting is widely used in the industry, but, usually, it is not useful to extract quantitative data about the players. Instead, the second approach provides to the researchers quantitative results based on players' physiological conditions. The emotions have a physiological reaction in humans. Dalgleish [19] has presented an overview of theories on the relation between emotions and physiological conditions. Thus, physiologic reactions to emotions can be described and acquired by at least three human output systems [61]: self-report measure (e.g., through verbal expressions), behaviors (e.g., facial expressions), and physiological reactions of Autonomic Nervous System (*ANS*), like, e.g., Heart Rate, Brain Activity, etc. The physiological information can be used to identify the user's emotions [32], or to provide an input to a software or a device (e.g., [8, 31]). This is one well-known application of the Affective Computing field. However, to our knowledge, this research field, applied to video games, is not deeply investigated.

Two studies [65, 86] have used physiological data for the interaction with video games: in particular, they have considered the respiration signal to infer players' emotions and to adapt accordingly the video game status. A different signal ascribed to the players' emotions is the electrocardiogram. For the sake of brevity, we can define the electrocardiogram as a technique that provides information about electrical response conforming with the heart contractions. This kind of information has been adopted to provide an input in a video game [30] and to expand the game log during the playtest [12]. Hazlett [47] has applied, during a racing game session, two EMG sensors on the face of 13 teenage players. Through this study, the author has shown that the Zygomaticus muscle is more involved during positive events, while the Corrugator muscle is involved in the negative ones. Also, Tognetti et al. [80] have considered a racing game in order to understand the players' emotion under different game events. They have recorded 5 physiological signals - Blood Volume Pressure

(BVP), ECG, GSR, Respiration, and Temperature - and they have used them to classify the information collected using a survey. Unfortunately, none of the mentioned works have made publicly available the collected database.

In [39], the authors of this paper have proposed a preliminary framework to determine the players' emotions through physiological data. In this work, we have extended the results obtained in [39], carrying out a more rigorous experimental setup and proposing an enhanced hardware architecture to collect the physiological signals. Moreover, we have studied the emotions in players during game sessions in VR.

3 Methodology and tools

In the following sub-sections, we present the methodology and tools used to collect the physiological RAcing GAme dataset (RAGA). Then, we discuss the different typologies of emotion identification, and the methodologies used for the selection of the stimuli and physiological signals that we have considered as the most informative to infer emotions from video games players.

3.1 Available affective datasets

The reason for the creation of a novel physiological dataset is given mainly to the lack of affective data regarding video game players. In fact, the most common available datasets [1, 6, 54, 67, 70, 72, 78] are mainly based on different kind of stimuli, like, e.g., video, images, etc. They provide info on the physiological signals acquired from the subjects, the method of the emotions annotation (self-assessed or using external annotators), and the representation of emotions in a 2D or 3D vector space. At the best of our knowledge, RAGA is the first freely available dataset, developed for academic research, based on the use of video games as stimuli: in particular, we have considered game sessions based on a standard monitor and a VR headset.

3.2 Physiological data acquisition for RAGA

For each player participating to the experimental setup (described in Section 4), we have acquired 4 groups of physiological signals: electrocardiography (ECG), electromyography (EMG) on 5 facial muscles, Galvanic Skin Response (GSR), and Respiration Rate. We have acquired signals produced by the EMGs sensors, rather than the implementation of others camera-based face expression recognition techniques (e.g., [44]), in order to permit the game fruition also with emergent video game technologies (i.e., VR) that cover a relevant area of the players' face. These analogical data are collected using an Arduino Due that uses its built-in ADC to perform a 12-bit quantization at 556 Hz. The data is sent through Bluetooth protocol (using the HC-06 peripheral) to a computer. We have used a wireless protocol in order to remove the noise produced by the power grid, and we have provided power to the Arduino and the sensors through a stand-alone 5 V battery.

Thus, for ECG and EMG signals acquisition, we have involved six Olimex-EKG-EMG shields, a device compatible with Arduino, already used in biomedical engineering [79]. It is able to read a 3-lead electrode connector via 3.5 jack. One Olimex is used to acquire the ECG data connecting three disposable electrodes Fiab F9079/100 (36x40 mm) on clean skin, following the Einthoven triangle guidelines [16]: two at both wrists and one at the left ankle. The other 5 Olimex sensor are used to collect the data on 5 different areas on the right side of participants' face using disposable electrodes of size 32x32 mm, (Fiab F9053N) as illustrated in Fig. 1: on *Zygomaticus Major* (EMG1), *Corrugator Supercilii* (EMG2), *Nasalis* (EMG3), *Orbicularis Oculi* (EMG4), and *Temporalis* (EMG5) muscles. These pairs of electrodes used to acquire the EMG signals have a common reference electrode, placed on the forehead, near the hair border, as suggested in [81].

The GSR signal, also named Electrodermal Activity (EDA), is acquired by placing two electrodes (Fiab F9053N) on two distal phalanxes of the left hand. Usually, the players control racing games with the left area of a gamepad, using the left-hand thumb to steer through an analog joystick and the left index to brake using a trigger. Thus, we have connected the electrodes on two fingers, middle and ring, which usually are not used to control racing games. The potential difference is amplified using an LM324 Surface Mount Device (SMD) installed on a Grove GSR sensor. GSR can be considered as a reflection of the sympathetic axis that produces an eccrine sweat gland [17]. In humans, research has shown that the sweating has also a function of emotional expression [20]. Thus, the sympathetic activity can be considered linked to emotions and, therefore, GSR is often suggested as emotions index [82].

Finally, we have acquired the respiration intensity/rate signal placing an NTC Thermistor (NTCLE203E3 SB0) under the participant's nose. The base of the sensor has been isolated using insulating tape and it has been placed avoiding contact with the user's skin in order to limit the noise involved with the epidermal temperature. Thus, when the user exhales, she increases the temperature under the nose area and, as a consequence, reduces the tension information acquired by the sensor (vice versa when she inhales). The thermistor has an accuracy of ± 0.5 C$*\circ$ in a range between 25 C$*\circ$ and 85 C$*\circ$, as declared by the manufacturer [83].

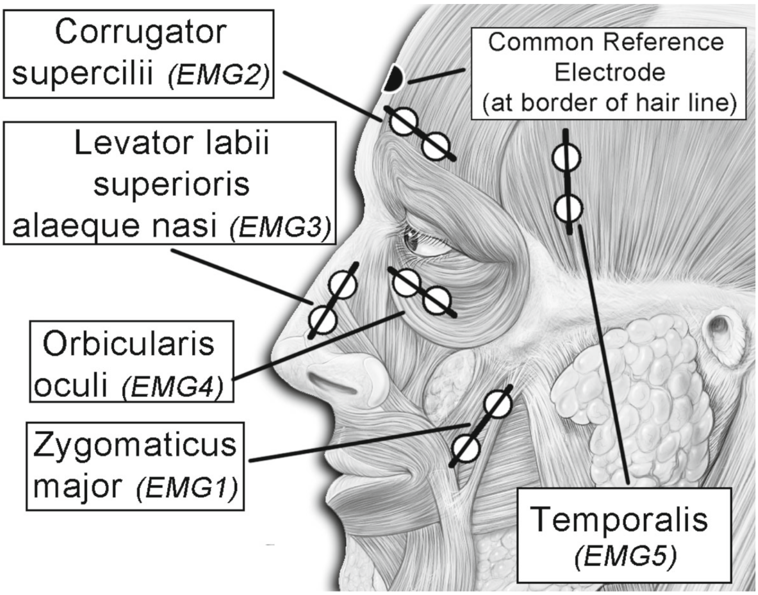

Fig. 1 Position of facial electrodes used to acquire the EMG signals. Original medical illustration from Patrick J. Lynch (https://goo.gl/ttgxo6)

All these sensors have received a tension of 3.3V directly by the Arduino Due, thus the ADC step is equal to 805.66 $\mu V/bit$. An architecture overview is presented in Fig. 2.

3.3 Emotion definition and tagging in RAGA

During the experiments, we need also to acquire information about players emotions. After the game session, we directly ask the participants to self-assess their emotional states during the video game fruition. Thus, an important decision to take is the type of emotion markers that must be applied.

We have considered acquiring the emotion self-assessment values in a continuous time over all the game levels. This has created a novel challenge in the video game experimental design, since the majority of the researches in affective video game field have identified the emotions in the discrete time, labeling the game sessions (or highlights) through a final survey (e.g., [80]). We agree that using a discrete time can provide a greater amount of data able to predict the emotional state, allowing an overall better accuracy on the ML model. However, to self-assess the overall game session with a single evaluation can be reductive, since a game stage can provide different and conflicting emotions. In contrast, if the research provides different evaluations using the game highlights, they may lose the focus on the overall emotions elicited by all the game session. In addition, defining the highlights as the parts of the game able to arouse emotions may be reductive for mainly two reasons: the game highlights may not affect the mental state of the player, and

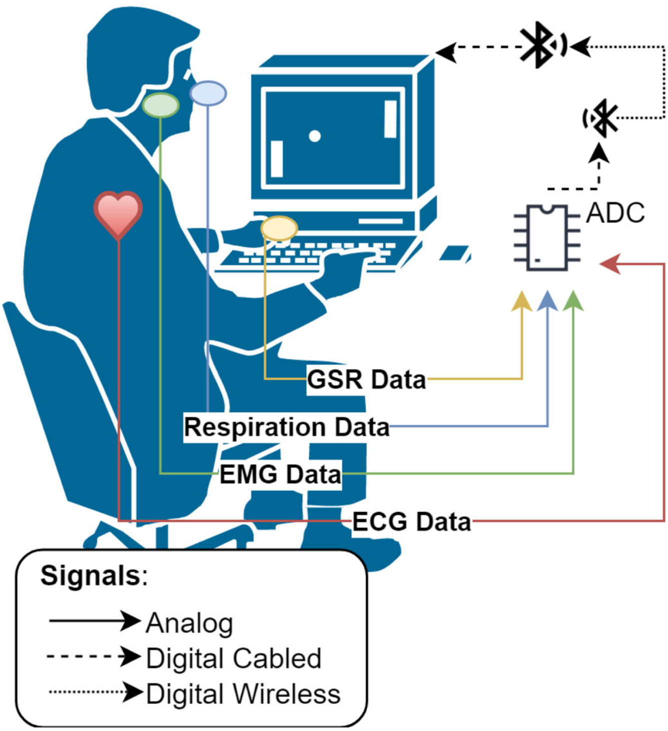

Fig. 2 Overview of the hardware architecture used to acquire physiological information for RAGA

not all the games have emotional highlights. For example, some repeated mechanics in a game may affect the game ability to elicit emotions in players, and as a consequence, may not provide the desired emotion. Moreover, to evaluate only the salient parts of the game (excluding the context) can induce the player to assess the emotion following what he thinks it is "expected", to the *emotional bias* [2] and *response bias* [36], or to confuse the events if they are similar. Considering, for example, a basketball video game, where the player makes and receives several points. This example describes the importance to examine the overall match, since if only the most salient moments of the match are shown to the player, she looses the context. Consequentially, an annotator may confuse the affective state elicited by a particular action, due to the similarity of the different events. Thus, to consider only the highlights may make the whole vision of the game lost. On the other side, some emotional games may not provide game highlights, but the implicit narrative and the environment are able to elicit the players' emotions. For example, *Journey* (developed by *Thatgamecompany*) does not provide any particular game highlight, however it is able to provide strong emotions to players [76]. A novel approach for the emotion classification may be to evaluate the physiological changes rather than game highlights. Of course, it does not resolve the issues based on biases and on the loss of the overall vision of the game stage, however it may assess the relevant information with the player point of view.

The most common way to describe an emotion is to identify it among a discrete set of labels (e.g., joy, boredom, etc.). Scientific researches have identified several markers, anyway, they are often not uniformly defined across the different cultural background. Ekman [27] has defined six basic and universal human emotions: *anger*, *fear*, *sadness*, *happiness*, *disgust*, and *surprise*. These emotions have the same intrinsic meaning and physiological outcome across the different cultures. Despite the interesting results, Russell [73] has underlined the gaps of Ekman's research by showing that the names of different emotions have an overlapped meaning in some languages. Therefore, psychologists have developed an alternative methodology to identify the emotions, by distributing them on an n-dimensional space. The most common approach is to map the emotions into a 3D vector space, considering *Pleasure*(*Valence*)-*Arousal*-*Dominance* (PAD), as axes [74]. Following the approach of other datasets [6, 70, 72], in the present work we have considered, for the emotions identification, only the *Valence* and *Arousal* vectors (VA). The former defines the emotion "quality" (from *averseness* to *attractiveness*) and the latter defines the emotion "intensity" (from *very calm* to *very excited*). This helps to reduce the time of the experimental sessions, and to minimize the participants' bias.

A common practice to allow the players to self-assess their emotions is to use the Self-Assesment Manikins (SAM) [9], which are structured by a series of anthropomorphic figures representing different human emotions. They are able to map the 2-dimensional space defined by the *Valence* and *Arousal* vectors. Betella and Verschure [5] have developed the Affective Slider (AS) as an alternative to the SAM. AS is structured by a set of emoticons that represent the emotions limits of PAD vectors. It also uses a bow-tie graph where the narrow area indicates a neutral emotion. Through their research, the authors have shown that AS can be a valid alternative to SAM. In our work, we have used a combination of SAM and AS in order to maximize the user precision in the labeling phase (see Fig. 3). By moving the AS slider, a square indicator moves over the corresponding SAM, as a reinforcement tool for self-assessment. A detailed explanation on the hardware architecture and the software developed for the experimental setup can be found in [41].

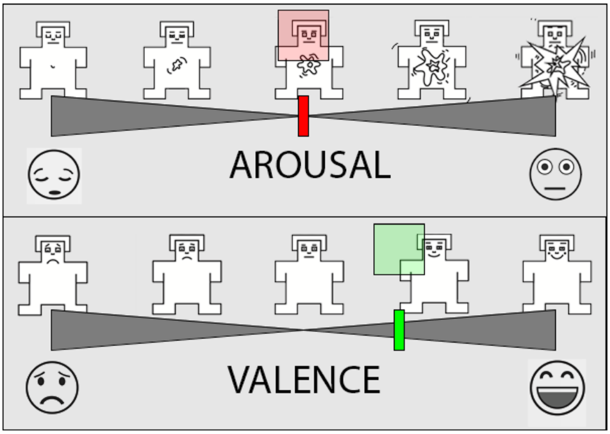

Fig. 3 SAM and AS self assessment tools used for RAGA dataset

3.4 Game genre used for RAGA stimuli

One of the first decision to take, when designing the creation of a physiological dataset based on video games, is which kind of game is the more adequate for the final purpose. To focus only on the influence of the game mechanics, and to avoid the emotional effect of narrative elements (like e.g., cinematic sequences), we have decided to consider racing video games. In racing video games, the users start from a point A, and they must arrive at a point B, usually driving vehicles, in the shortest possible time or overcoming the opponents. Racing games usually have a set of possible events that can arouse players' emotions. In fact, players are involved in high-speed races where usually: accidents, overtaking, high-speed corners, etc. can occur. Therefore, we have hypothesized a high variability of players emotions during a racing game and, as a consequence, of the corresponding VA values. This genre can be further divided in three sub-categories: arcade racing games, where the priority is given to fun and feel of speed; simulation games, designed to guarantee a user experience similar to the reality (goal of the game engine is to simulate a truthful vehicles physics); and kart games, based on a simplified driving mechanic characterized by features that usually do not appear in other racing sub-genres (e.g., obstacles, weapons, possibility to jump, power-up, etc). We have decided to consider 2 racing games from different sub-genres. We have defined a set of constraints for the games' selection: they should have an intuitive and straightforward game mechanic and environment; they must have comparable input controls and level length; they can be played either in VR or using a standard monitor. Thus, we have selected a simulation driving game, Project Cars[1] (*PCars* from now on), and an arcade driving game, RedOut[2] (*RO*). PCars, a game released to the market in 2015, is developed by *Slightly Mad Studios* and published by *Bandai Namco*. It is a simulation game, where the drivers are involved in races on virtual reproductions of existing cars that compete on famous circuits. RO is a futuristic racing game developed by an Italian company (*34BigThings*) and released in September 2016. It is an independent game where futuristic shuttles compete in full acrobatics tracks. The main inputs for the above-mentioned games

[1] https://www.projectcarsgame.com

[2] https://34bigthings.com/portfolio/redout/

are quite simple and symmetrical: a steering input, an input to accelerate, and an input to brake. RO, due to its arcade nature, has two other inputs: the first dedicated to control the shuttle inclination, and the second used to activate the turbo speed. Both games can be played from the driver point of view, with or without VR headset.

4 Experimental setup

The group of participants was composed of 33 players (29 males and 4 females), between 18 and 40 (μ = 24.66 and σ = 5.15) years old. They usually play video games in average 3.6 days per week ($\sigma = 2.16$), and the games session of more than half of participants are longer than 2 hours. All participants were Italian, and they have not received any monetary contribution for the experiment.

4.1 Procedure

Upon arrival in the laboratory, the participants have been invited to sit on a comfortable chair. They have been informed about the experimental procedure and they have been invited to read and sign an informed consent, and a permission to use the data and video recorded during the experiment for research and academic purposes. The acquired data have been collected anonymously, and we have assigned, at each participant, an incremental identifier for the analysis and for future experiments. The participants with ID lower than 10 are the same users who participated at our previous experiment [39]. Lastly, each user has performed the test in a single daily session. During the experiment, a 5 MP camera has been placed under the 32" display used during the gaming sessions. The games have been played using a gamepad on a computer with Windows 10 OS, i7 6700k CPU, 32gb RAM (DDR4), and NVIDIA GeForce GTX 1080 GPU. For the experiment, we have also used the headset Oculus Rift DK2 for the VR sessions.

The experiment has consisted of three stages: electrodes placement and test presentation, main test, and final survey. We have connected the electrodes and the thermometer on the participant's skin, taking care to not bother her view or attention. After the electrodes' placement, we have powered the Arduino Due and, as a consequence, all the sensors. After checking that all the sensors were working, we have checked the signal and communication quality asking the participants to perform facial movements, in order to check the EMGs signals, and to have a deep breath, which usually causes a sweating alteration, to check the GSR signal.

A video with the demo of the games, the tracks, and the vehicles have been also shown to the participants. For all the gaming session, we have selected a *McLaren, 12C* in the *California Highway Stage 2* track for PCars, while we have used *Asera, Yoshinobu* shuttle on *Alaska, Airbone* on RO. Furthermore, a member of the laboratory staff has explained the game mechanics for both games and he made sure that the participant has understood how to interact with the games.

The software for the emotion annotation has been presented to each participant, and a short training to familiarize with the input system have been conducted. The assessment software shows to the participant a video with the information of the player's face, a video of the gameplay, and a data synchronization graph. We have asked the participants to identify their emotions during the entire video playback using the SAM/AS tool described in Section 3.3. In Fig. 4, we show the six different areas of the application GUI:

(a) the bottom left bar indicates the Valence values: the participant has to move the green rod using the gamepad left analog joystick to self-assess her own emotional state
(b) the bottom right bar is used to collect the Arousal values. In this case, the player uses the right analog joystick to move the rod
(c) In the bottom right area, the player's face during the game session is shown: we have asked the participant to focus on this area in order to re-evoke the emotions experienced during the game session
(d) In the top right area, the gameplay useful to support the emotions recall is shown
(e) In the top left area, the acquired physiological information is shown. It is used only to synchronize, through its top and bottom bars, the collected data with the Valence and Arousal information acquired by the software
(f) Lastly, the red bar indicates the time position within the video.

The *main test* has been structured in two stages, each repeated two times (i.e., with and without VR): game session and emotion tagging. Randomly, the participants have started the game session with or without VR: 14 participants have started with VR, while 19 have performed the first game session without VR. Each player, whether she has used VR or not, played to PCars as the first game. The beginning and end of each race have been synchronized to physiological and assessment data by pressing two different buttons, connected to the Arduino Due. The synchronization starts/ends at the beginning of the Arduino clock cycle. The first button pressure inserts in a specific column of the physiological dataset the value 14 (a number used only to identify the session beginning), and, at the same time, it

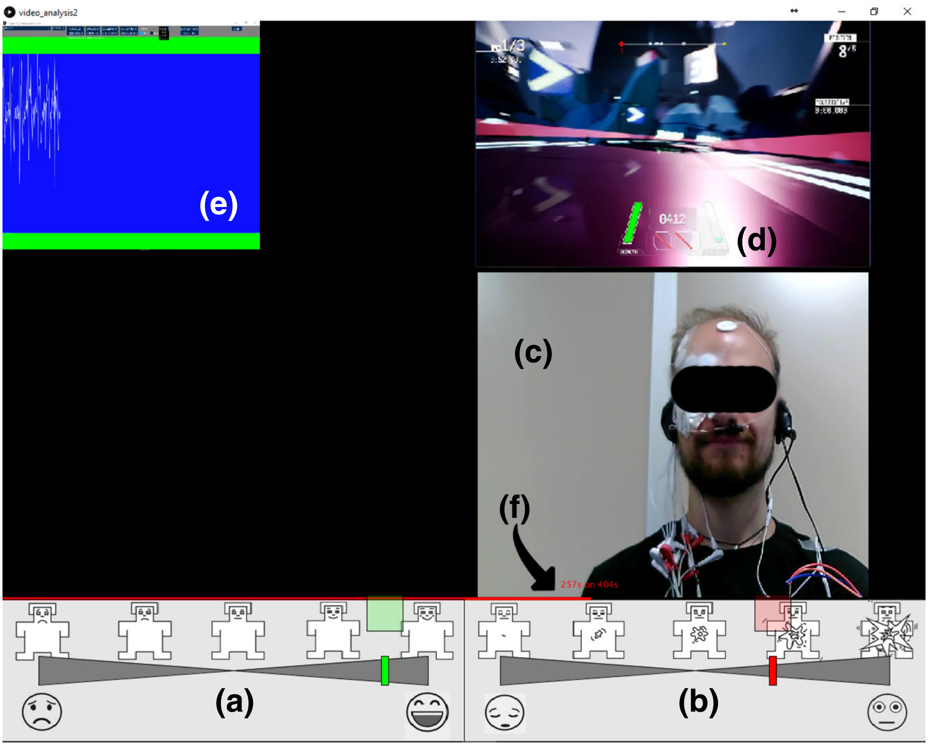

Fig. 4 GUI of the emotion tagging software used after game session

switches the color of the two bars (e) to green. In the same way, the second button pressure is used to insert the value 15 and to switch the color of the bars to red. Thus, the two buttons are used to synchronize the physiological data with the first emotion tagging performed immediately after the game session. The second game session has been performed right after the first, followed by another emotion tagging stage. Before the VR stage, the laboratory staff has explained the potential risks related to the device. The possible motion sickness deriving from the VR device may arise due to a not accurate settings of the VR parameters [38, 84] (like e.g., the inter-ocular distance considered in the virtual scene, or the field of view of the virtual camera), or to physiological issues of the players with stereoscopic vision [37].

Lastly, we have asked the participants to fill a survey with questions about their habits and game skills, and regarding the overall experiment considerations.

4.1.1 Practical consideration

Albeit the experiments have been conducted on 36 people, we have not considered in the analysis the data of participants with ID 11, 12 and 15 due to technical issues. Furthermore, the participants 20, 23, and 40 have experienced sickness during the VR session, although they have completed the experiment and their data have been considered during the data analysis. The acquired RAGA dataset and the data analysis source code are freely available to the academic community.[3]

5 Data analysis

The data analysis has consisted in four main steps: data filtering, extraction of features from raw data, selection of the most informative features, and application of an algorithm of supervised learning, in order to understand if these data can be used to predict the players' emotions during the game sessions.

Since the physiological data and the emotions self-assessment information have been acquired at two different frequencies (respectively 556Hz and 60Hz), it has been necessary to uniform the number of instances. We have decided to consider as final frequency the same frequency used to acquire physiological data (i.e., 556Hz). Thus, let a_f and v_f be respectively the values of Arousal and Valence at the instance $f \in \{1, 2, 3, .., F\}$, where F is the length of Arousal/Valence arrays, and let N the number of instances of their corresponding physiological data, we have calculated the new points in the interval $[1, F]$ to have a length equal to N. Consequentially, let $n1$ be a set of integers $\{0, 1, 2, .., N - 1\}$, the new set of points $nf = [1, F]$ is:

$$nf = 1 + n1 \odot ((F - 1) \oslash (N - 1)) \tag{1}$$

Where the $\odot$ and $\oslash$ are an Hadamard, respectively, product and division.

As a consequence, we have two arrays: $f \in N$ of length F, and $nf \in Q$ of length N, with values in $[1, F]$. Thus, we have applied a linear interpolation on each Arousal and Valence data in nf. For each element in nf, we have considered the integer neighbors in f

[3] https://github.com/grano00/GameVRRacingPhysioDB

such that $f_j \leq nf_i \leq f_k$. Lastly, we have generated the new Arousal (an) and Valence (vn) points for every i in $\{1, 2, 3, ...N\}$.

$$an_i = \frac{a_j(f_k - nf_i) + a_k(nf_i - f_j)}{f_k - f_j} \tag{2}$$

$$vn_i = \frac{v_j(f_k - nf_i) + v_k(nf_i - f_j)}{f_k - f_j} \tag{3}$$

Where a_j/v_j and a_k/v_k are the annotation values in the corresponding position of the elements in f vector (i.e., f_j, f_k).

As a consequence, we have calculated two new sets of values with the same length, and frequency, of the physiological data. Through these operations, we have aligned the self-assessment vectors with the acquired physiological signals. Having the same sample frequency, we can process the features (i.e., the information extracted from the physiological signal) and the ground truth (i.e., the arousal and valence vector) coherently. Consequentially, in the following session, we will refer with the term $SampleRate$ at the frequency of both kind of vectors, i.e. 556 Hz.

5.1 Data filtering

All the physiological and emotional tagging data have been separated for each game, thus getting 4 sets of data for each participant: RO, PCars, RO in VR, PCars in VR. Thus, we have filtered the groups of data separately in order to minimize the probability to introduce data patterns that may alter the future analysis. For each type of physiological information, we have centered and scaled, using the standard deviation, the acquired data [57]. Let x be a set of physiological data (e.g., ECG signal). The centered and scaled of a generic physiological signal x will be equal to the outcome of the function $cs(x)$:

$$cs(x) = \frac{x - \mu(x)}{\sqrt{\frac{\sum_{i=1} N(x_i - \mu(x))^2}{N-1}}} \tag{4}$$

where $\mu(x) = \frac{1}{N}\sum_{i=1}^{N} x_i$; $x \in [1, N]$.

The emotional tagging data is expressed as an integer value from 0 to 100. Thus, the value 50 underlines a neutral emotion and, as a consequence, we have centered the data at this value, scaling them in order to create a signal in the interval [-1,1]. Furthermore, we have filtered some frequencies of the physiological data in order to remove the noise generated by data acquisition. Starting from ECG data, we were interested to collect the Heart Rate (HR) of the players during video games fruition. As suggested by Fedotov [29], the most informative frequencies to understand the HR are between 5 to 30 Hz. We have considered an upper band frequency of 35 Hz, filtering the data using an Equiripple FIR band-pass [48], in order to consider also a possible excessive increase of heartbeats. All the EMG signals have been filtered using a high-pass Equiripple FIR with the cut-off band at 20Hz as suggested in [7]. As breathing produces a low-frequency signal alteration, we have applied a moving average filter [60] with a window length equal to $SampleRate$ (1 second in time domain). This value has been selected to not deeply alter the signal and, at the same time, to be able to smooth it in the high frequencies (i.e., noise). As a consequence, we have obtained the average temperature, under the participants' nose, over a period of 1 second. For the GSR, we have applied a 1^{st} order Butterworth low-pass filter with 5Hz cutoff. Through Ledalab [4], we have also extracted the Tonic (SCL) and Phasic (SCR) information [59], and we have added them to the set of physiological signals in RAGA.

Furthermore, as said in Section 4.1, the emotion tagging phase have been performed during the video playback. In order to improve the emotion recall, we have suggested to the users to not stop or rollback the video. Anyway, in some cases, some participants have performed a wrong assessment, and they have quickly corrected the mistake by moving back the SAM/AS slider pointer. This action has produced high-frequency noise in the final assessment data as in the respiration signal. Consequentially, we have removed this high-frequency information applying a moving average filter with a coefficient equal to $1/Samplerate$.

5.2 Features extraction

After the data filtering, we have proceeded to extract features from the physiological data. We have considered to analyze and to predict the data at 1-second precision, thus the variables output have been structured considering 1 data per second. The ECG data presents three signal deflections that repeat over time and occur in rapid succession. They correspond to the depolarization of the heart ventricles. The name of these signal deflections is *QRS complex* [26], where Q wave is a downward deflection, followed by R wave, an upward deflection, and in turn followed by a second downward deflation, the S wave. The distance between two equal points in two repetitions of QRS complex provides the necessary information to calculate the HR. Usually, the R wave is used for this purpose, thus we have detected the RR interval on ECG signal using OSEA algorithm [45]. The algorithm returns the points where the QRS waves are located, we have thus selected only the fiduciary points of R waves. Anyway, the algorithm is not able to identify the RR information in noisy signals, and ECG information can be subject to noise generated by motion artifacts. Usually, during the ECG, the clinical staff ask to reduce the movement in the specific locations where the electrodes are placed; anyway, this is not applicable to our experiments because the participants should not have movements limitations in order to better interact with the games. Thus, we have designed an algorithm aimed at creating or removing the R points on noisy signals according to the RR interval dimension. We have set a score for each RR interval (rr) according to:

$$rrscore = \|rr/\mu(rr)\| - 1 \tag{5}$$

where $rrscore = 0$ indicates a correct interval dimension, $rrscore > 0$ means an underestimation, and $rrscore < 0$ underlines an overestimation (Fig. 5). Thus, we have removed the overestimated points, and we have created new points in the underestimated area. To add new points, we have controlled each $rrscore$, adding virtual points when $rrscore \geq 0$ and deleting the points when $rrscore \leq 0$. In particular, the number of points added or deleted are in accordance with the $rrscore$ value. If the score is greater than 0, we add $rrscore$ points with the following position:

$$r_{new} = r_1 + ((r_2 - r_1) * ((i + 1)/(rrscore + 1))) \tag{6}$$

Where: $r_1 = r$ point at beginning of the rr interval; $r_2 = r$ point at the end of the rr interval; i is an incremental counter in $[1, 2, ..rrscore]$. On the contrary, if the score is less than 0, we removed n points at the beginning of r_1.

To have RR data in a more understandable measure unit, we have converted the RR distances in second domain, and we have multiplied the result by 60 in order to obtain Beats Per Minute (BPM): $bpm = 60 * SampleRate/rr$. Lastly, we have applied a moving average filter with a coefficient equal to $1/4$ on the BPM data, and we have acquired the information at second precision extracting the average BPM value.

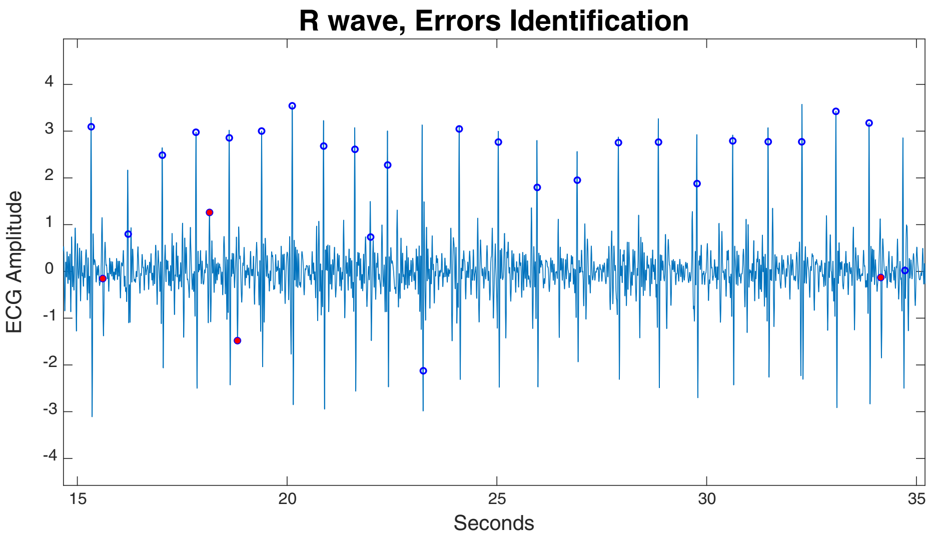

Fig. 5 Example of overestimation points due to noise in ECG signal. The red points are considered as overestimated

For the respiration, we were interested in the ratio between the exhalation and inhalation. Thus, we have analyzed the breath signal, and we have extracted all the upper and lower peaks. The difference between the peaks gives the time spent between one breath and another.

Finally, we have extracted a set of features from the EMGs, GSRs (Raw, Tonic, and Phasic), and respiration data using the previous information. Usually, this kind of analysis is focused on the central area of an approximation window, anyway, we were interested to understand if it is possible to predict the players' emotion during the game session, thus only the past information of these signals was available. As a consequence, we have designed an approximation window of 3 seconds in order to analyze the last window area. We have grouped the extracted features in Table 1. In literature [28, 46, 50, 64, 66, 69], these features are used for the EMG signal analysis, anyway, most of them can be considered also for other physiological signals. For sake of brevity, we present only the features that are not explained in the above-mentioned papers or those extended to fit with the experiments purpose. A detailed explanation of each feature is available in [40].

For each approximation window in EMGs, GSRs, and Respiration data, we have acquired the average spectrum power, and we have also extracted the average power in a subset of the frequency range, in particular in steps of 5Hz regarding EMGs, and 0.05Hz for GSRs and Respiration. Let $x \in [x_1, x_2, .., x_N]$ a generic physiological signal with N samples, we have also calculated two different Modified Mean Absolute Value (ModMAV) functions. ModMAV is comparable to the mean of absolute values ($MAV(x) = 1/N \sum_{n=1}^{N} |x_n|$), but it is applied to the results of an elaboration of the original data. Considering an approximation window of length N, the first modified function ($ModMAV1(x1_n)$) implements a *fade in*, where:

$$x1_n = \begin{cases} x_n/2 & \text{if } n < 0.75N \\ x_n & \text{otherwise} \end{cases} \tag{7}$$

Table 1 Features extracted by raw data

Feature Collected without Approximation Window	
ECG	BPM
Respiration	Breath Rate
EMGs, GSRs, and Respiration	Raw Data
Features Collected with 3 sec Approximation Window	
Common Features for EMGs, GSRs, and Respiration data	Band Power, Power, Integral [50], Mean Amplitude, Mean Absolute Value MAV [28], Precise Mean, Mod MAV 1 [66], Mod MAV 2 [66], MAV Slope MAVS [66], Root Mean Square RMS [66], Variance σ^2 [64], Waveform Length WL [28], Zero Crossings ZC [28], Slope Sign Changes SSC [28], Willison Amplitude WAMP [66], Simple Square Integral SSI [66], Frequency Median FMD [64], Frequency Mean FMN [64], Modified Frequency Median MFMD [66], Modified Frequency Mean MFMN [66], Frequency Ratio FR [46]
Features of GSRs data	MIN, MAX, # of Peaks, Mean Amplitude of Peaks

while, the second modified function ($ModMAV2(x2_n)$) implements a more gradual *fade in* applying an incremental weight between 0 to 1, with:

$$x2_n = \begin{cases} x_n * n/0.75N & \text{if } n < 0.75N \\ x & \text{otherwise} \end{cases} \tag{8}$$

Lastly, only for GSRs data, we have collected the minimum and maximum values, the number of peaks, and their average amplitudes as suggested in [10].

After the feature extraction process, we have thus collected: 1 feature for ECG, 38 features for Respiration, 62 features for each GSR signals, and 77 features for each EMG signal. Therefore, for each experiment, we have a set of 610 variables.

5.3 Machine learning approach: feature selection and supervised learning

For each participant, we have consider 9 groups of analysis:

RONOVR : RO game session with a standard monitor
PCarsNOVR : PCars game session with a standard monitor
ROVR : RO game session in VR
PCarsVR : PCars game session in VR
NOVR : the merged data of game sessions with a standard monitor
VR : the merged data of game sessions in VR
RO : the merged data of RO in both configurations
PCars : the merged data of PCars in both configurations
Player : all the data collected on each participant

For each group, we have selected a subgroup of the extracted features, and we have used them to train a Machine Learning (ML) algorithm that uses as target variable the Arousal and Valence values, provided by participants during the emotional tagging to produce a regression hypothesis.

We have decided to consider only a subgroup of features, in order to avoid the *curse of dimensionality* [23]. In addition, we have removed redundant or irrelevant variables to improve the ML algorithm performance. In order to not manipulate the data, we have designed an automatic procedure to select, for each analysis, the most informative features. As a first step to remove features that are not relevant for the ML method, we have calculated the Pearson linear correlation between each feature and the Arousal and Valence arrays. Thus, we have tested the hypothesis of no correlation and we have stored only the features rejecting the hypothesis (*p-value* < 0.05) in F. On the subgroup of features that respect this constraint, we have applied a method able to identify only a restricted number of variables identified as most informative. Our algorithm is a modified Sequential Floating Forward Selection (SFFS) [14] method, which returns the set of features able to minimize the regression error. The pseudo-code of the method is shown in Algorithm 1. In the algorithm, the ERR(x) function returns the error value of a generic predictor method. In our specific case, we have used a 10 Cross Validation (10CV) [55], where a set of Random Forest (RF) [49] are trained: we have used 100 trees, as suggested in [63], with 1/3 of features for each decision split. For each fold, we have extracted the RMSE index, and we have divided it by the number of the elements available in the test set.

Algorithm 1 Modified version of SFFS. F is the subset of features selected in the first step of feature selection.

```
let Y = {Ø}
let F = {ExtractedFeatures with corr. p-value < 0.05 }
let oldY an empty set of Y
while length(Y) < length(F) do
    ERRListOne = {Ø}
    for each V ∈ (F − Y) do
        ADD ERR({Y ∪ V}) to ERRListOne
    end
    Y = FEATURES that MIN(ERRListOne)
    if Y ∈ oldY OR
    ERR(Y) = 0 OR
    ERR(Y) is a local minimum then
        RETURN Y
    end
    ADD new line in oldY with Y features

    ERRListTwo = {Ø}
    for each V = element ∈ Y do
        ADD ERR({Y − V}) to ERRListTwo
    end
    if MIN(ERRListTwo) < MIN(ERRListOne) then
        Y = FEATURES that MIN(ERRListTwo)
    end
end
```

This algorithm starts with an empty set of features Y. It trains a different model for each feature in F (using only one feature at time), and it collects in the Y array the features that minimize ERR(x). The method then dynamically adds and removes a feature at time, collecting the subset which minimizes the regression error. The algorithm stops when finds a local minimum error, or if, in the process, it collects a subset of features already used in previous iterations.

Considering only the selected features, acquired through the feature selection process, we have tested different supervised learning techniques in order to verify which one performs better on our dataset. This preliminary comparison has been performed on the group which consider all the player's data (i.e., group 9). As a consequence, we have listed a set of potential regression algorithms:

– **Support Vector Machines** [25] with *Linear* kernel
– **Support Vector Machines** [25] with *Gaussian* kernel
– **Random Forest** (RF) [11]
– **Gradient Boosting** [33] of trees (GBoT)
– **Gaussian Process Regression** (GPR) [85]

Each algorithm has been tested using a 5CV (5 fold Cross Validation) on each group of analysis. Thus, we have used the algorithm which, in average, has provided a model reaching a better accuracy on the depended variables prediction (i.e., VA self-assessment values) to infer the players' emotions. If the error levels of the subgroup of models that provide the better accuracy do not present a significant difference, all the models have been trained n times, and the algorithm with the average better accuracy have been selected.

In our specific case, the GPR, also known as *Kriging*, has provided the better results (Fig. 7). Moreover, it is a different method than the ML algorithm used in the second step of feature selection: this should minimize the probability to obtain biased results. Although the GPR computational cost is quite high ($O(n^3)$), the limited number of instances for each experiment does not affect excessively the overall performances of the method. In our approach, to reduce computational time, we have lowered the number of cross-validation folds from 10, used for the feature selection, to 5. Reducing the number of folds, we have increased the number of elements in the test set and, consequentially, the algorithm returns a more pessimistic error. For each experiment, the average time required for a single core to train the *Kriging* algorithm, on the computer configuration presented in Section 4.1, is $\approx$ 4 seconds, and the 5CV, computed in multi-thread, requires in average $\approx$ 12 seconds. Anyway, as we discuss in Section 6, the final results show how this choice does not affect excessively the accuracy of the prediction of *Kriging*.

6 Data analysis results

In the following subsections, we present the data analysis outcomes. In Section 6.1, we compare the importance of the different signals, and their related features. Lastly, in Section 6.2, we present the results of the different prediction models, with a particular attention on the results of the final model.

6.1 Features for players' emotions prediction

In literature, there are some feature selection models able to define, for each feature, which is more effective in the computation of an accurate prediction. For example, RF provides

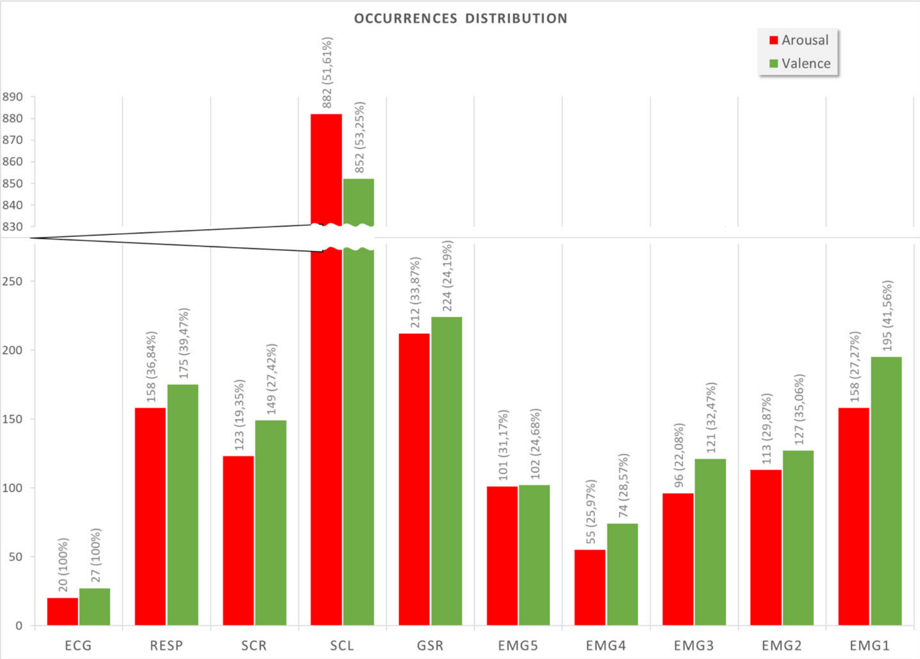

Fig. 6 The figure presents, for each feature, the number of occurrences, grouped by their original signals. On the top of each bar is presented the percentage of features involved in the process. The correspondence of EMGs number are shown in Fig. 1

an index for each feature, considering which split of the tree will be the most effective to distinguish the classes, and reporting the importance through a standard index, like e.g., Gini index [62]. Unfortunately, the feature selection algorithm designed starting from SFFS does not provide an importance ranking of each feature using a standard index. As a consequence, we are not able to identify accurately which feature is the most important. Anyway, we are able to obtain the information about the features more involved in emotions recognition during racing video games, considering the number of times that a particular feature is selected by our algorithm. In particular, the proposed method selects an average of 6.46 features to analyze the Arousal data, and 6.89 for the Valence, for a total of, respectively, 185 and 206 variables involved in the prediction process. In Fig. 6 the number of occurrences of each feature is shown, grouped by the signal of origin.

6.2 Emotions prediction outcomes

We have also evaluated the efficacy of the ML algorithm to predict the emotions, in the Valence and Arousal space, during the video games sessions. As stated in Section 5.3, after a comparison of the prediction errors between different ML algorithms (see Fig. 7), we have selected the model which has minimized the regression error (i.e., GPR).

Considering the GPR, we have designed a hypothesis testing a set of σ_{gpr} values in a range $[10^{-3}, \sigma(x)]$ for each participant and experiment as suggested in [77]. Then, we have calculated the results of 5CV, and we have acquired the Root Mean Square Error (RMSE) between observed and estimated data. Thus, we have calculated the Normalized RMSE

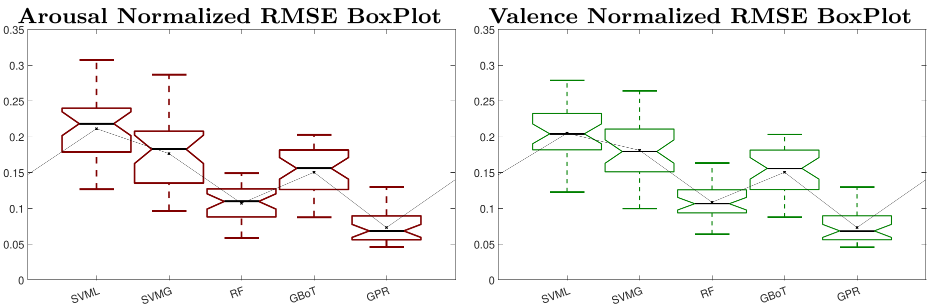

Fig. 7 The box plots present the NRMSE distribution across the experiment "**Player**" (see Section 5.3) using different ML algorithms. The cross inside the boxes underlines the average value, while the outliers are represented with the stars

(NRMSE), presented in (9), in order to have scale-free results.

$$NRMSE(y, \hat{y}) = \frac{RMSE(y, \hat{y})}{max(\hat{y}) - min(\hat{y})} \tag{9}$$

Where y are the predicted values, while $\hat{y}$ is the observed array (i.e., the ground truth).

In Fig. 8 and in Table 2 are shown, respectively, the box plots that indicate the NRMSE dispersion of Valence and Arousal, and the experiments numerical results.

The NRMSE may vary between 0.0, that indicates a perfect overlap between the estimator $(\hat{y})$ and estimated (y) sets (i.e., a perfect prediction, where $\hat{y} = y$), and 1.0 that indicates two divergent sets of data. In latter case, we can consider, in a simplified example, that $\hat{y}$ and y may have only two values (e.g., 0 and 1). If, for each element of the sets, $\hat{y} \neq y$, the NRMSE will be equal to 1.0. This range is respected only under two constraints: if $max(\hat{y}) \geq max(y)$ and $min(\hat{y}) \leq min(y)$, which are satisfied in our case study. In Fig. 8, we present the box plots of the acquired NRMSE data for each kind of experiment. Plots

Table 2 NMRSE results for each experiment

Experiment	Arousal				Valence			
	Mean	*Sigma*	*MIN*	*MAX*	*Mean*	*Sigma*	*MIN*	*MAX*
PCars with Monitor	0.090	0.037	0.036	0.241	0.095	0.036	0.041	0.236
RO with Monitor	0.089	0.032	0.034	0.152	0.097	0.033	0.037	0.162
Monitor	0.086	0.030	0.045	0.195	0.090	0.024	0.050	0.158
PCars in VR	0.091	0.035	0.036	0.189	0.096	0.045	0.028	0.195
RO in VR	0.091	0.031	0.041	0.167	0.093	0.029	0.045	0.165
VR	0.087	0.034	0.042	0.187	0.080	0.028	0.051	0.176
PCars	0.079	0.026	0.035	0.135	0.082	0.025	0.044	0.147
RO	0.082	0.032	0.034	0.192	0.089	0.028	0.037	0.161
Player	0.073	0.020	0.046	0.125	0.078	0.021	0.050	0.121
AVERAGE	0.085	0.031	0.039	0.176	0.089	0.030	0.043	0.169
MIN	0.073	0.020	0.039	0.125	0.078	0.021	0.028	0.121
MAX	0.091	0.037	0.046	0.241	0.097	0.045	0.051	0.236

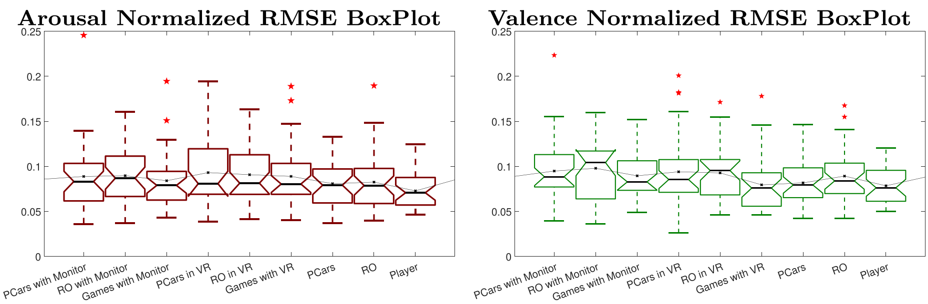

Fig. 8 The box plots present the NRMSE distribution between the estimated data collected using the GPR (trained with different hyperparameters) and the observed self-assessment data across the experiments (see Section 5.3). The cross inside the boxes underlines the average value, while the outliers are represented with the stars

related to each single prediction or box plots that use different indexes are available on the RAGA homepage presented in Section 4.1.1.

7 Discussion

In the following subsections, we discuss about the different experiment results. In Section 7.1, we focus on the outcome of the dataset and on the analysis about the data acquired by the survey, evaluating the reliability of the emotion self-assessment provided by the participants. In Section 7.2, we discuss about the important features acquired by the physiological signals. Finally, in Section 7.3, we provide an interpretation of the model outcome.

7.1 RAGA dataset and overall experiment outcomes

Summarizing, RAGA is a public dataset, available to the scientific community, which can be adopted as a benchmark or for research proposes. It contains physiological information (see Section 3.2 and Table 4 for the complete description and the list of the collected data) and a ground truth, represented in a 2-dimensional space (valence and arousal). Albeit the physiological signals, and the self-assessment have been acquired at different sample frequencies, respectively 556 Hz and 60 Hz, we have aligned they and we have provided both, aligned and raw data, in the dataset.

Table 3 presents a comparison of different datasets, considering the features (i.e., kind of stimuli, physiological data acquired, type of emotions identification) and the annotation methods between RAGA and the datasets mentioned in Section 3.1.

After each session, we have asked each participant to evaluate the overall experience of the experiment. Almost no participant has reported a significant discomfort due to the motion sickness ($\mu = 2.85$, $\sigma = 2.59$, in a rank between 1 to 10), and the 82% have claimed they have not been disturbed by the sensors used in the experiment. One of the crucial hypothesis at the basis of the proposed method is that the affective self-assessment performed by the participants, using the video tagging procedure described in Section 4, is reliable. Since the self-assessment values have been collected in continuous, over all the

Table 3 Comparison among available affective datasets and RAGA dataset. The physiological signals considered are: EEG = electroencephalography (with the number of channels), ECG = electrocardiogram, BVP = blood volume pulse, GSR = galvanic skin response, Facial EMG = electromyography placed on participant face (with the number of muscles considered), Resp = respiration, Temp = temperature, Gaze = eye gaze tracking. The last 3 columns define the type of emotion identification, where: ES = Emotion Space, in type column D = Discrete, C = Continuous, and in Annotator column S = Self Report, E = External Report with, in brackets, the number of annotators

Dataset	**Eight-Emotion Sentics Data [67]**	**MAHNOB-HCI [78]**	**DEAP° [53]**	**RECOLA [70]**
Stimulus	Sentic [14]	Video, Images	Music Video	Collaborative Work
Subj.	1	30	33	16
EEG	-	32 Ch.	32 Ch.	-
ECG	-	✓	✓	✓
BVP	✓	-	✓	-
GSR	✓	✓	✓	✓
Facial EMG	1	-	1	-
Resp	✓	✓	✓	-
Temp	-	✓	✓	-
Gaze	-	✓	-	-
ES	8-D emot.	PAD	PAD + Liking	VA
Type	D	D	D	C
Annotator	S	S	S	6(E) + S

Dataset	**DECAF [1]***	**OPEN EmoRec II [72]**	**AMHUSE [6]**	**RAGA**
Stimulus	Movie Clips/Music Movie	Mental Puzzles	Movie Clips	Racing Video Games in VR
Subj.	30	30	36	33
EEG	-	-	-	-
ECG	✓	-	-	✓
BVP	-	✓	✓	-
GSR	-	✓	✓	✓
Facial EMG	-	2	-	5
Resp	-	✓	-	✓
Temp	-	-	✓	-
Gaze	-	-	-	-
ES	PAD	VA	VA	VA
Type	D	D	C + D	C
Annotator	S	4(E) + S	4(E) + S	S

*The dataset provides also the signals of: EMG on Trapezius muscle, MEG in order to measure the brain activity, and EOG to investigate the eye movements

°The dataset provides also the signals of: EMG on Trapezius muscle, and EOG to investigate the eye movements gameplay session, we designed a set of discrete markers in order to determine if the participants were able to label accurately their emotions. It has been performed as post-experiment evaluation of the participants' self-assessment accuracy. Actually, we have asked the participants to fill a survey with a set of questions aimed to evaluate how accurate they think they have been during the emotion tagging stage. From the data of these surveys, their perceived precision during the self-assessment emotion tagging has been equal to 7.48 for Arousal data, and 7.30 regarding Valence data (in a rank between 1 to 10). Furthermore, the participants have ranked their concentration (Arousal) during racing games equal to 0.56 (in a rank between -1.0 and 1.0). They have also evaluated the games ability to arouse emotions (Valence) equal to 0.28. The ability of the games to raise a positive emotion has been evaluated 0.34, while 0.14 is the evaluation for a negative emotion. Considering the data acquired from each participant during the emotions tagging phase, the mean value of Arousal is 0.41 ($\sigma = 0.44$), and for Valence is 0.18 ($\sigma = 0.49$). Thus, we can consider the answers, although slightly overestimated, in line with the information given by emotion tagging, validating thus the reliability of the self-assessment stage. These scores have been summarized in Table 4.

The self-assessment data distribution has been reported in Fig. 9. Albeit the self-assessment data are slightly contracted at the ends of valence and top of arousal, and in positive values for both type of emotions, we can consider the data balanced in the 2-dimensional space. For this reason, we were able to investigate accurately almost all type of players status in the 2-dimensional model.

Lastly, we can propose a potential interpretation between some of the acquired results, and the *Spatial Presence* and theory of *Flow* concepts, introduced in Section 1. In the final survey, the 88% of participants have reported more intense emotions during the sessions with VR games; however, we have not found a significant difference in the Valence and Arousal values between the sessions with or without VR. However, some researches have assumed higher level of *Spatial Presence* condition when using immersive devices. Our results seem to suggest that the evaluation of the *Spatial Presence* condition can not be defined using the 2-dimensional Valence/Arousal space used in our experiment. As a consequence, future analysis may investigate the use of a more complex system for the identification of emotions (like e.g., PAD). Moreover, analyzing the Valence and Arousal values for each game, we can notice how these values are, in average, quite positive. In fact, PCars has collected $\mu = 0.38$ for Arousal and $\mu = 0.15$ for Valence, while RO has obtained $\mu = 0.44$ for Arousal and $\mu = 0.21$ for Valence. Considering the average absence of emotions tending to averseness, and also the high level of positive feedback provided by the players through the post-experiment survey, we can hypothesize that the players have been relevantly absorbed (thus, in the *Flow* state) during the game sessions.

7.2 What the players' body said

According with the results presented in Section 6.1, the tonic component (Skin Conductance Level, SCL) of GSR is one of the most informative signals: its features are selected 882

Table 4 Self-Assessment perception of participants compared with the labeled data

	Self-assessment	Post-experiment evaluation
Average arousal	0.41	0.56
Average valence	0.18	0.28

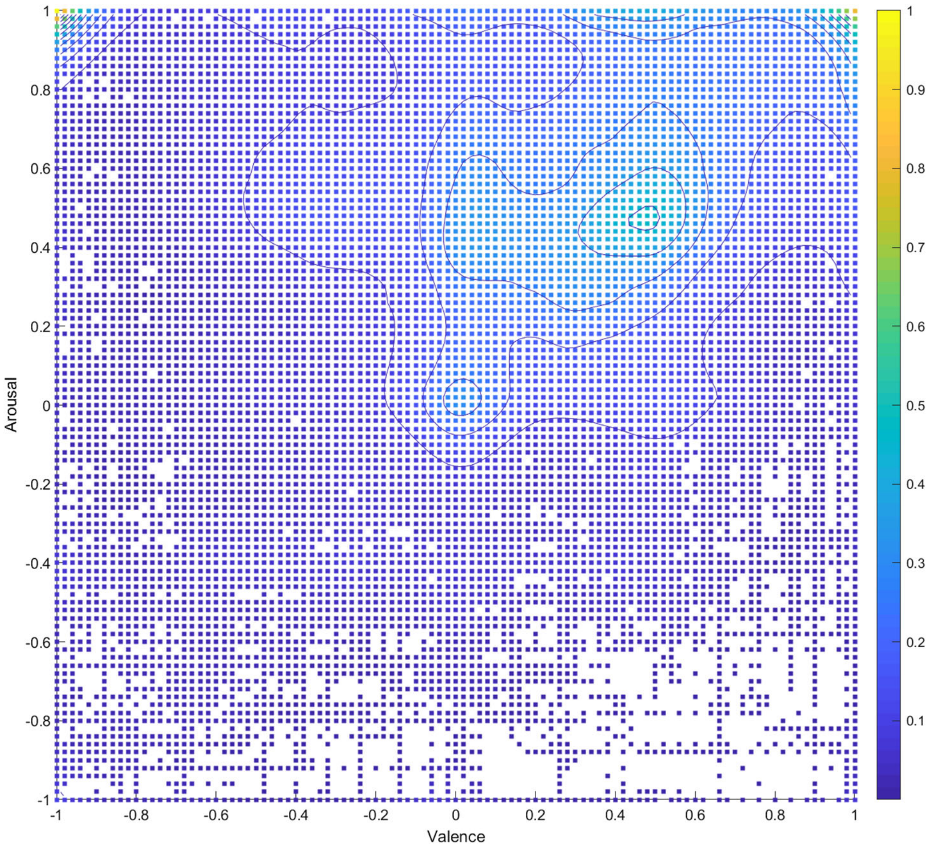

Fig. 9 Distribution of self-assessment evaluations collected on RAGA participants

times for Arousal and 852 times for Valence analysis. Since the "tonic value stands for an activity that shows a certain amount of continuity over time" [75] and it "is related to a person's overall arousal" [75], we can consider the racing games consistent with the achieved results, since they are designed to maintain the player's attention over the time, increasing stress levels as the player approaches the end of the race.

Focusing, for example, on Player analysis, which considers all the data of each participant during the overall experimental session (see "**Player**" in the list presented in Section 5.2), the selected variables used to predict the VA target values are in line with the features occurrences distribution above mentioned (e.g., in Fig. 10 are shown the occurrences of the different features during the Valence analysis).

The proposed feature selection method can be also used to validate novel features, since it compares all the features and it extracts only the most interesting. It is also designed to work autonomously, looking for the best set of features that maximizes the data prediction, and, as a consequence, the ground truth of the hypothesis. Lastly, it provides a history of the selected variables, structured in the order in which the features are selected. As a consequence, during the test of a novel feature, the algorithm may provide (with a certain approximation) an index of informative level which contains the variable, according to its position in the history, beside the boolean information of feature importance (i.e., considered or not considered).

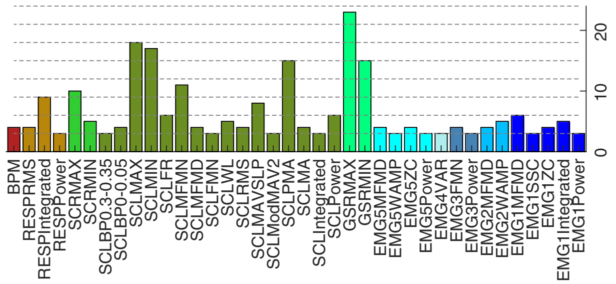

Fig. 10 Valence features occurrences on player analysis. The number of occurrences of each selected features are shown on the vertical axis. The figure shows only the features that have been selected in 3 or more occurrences

For a detailed description of the different features and an exhaustive discussion of their importance see [40].

7.3 Model interpretation

According to the results described in Section 6.2, the average regression error is quite low, which underlines the ability to design a hypothesis able to predict the values of Arousal and Valence vectors during a video game session with a precision of 1 second. Moreover, in both dimensions, the cumulative participant's data (i.e., the Player analysis) provide, on average, a smaller regression error. These results suggest that the model is sensitive to the amount of acquired from a participant data, independently by the context (i.e., VR and Monitor, and the selected video game). Consequentially, this method should produce better prediction according to the time employed to play at games of the same genre.

The GPR has a non-parametric approach, which only assumes that similar data points, defined by a covariance function, are close in the output space. As suggested in [13], GPR is robust to errors in input sources (e.g., loss of an electrode contact), since it depends directly on data, and not on the features' relationship. Consequentially, its characteristics and the achieved results in the experimental session have supported our decision to consider this algorithm reliable in our experimental procedure.

8 Conclusion and future works

In this paper, we have presented RAGA, an affective dataset based on the acquisition of physiological signals from video game players. We have provided an overview of the considered physiological data and of the hardware setup used to acquire it. We have collected the physiological signals and self-assessment information from a set of participants playing racing games. The players have played in two different environments, using a monitor, and a VR headset. Furthermore, we have provided an analysis of the relevance of each signal and of their contribution to predict the players' emotional state. Lastly, we have described the ML algorithm used to design a hypothesis able to predict the emotions of participants at second precision. The results seem to confirm the validity of the experimental framework, since the ML model obtain low errors in each experiment. Thus, the considered algorithm is able to predict, with a certain precision, the self-assessment emotion into the Valence/Arousal

2-dimensional space. Moreover, to the best of our knowledge, there are not freely available databases of affective annotation using video game as stimuli.

In a future extension of this work, we will design a racing game based on the prediction features introduced in the current research: this game will adapt its difficulty on the basis of the players' emotions, trying to keep the players' entertainment at a qualitative standard required by the video game industry. Moreover, future works will address two challenges: to design an integrated set of gaming devices able to reveal the physiological data of the players, and to identify game features able to provide the evaluation of players mental state, avoiding a specific self-assessment for the players. The former can be achieved by the integration of the sensors in game devices. Almost all the electrodes used in the experiments are located near the hands and face of the players; therefore, it is possible to design their integration into, respectively, a gamepad and a VR headset. Considering the second challenge, currently the prediction model is linked to the individual player, as it has to be trained on data collected by each player. As a consequence, our algorithm needs to acquire, for each subject, the self-assessment data in order to provide the target variables to the ML algorithm. Anyway, during the game design process, different options aimed at the acquisition of the emotion labeling can be considered, for example borrowing gamification elements in the final game [22, 53]. A possible approach could be the design of a mini-game that asks to evaluate the emotions on the game highlights, providing, after the identification, an in-game reward. Another solution can be represented by the adoption of a different ML algorithm, with a general hypothesis aimed at predicting the emotion on a wide range of players avoiding the limitation to request the self-assessment information by each player. It could be supported by external annotators in order to decrease the noise due to the emotion tagging variability. Albeit it is commonly used in affective computing research, the time required to provide all the evaluations may not comply with industry production standards. Furthermore, an additional difficulty to use external annotators is caused by the VR headset during some experiment phases. During the experiment, we have curbed the problem asking the participants to evoke their emotional state immediately after the game session. Anyway, for an external annotator, it could be a challenge to identify the emotions of players with part of the face covered. The present work can also contribute to introduce a modality to study the players' affective, widely used in *Affective Computing*, in the video game research field.

Acknowledgments The authors want to thank 34BigThings for having provided a sample copy of RedOut. We also thank Prof. Nicolò Cesa-Bianchi, Dr. Vittorio Cuculo, and Prof. Giuseppe Boccignone for their comments and suggestions, that greatly improved the overall research.

References

1. Abadi MK, Subramanian R, Kia SM, Avesani P, Patras I, Sebe N (2015) Decaf: meg-based multimodal database for decoding affective physiological responses. IEEE Trans Affect Comput 6(3):209–222
2. Angie AD, Connelly S, Waples EP, Kligyte V (2011) The influence of discrete emotions on judgement and decision-making: a meta-analytic review. Cogn Emotion 25(8):1393–1422
3. Bartle RA (2003) Designing virtual worlds. New Riders
4. Benedek M, Kaernbach C (2010) A continuous measure of phasic electrodermal activity. J Neurosci Methods 190(1):80–91
5. Betella A, Verschure P (2016) The affective slider: a digital self-assessment scale for the measurement of human emotions. PLoS ONE, 11(2)
6. Boccignone G, Conte D, Cuculo V, Lanzarotti R (2017) AMHUSE: a multimodal dataset for humour sensing. In: Proceedings of the 19th ACM international conference on multimodal interaction. ACM, pp 438–445

7. Boccignone G, Cuculo V, Grossi G, Lanzarotti R, Migliaccio R (2017) Virtual EMG via facial video analysis. In: International conference on image analysis and processing. Springer, pp 197–207
8. Boccignone G, Conte D, Cuculo V, D'Amelio A, Grossi G, Lanzarotti R (2018) Deep construction of an affective latent space via multimodal enactment. IEEE Trans Cogn Develop Syst 10(4):865–880
9. Bradley MM, Lang PJ (1994) Measuring emotion: the self-assessment manikin and the semantic differential. J Behav Therapy Exper Psych 25(1):49–59
10. Braithwaite JJ, Watson DG, Jones R, Rowe M (2013) A guide for analysing electrodermal activity (EDA) & skin conductance responses (SCRs) for psychological experiments. Psychophysiology 49(1):1017–1034
11. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
12. Castellar EN, Oksanen K, Van Looy J (2014) Assessing game experience: heart rate variability, in-game behavior and self-report measures. In: 2014 Sixth international workshop on quality of multimedia experience (QoMEX). IEEE, pp 292–296
13. Caywood MS, Roberts DM, Colombe JB, Greenwald HS, Weiland MZ (2017) Gaussian process regression for predictive but interpretable machine learning models: an example of predicting mental workload across tasks. Frontiers Human Neurosci 10:647
14. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28
15. Clynes M (1978) Sentics: the touch of the emotions. Anchor Press/Doubleday
16. Conover MB (2003) Understanding electrocardiography. Elsevier Health Sciences
17. Critchley HD (2002) Electrodermal responses: what happens in the brain. Neuroscientist 8(2):132–142
18. Csikszentmihalyi M, Nakamura J (1979) The concept of flow. Play Learn, 257–274
19. Dalgleish T (20za04) The emotional brain. Nat Rev Neurosci 5(7):583
20. Darwin C (1872) The expression of the emotions in man and animals. Chicago University Press, USA
21. De Felice D, Granato M, Ripamonti L, Trubian M, Gadia D, Maggiorini D (2017) Effect of different looting systems on the behavior of players in a MMOG: simulation with real data, vol 181 LNICST
22. Deterding S, Dixon D, Khaled R, Nacke L (2011) From game design elements to gamefulness: defining gamification. In: Proceedings of the 15th international academic MindTrek conference: envisioning future media environments. ACM, pp 9–15
23. Donoho DL et al (2000) High-dimensional data analysis: the curses and blessings of dimensionality. AMS Math Challenges Lect 1:32
24. Draper JV, Kaber DB, Usher JM (1998) Telepresence. Human Factors 40(3):354–375
25. Drucker H, Burges CJ, Kaufman L, Smola AJ, Vapnik V (1997) Support vector regression machines. In: Advances in neural information processing systems, pp 155–161
26. Dubin D (2000) Rapid interpretation of EKG's, 6th edn. Cover Pub Co
27. Ekman P (1971) Universals and cultural differences in facial expressions of emotion. In: Nebraska symposium on motivation. University of Nebraska Press, pp 207–282
28. Englehart K, Hudgins B (2003) A robust, real-time control scheme for multifunction myoelectric control. IEEE Trans Biomed Eng 50(7):848–854
29. Fedotov A (2016) Selection of parameters of bandpass filtering of the ECG signal for heart rhythm monitoring systems. Biomed Eng 50(2):114–118
30. Fernandes T, Chec A, Olczak D, Ferreira H (2015) Physiological computing gaming: use of electrocardiogram as an input for video gaming. In: Proceedings of the international conference on physiological computing systems (PhyCS), SCITEPRESS (Science and Technology Publications, Lda), Angers, pp 11–13
31. Folgieri R, Lucchiari C, Granato M, Grechi D (2014) Brain, technology and creativity. Brainart: a BCI-based entertainment tool to enact creativity and create drawing from cerebral rhythms. In: Digital Da Vinci. Springer, pp 65–97
32. Fourati N, Pelachaud C (2018) Perception of emotions and body movement in the emilya database. IEEE Trans Affect Comput 9(1):90–101
33. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat, 1189–1232
34. Freeman DE (2003) Creating emotion in games: the craft and art of emotioneering. New Riders
35. Fullerton T (2014) Game design workshop: a playcentric approach to creating innovative games. CRC Press
36. Furnham A (1986) Response bias, social desirability and dissimulation. Person Individ Diff 7(3):385–400
37. Gadia D, Garipoli G, Bonanomi C, Albani L, Rizzi A (2014) Assessing stereo blindness and stereo acuity on digital displays. Displays 35(4):206–212

38. Gadia D, Granato M, Maggiorini D, Ripamonti L, Vismara C (2018) Consumer-oriented head mounted displays: analysis and evaluation of stereoscopic characteristics and user preferences. Mob Netw Appl 23(1):136–146
39. Granato M, Gadia D, Maggiorini D, Ripamonti L (2017) Emotions detection through the analysis of physiological information during video games fruition. In: Springer lecture notes in computer science (proceedings of 6th international conference of games and learning alliance - GALA 2017), vol 10653, pp 197–207
40. Granato M, Gadia D, Maggiorini D, Ripamonti L (2018) Feature extraction and selection for real-time emotion recognition in video games players. In: Proceedings - 14th international conference on signal-image technology and internet-based systems, SITIS 2018. IEEE
41. Granato M, Gadia D, Maggiorini D, Ripamonti L (2018) Software and hardware setup for emotion recognition during video game fruition. In: Proceedings of the 4th EAI international conference on smart objects and technologies for social good, Goodtechs '18. ACM, New York, pp 19–24
42. Granic I, Lobel A, Engels RC (2014) The benefits of playing video games. Am Psychol 69(1):66
43. Grodal T (2000) Video games and the pleasures of control. Media Entertainment: The Psychology of its Appeal, 197–213
44. Grossi G, Lanzarotti R, Lin J (2016) Robust face recognition providing the identity and its reliability degree combining sparse representation and multiple features. Int J Pattern Recogn Artif Intell 30(10):1656007
45. Hamilton P (2002) Open source ECG analysis. In: Computers in cardiology, 2002. IEEE, pp 101–104
46. Han JS, Song WK, Kim JS, Bang WC, Lee H, Bien Z (2000) New emg pattern recognition based on soft computing techniques and its application to control of a rehabilitation robotic arm. In: Proc. of 6th international conference on soft computing (IIZUKA2000), pp 890–897
47. Hazlett RL (2006) Measuring emotional valence during interactive experiences: boys at video game play. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 1023–1026
48. Herrmann O (1970) Design of nonrecursive digital filters with linear phase. Electron Lett 6(11):328–329
49. Ho TK (1995) Random decision forests. In: Proceedings of the Third international conference on document analysis and recognition, 1995, vol 1. IEEE, pp 278–282
50. Huang HP, Chen CY (1999) Development of a myoelectric discrimination system for a multi-degree prosthetic hand. In: 1999 IEEE International conference on robotics and automation, 1999. Proceedings, vol 3. IEEE, pp 2392–2397
51. Huizinga J (2014) Homo Ludens: a study of the play-element in culture. Martino Fine Books
52. Isbister K (2016) How games move us: emotion by design. Mit Press
53. Knutas A, Van Roy R, Hynninen T, Granato M, Kasurinen J, Ikonen J (2019) A process for designing algorithm-based personalized gamification. Multimed Tools Appl 78(10):13593–13612
54. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I (2012) DEAP: a database for emotion analysis; using physiological signals. IEEE Trans Affect Comput 3(1):18–31
55. Kohavi R et al (1995) A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Ijcai, vol 14, Stanford, pp 1137–1145
56. Koster R (2013) Theory of fun for game design. O'Reilly Media Inc.
57. Kreyszig E (2011) Advanced engineering mathematics. Wiley
58. Lazzaro N (2004) Why we play games: four keys to more emotion without story. Games Developer Conference
59. Lykken DT, Venables PH (1971) Direct measurement of skin conductance: a proposal for standardization. Psychophysiology 8(5):656–672
60. Lyons RG (2011) Understanding digital signal processing. Pearson Education India
61. Mauss IB, Robinson MD (2009) Measures of emotion: a review. Cogn Emot 23(2):209–237
62. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA (2009) A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. BMC Bioinform 10(1):213
63. Oshiro TM, Perez PS, Baranauskas JA (2012) How many trees in a random forest? In: International workshop on machine learning and data mining in pattern recognition. Springer, pp 154–168
64. Oskoei MA, Hu H (2006) GA-based feature subset selection for myoelectric classification. In: IEEE International conference on robotics and biomimetics, 2006. ROBIO'06. IEEE, pp 1465–1470
65. Parnandi A, Gutierrez-Osuna R (2018) Visual biofeedback and game adaptation in relaxation skill transfer. IEEE Trans Affect Comput, 1–1
66. Phinyomark A, Limsakul C, Phukpattaranont P (2009) A novel feature extraction for robust EMG pattern recognition. J Comput 1(1):71–80

67. Picard RW, Vyzas E, Healey J (2001) Toward machine emotional intelligence: analysis of affective physiological state. IEEE Trans Pattern Anal Mach Intell 23(10):1175–1191
68. Ravaja N, Salminen M, Holopainen J, Saari T, Laarni J, Järvinen A (2004) Emotional response patterns and sense of presence during video games: potential criterion variables for game design. In: Proceedings of the third Nordic conference on human-computer interaction. ACM, pp 339–347
69. Rechy-Ramirez EJ, Hu H (2011) Stages for developing control systems using EMG and EEG signals: a survey. School of Computer Science and Electronic Engineering. University of Essex, pp 1744–8050
70. Ringeval F, Sonderegger A, Sauer J, Lalanne D (2013) Introducing the recola multimodal corpus of remote collaborative and affective interactions. In: 2013 10th IEEE International conference and workshops on automatic face and gesture recognition (FG). IEEE, pp 1–8
71. Ripamonti L, Granato M, Trubian M, Knutas A, Gadia D, Maggiorini D (2018) Multi-agent simulations for the evaluation of looting systems design in MMOG and MOBA games. Simul Model Pract Theory 83:124–148
72. Rukavina S, Gruss S, Walter S, Hoffmann H, Traue HC (2015) Open_emorec_ii-a multimodal corpus of human-computer interaction. Int J Comput Electr Autom Control Inform Eng 9(5):977–983
73. Russell JA (1991) Culture and the categorization of emotions. Psychol Bull 110(3):426–450
74. Russell JA, Mehrabian A (1977) Evidence for a three-factor theory of emotions. J Res Person 11(3):273–294
75. Schmidt S, Walach H (1999) Electrodermal activity (eda): state-of-the-art measurement and techniques for parapsychological purposes. J Parapsychol 63(3):221–221
76. Sheffield B (2013) The journey to create journey – the quest for emotion. https://www.gamasutra.com/view/news/186282/The_journey_to_create_Journey_the_quest_for_emotion.php
77. Snoek J, Larochelle H, Adams RP (2012) Practical Bayesian optimization of machine learning algorithms. In: Advances in neural information processing systems, pp 2951–2959
78. Soleymani M, Lichtenauer J, Pun T, Pantic M (2012) A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput 3(1):42–55
79. Stojanović R, Čaplánová A, Kovačević Ž, Nemanja N, Bundalo Z (2015) Alternative approach to addressing infrastructure needs in biomedical engineering programs (case of emerging economies). Folia Medica Facultatis Medicinae Universitatis Saraeviensis 50(1):29–33
80. Tognetti S, Garbarino M, Bonarini A, Matteucci M (2010) Modeling enjoyment preference from physiological responses in a car racing game. In: 2010 IEEE Symposium on computational intelligence and games (CIG). IEEE, pp 321–328
81. Van Boxtel A (2010) Facial EMG as a tool for inferring affective states. In: Proceedings of measuring behavior. Noldus Information Technology Wageningen, pp 104–108
82. Venables PH, Christie MJ (1980) Electrodermal activity. Techniques Psychophysiol, 54(3)
83. Vishay (2017) Datasheet. NTCLE203E3...SB0. http://www.vishay.com/docs/29118/ntcle203.pdf
84. Vismara C, Granato M, Ripamonti L, Maggiorini D, Gadia D (2017) Analysis of stereoscopic visualization in a consumer-oriented head mounted display. In: Lecture notes of the institute for computer sciences, social-informatics and telecommunications engineering, LNICST, vol 195 LNICST, pp 274–283
85. Williams CK, Rasmussen CE (2006) Gaussian processes for machine learning the. MIT Press
86. Zafar M, Ahmed B, Al-Rihawi R, Gutierrez-Osuna R (2018) Gaming away stress: using biofeedback games to learn paced breathing. IEEE Transactions on Affective Computing - In Press

Marco Granato has received his B.Sc. in Computer Science (2013) and M. Sc. in Cognitive Science and Decision Making (2015) from the University of Milan. He is now a doctoral student at the Department of Computer Science at University of Milan and member of PONG - Playlab fOr inNovation in Games Laboratory. His actual research interest includes emotivity in video games.

Davide Gadia has received his M.Sc. (2003) and Ph.D (2007) in Computer Science from the University of Milan. Currently he is Assistant Professor at Department of Computer Science at University of Milan. His research interests concern mainly Computer Graphics, Video Game programming, Virtual Reality, and Digital Imaging. He is also interested in the processing of stereoscopic images and videos, and in the analysis of perceptual issues in the interaction with Virtual Environments. He is involved in the research activities of the PONG - Playlab fOr inNovation in Games Laboratory.

Dario Maggiorini is Associate Professor at the University of Milan, Italy. He joined the Department of Computer Science in 2003 where he is now co-directing the PONG - Playlab fOr inNovation in Games. In the past, Dario has been working on the Quality of Service in IP networks, multimedia content delivery, application level networking, and software architectures for service provisioning. Currently, his research interests focus mainly on software and network architecture to support entertainment applications and on content, service, and gaming provisioning in distributed environments.

Laura A. Ripamonti has graduated in Engineering and Managerial Sciences at Politecnico di Milano and then got a Ph.D. in Computer Science at University of Milan (Italy). Currently, she is Assistant Professor at the Department of Computer Science of the University of Milan, where she is co-directing the PONG - Playlab fOr inNovation in Games and a master degree on Video Games for Computer Science. Her research interests focus on video games, with a special interest for MMORPGs (Massively Multiplayer Online Role-Playing Games) and Applied Games. In particular, she investigates the relations connecting game design issues to social interaction and AI applications. Since 2010, she lectures in game design to master students in computer science.