Physiological-Based Difficulty Assessment for Virtual Reality Rehabilitation Games

Pedro Rodrigues
pme.rodrigues@campus.fct.unl.pt
Department of Physics, NOVA School of Science and Technology
Caparica, Portugal

Micaela Fonseca
micaela.fonseca@ulusofona.pt
Lusófona University/HEI-Lab: Digital Human-Environment Interaction Lab
Lisbon, Portugal
LIBPhys – Laboratory of Instrumentation, Biomedical Engineering and Radiation Physics
Caparica, Portugal

Phil Lopes
phil.lopes@ulusofona.pt
Lusófona University/HEI-Lab: Digital Human-Environment Interaction Lab
Lisbon, Portugal

ABSTRACT

This paper proposes an empirical framework that aims to classify difficulty according to the player's physiological response. As part of the experimental protocol, a simple puzzle-based Virtual Reality (VR) videogame with three levels of difficulty was developed, each targeting a distinct region of the valence-arousal space. A study involving 32 participants was conducted, during which physiological responses (EDA, ECG, Respiration), were measured alongside emotional ratings, which were self-assessed using the Self-Assessment Manikin (SAM) during gameplay. Statistical analysis of the self-reports verified the effectiveness of the three levels in eliciting different emotions. Furthermore, classification using a Support Vector Machine (SVM) was performed to predict difficulty considering the physiological responses associated with each level. Results report an overall F1-score of 74.05% in detecting the three levels of difficulty, which validates the adopted methodology and encourages further research with a larger dataset.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; • **Computing methodologies** → *Supervised learning by classification.*

KEYWORDS

Affective computing, emotion assessment, multimodal dataset, virtual reality, games

ACM Reference Format:
Pedro Rodrigues, Micaela Fonseca, and Phil Lopes. 2023. Physiological-Based Difficulty Assessment for Virtual Reality Rehabilitation Games. In *Foundations of Digital Games 2023 (FDG 2023), April 12–14, 2023, Lisbon, Portugal.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3582437.3587187

1 INTRODUCTION

In recent years, research on the topic of emotions has integrated itself into the field of Human-Computer Interaction (HCI), and has led to the emergence of the study area described as Affective Computing (AC) [13]. One common application is through the use of dynamic difficulty adjustment (DDA) in videogames, which consists of constructing systems capable of recognizing player emotional states and dynamically alter itself to either reduce frustration or boredom [9]. Unlike traditional methodologies where difficulty adjusts itself linearly [1, 2], DDA is a dynamic solution that observes play and self-regulates itself to fit the necessities of the player.

This work further explores this relationship, more precisely that between emotion and diverging skill levels. Thus, this paper proposes a difficulty recognition system, based on physiological data collected during gameplay of a Virtual Reality (VR) puzzle game for a rehabilitation context. This work focuses on the classification of three classes of difficulty using a multimodal classifier based on Support Vector Machines (SVM).

2 RELATED WORK

In the context of videogames, affective computing emerges as a promising area of research to further enhance the player experience [16]. Videogames can elicit a wide range of affective states, however knowing all of them is often not necessary to evaluate a player's emotional experience. In fact, in the literature, the relationship between videogame difficulty and player emotions is usually described through the theory of flow [4, 9, 10, 14]. According to these models, strong involvement in the game occurs when the abilities of the player match the difficulty of the tasks. In response to the players' emotions and competence, the game would adjust the challenge so that it would be neither insufficiently nor excessively challenging and, thus, the most engaging possible [10].

This balancing act has to be continuous, as the player's skill level should naturally increase as they continue playing. For this, the automatic assessment of player emotion can be invaluable. Based on this argument, the present work developed an approach similar to that of [10], where three emotional states of interest were defined, each one corresponding to a different region of the valence and arousal space.

This work distinguishes itself from previous work by focusing on an unexplored task from those previously explored in the literature, whereas the task relates specifically to a rehabilitation exercise.

FDG 2023, April 12–14, 2023, Lisbon, Portugal

ACM ISBN 978-1-4503-9855-8/23/04.
https://doi.org/10.1145/3582437.3587187

More specifically, the contextual nature of this task is exploring the concept of DDA transferred into a therapeutic setting, allowing for the therapy itself to adapt (or as a tool for therapists), and optimize its effectiveness while maintaining patient motivation. Thus, the reason why the Trail Making Test (TMT) [3] served as the core task of this study.

3 METHODOLOGY

This study was approved by the Ethics and Deontology Committee for Scientific Research of the Lusófona University.

3.1 The Game: Wandering Druid

An exercise based on the TMT was added to the Wandering Druid (WD)[1], a VR Game for motor rehabilitation. By adding the TMT exercise to the WD game, players can perform the exercise in a way that feels more real and engaging. This can improve learning outcomes and motivation, and also provide more objective assessment of skills and behaviors [6].

In the WD players are prompted to connect a series of numeric and alphabetic dots in ascending order, while alternating between the two. Difficulty is directly proportional to the number of points on screen. Three difficulty levels were created (i.e., Easy, Medium, and Hard), comprising of a set of sequences that the player was required to complete in succession. To pin-point the difficulty "distance" between each level, a series of play-tests were conducted measuring the average completion time of a total of 15 participants between the ages of 20 and 35. Each participant played all three levels in succession, with each level consisting of 5 sequences. Considering player feedback and collected data each level was re-adjusted and fine-tuned for the following data collection task.

3.2 Experimental Settings

A Biosignal Plux[2], with a total of 3 channels, was used to collect eletrodermal activity (EDA), electrocardiogram (ECG) and respiratory activity (RSP). The VR headset consisted of a HTC Vive Pro Eye, where audio was turned off. All experiments were conducted in a sound-proof room without external distractions.

3.3 Acquisition protocol

The start of the experiment consists of consent form and demographics survey, in which during the participants are explained the task. Once the headset is placed participants play a tutorial demonstrating the core mechanics of the game, and subsequently play each of the three levels. Each level begins after a one-minute break, allowing players to rest. To assess the success of the emotional elicitation process and establish ground truth, participants were asked, after being exposed to each level, to self-annotate their emotional state, using the Self-Assessment Manikin (SAM) [7]. Furthermore, each participant was also requested to self-annotate the perceived difficulty experienced in each level using an ad-hoc questionnaire. Lastly, the Fatigue Assessment Questionnaire [12] was used to report their level of fatigue at that time.

Table 1: List of extracted features from each of the signals. Abbreviations: STD = standard deviation, MAD = median absolute deviation.

	Feature	Description
ECG	μ_{HR}, σ_{HR}	Mean, STD of HR
	μ_{HRV}, σ_{HRV}	Mean, STD of HRV
	rms_{HRV}	Root mean square of HRV
	SDSD	STD of successive differences between RR intervals
	$CVNN$	σ_{HRV} divided by μ_{HRV}
	$CVSD$	rms_{HRV} divided by μ_{HRV}
	M_{HRV}, MAD_{HRV}	Median, MAD of HRV
	$MCVNN$	MAD_{HRV} divided by M_{HRV}
	IQR_{HRV}	Interquartile range of HRV
	pNN20, pNN50	Percentage of HRV intervals differing more than 20 and 50ms
	$f^{x}_{x \in LF,HF}$	Energy in the low and high frequency component of the HRV
	LF_{norm}, HF_{norm}	Normalised LF and HF components
	$f^{LF/HF}_{HRV}$	Ratio of LF and HF components
EDA	μ_{SCL}, σ_{SCL}	Mean, STD of SCL
	M_{SCL}, MAD_{SCL}	Median, MAD of SCL
	$\int_{SCL}$	Area under the SCL curve
	μ_{SCR}, σ_{SCR}	Mean, STD of SCR
	M_{SCR}, MAD_{SCR}	Median, MAD of SCR
	$\int_{SCR}$	Area under the identified SCRs
	NP	Number of SCR peaks
	μ_{RET}	Mean rise time of SCRs
	μ_{RIT}	Mean recovery time of SCRs
	max_{SCR}	Maximum SCR amplitude
RSP	$\mu_x, \sigma_x, M_x, MAD_x$	Mean, STD, median, MAD
	$P80_x$	and percentile 80% of Inhalation (ID), exhalation (ED)
	$x \in [ID, ED, IE]$	duration and there ratio (IE)
	$\mu_x, \sigma_x, M_x, MAD_x$	Mean, STD, median, MAD
	$P80_x$	and percentile 80% of stretch (R) and first order
	$x \in [R_{RSP}, FDE_{RSP}]$	differences (FDE)
	RR	Respiration Rate
	$f^{x}_{x \in LF,HF}$	Energy in the low and high frequency component of the RSP
	LF_{norm}, HF_{norm}	Normalised LF and HF components
	$f^{LF/HF}_{RSP}$	Ratio of LF and HF components

3.4 Feature Extraction and Classification

After pre-processing, the features extracted from the different modalities are displayed in table 1.

To maximize the number of training instances, this work adopted a partitioning of 80:20, for training and testing respectively. Cross-validation (CV) was used for the tuning of the SVM hyperparameters for the linear and radial basis function (RBF) kernels. The purpose of which was to produce models with stronger generalization capacities [15]. As such, the selection of the best performing hyperparameters was achieved through an exhaustive search of several combinations of parameters. A total of 15 folds (5-Fold CV repeated over 3 random trials) were computed for each kernel, for testing the different combinations over the following set:

- C (regularization) parameter: $[10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$
- γ (Gamma) parameter, specific to the RBF Kernel: $[10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}, 10^{5}, 10^{6}, 10^{7}]$

Finally, the linear and RBF models determined as optimal were then fitted to the whole training set and tested on the remaining samples. The resulting predictions were used to compute the overall accuracy and F1-scores for each kernel.

4 EXPERIMENTAL RESULTS

4.1 Participants

Only healthy subjects with no history of psychological or neurological conditions were admitted. No participant reported suffering from any cardio-respiratory disease, as well as hyperhidrosis. A

[1] Gameplay: https://www.youtube.com/watch?v=zwi1RnEuCBc
[2] Explorer Kit: https://www.pluxbiosignals.com/collections/biosignalsplux/products/copy-of-explorer

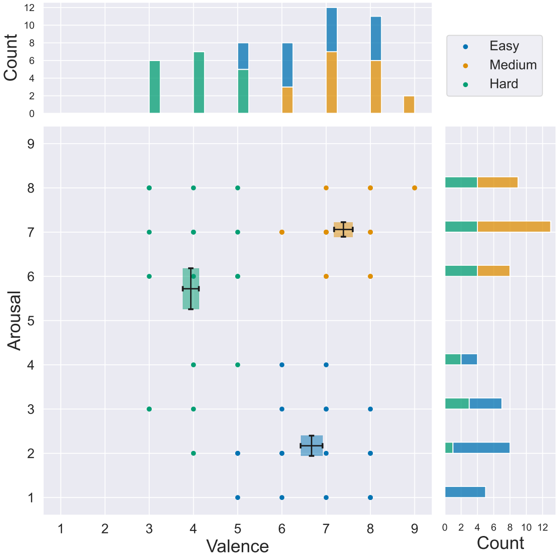

Figure 1: Self-Assessment of 18 participants for Easy (Blue), Medium (Yellow) and Hard (Green) reported on a 9-points scale mapped to the valence-arousal space.

total of 32 individuals (34.4% female) aged between 18-30 years (22.5±1.8 years old) participated in this study.

Preliminary assessment revealed that all players were able to identify the respective difficulty assigned to each level without any prior information. We excluded 7 participants from the data due to high fatigue levels and abnormal EDA signal noise. The experiment was conducted in summer which may have caused more noise in the EDA signal due to higher sweat activity.

4.2 Emotion Elicitation Results

Preliminary data analysis revealed the presence of an additional 7 outliers, whose responses deviated significantly from the rest of the participants. These outliers were identified as introducing noise into the data set and were subsequently removed to maintain data integrity and accuracy. This step was crucial as it improved the overall quality and reliability of the data, resulting in 18 remaining participants' data being used for further analysis and more accurate conclusions.

Figure 1, showcases the self-reported ratings reported for the easy, medium or hard levels. Each level is described by a box, whose center (C_{level}) is determined by computing the mean of the reported valence (M_V) and arousal (M_A) values. The dimensions of each box are bounded by the standard deviation estimated using bootstrap on the ratings given for each emotional dimension. By taking advantage of dimensional nature of the emotion annotation items, self-reports were categorized into 4 groups: positive valence (>5) and low arousal (<5) (PVLA), positive valence and high arousal (>5) (PVHA), negative valence (<5) and low arousal (NVLA), and negative valence (<5) and high arousal (NVHA). Thus, observing the following:

- The ratings reported for the easy difficulty level are concentrated in the lower right region of the valence and arousal space (PVLA) - $C_{easy}(M_V = 6.67 \pm 0.25,\ M_A = 2.17 \pm 0.23)$.
- The upper right region of the valence and arousal space, is where most of the ratings reported for the medium difficulty level are concentrated (PVHA) - $C_{medium}(M_V = 7.39 \pm 0.21,\ M_A = 7.06 \pm 0.17)$.
- The ratings reported for the hard difficulty level are divided between the upper left, and the lower left regions of the valence-arousal space (NVHA and NVLA) - $C_{hard}(M_V = 3.94 \pm 0.18,\ M_A = 5.72 \pm 0.47)$.

Results also showcased how each dimension tends to vary based on difficulty, where an increase in arousal was consistently observed when players switched from the easy to the medium difficulty. Additionally, when comparing the medium and hard it is observed that the mean value for both dimensions decreases.

A pairwise T-test was used to evaluate the existence of significant differences between group pairs, where statistical significance consists of $\alpha = 0.05$ and corrected for multiple comparisons, using a Bonferroni correction [11]. The results are summarized as follows:

- Easy-Medium: The T-test revealed that the average of arousal for the medium level was significantly higher compared to the easy level ($T = -16.84,\ p_{corrected} << 0.016$), suggesting that medium levels were more arousing than the easy one. Applying the same test to valence data yielded a similar conclusion ($T = -2.72,\ p_{corrected} = 0.007$).
- Medium-Hard: The averages for arousal and valence were both significantly higher (valence, $T = 12.17,\ p_{corrected} << 0.016$; arousal, ($T = 3.23,\ p_{corrected} = 0.007$)), suggesting that the medium level provided a more arousing and positive experience than the hard level.
- Easy-Hard: The comparison of the arousal averages for the hard and easy levels revealed that the first was significantly higher than the second ($T = 7.21,\ p_{corrected} << 0.0016$), suggesting that easy levels were less arousing than hard. Contrarily valence gave an opposite observation with valence averages indicating the easy level provided a more enjoyable experience than the hard level ($T = 7.38,\ p_{corrected} << 0.016$).

4.2.1 *Summary Results and Considerations.*

Self-reports indicated that the game was effective in eliciting different emotions for each of the three levels, however only the emotional responses reported for the medium level coincide with the targeted emotions of the level (High Valence and Arousal). For the easy level, participant ratings tended to indicate a state of relaxation contrarily to the common statement from literature (i.e. boredom). On the other hand, for the hard level there was a divide amongst participants with reports suggesting a mix of high-low arousal states.

4.3 Classification Results

To assess the physiological responses and their relation to the user reported affect an SVM-based classifier was built. Features were extracted by windowing the signal as a means of improving the overall performance of the SVM classifier. This method increases

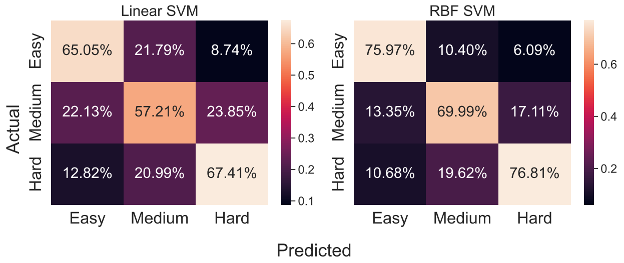

Figure 2: Confusion matrices obtained for the Linear SVM and RBF SVM. Values were normalized by the predicted values, so that the diagonal of the matrices provides information about the precision of both models.

the number of feature vectors for training and testing. Internal testing also suggested this method as superior to the non-windowing method. Classification scenarios are repeated 30 times to calculate both mean and standard deviation relative to the accuracy and f1-score. The Linear and RBF SVM variants obtained an accuracy of 63.24 ± 5.49 and 74.22 ± 6.01, and an F1 of 62.89 ± 5.4 and 74.05 ± 6.03, respectively.

Figure 2, showcases the confusion matrix for both Linear and RBF variants. Results showcase that the model had a more difficult time predicting the medium difficulty class, compared to other classes. This is inline with the results observed in [10], highlighting the variability between players regarding the notion of flow or engagement.

Furthermore, the lack of precision from the model also invalidates any firm conclusion about the relationship between emotions and difficulty. The low performance in each of the three classes can be attributed to the context of high dimensionality and low sample size associated with both classification tasks, as noted in [5]. It is thus believed that the existence of a larger dataset could yield better results.

5 CONCLUSIONS AND FUTURE WORK

This study created a virtual reality puzzle game with three difficulty levels to assess the difficulty using physiological data and a SVM algorithm. The game was added to an existing VR game and tested with 32 participants who provided physiological and self-report data for each level played.

Results obtained from statistical analysis indicate that playing the WD at different difficulty levels gave rise to different emotional states. The easy level was related to a state of positive valence, and low arousal. In comparison, the medium level was regarded as a more arousing and positive experience than the easy level. Finally, for the hard level, participants reported the experience as negative and less arousing than the medium level, although compared to the easy level it was more exciting. The results indicate that despite the easy and hard levels of the game not being able to elicit the emotions initially targeted, the adopted protocol was successful in eliciting different emotions for each level, thus validating the usability of the WD as a tool to explore players' emotions.

The automatic detection of the three levels of difficulty through the peripheral signals recorded during each of the conditions was analyzed for different SVM-based classifiers. The results obtained indicate that the RBF SVM (F1-score = 74.05%) is more suitable for the prediction of the three levels of difficulty, compared to the Linear SVM (F1-score = 62.89%), however both models share a lower detection rate of the medium difficulty, when compared to the other classes. Since different players may have different ways of approaching a game, this can cause variability in their emotional and physiological responses even when playing at the same difficulty level. For instance, some players like to be challenged slightly beyond their abilities, while others prefer a more balanced or easier task. In the future, researchers should use a mix of established and innovative methods, like the GEQ [8] and physiological measures, to capture the various differences in players' experiences. This will provide a better understanding of the relationship between flow and player experience and how it affects the classification task.

Regarding the classification task, future work should focus on testing the classification performance of other supervised learning methods such as Random Forests. Moreover, techniques for dimensionality reduction or feature selection should be integrated into the classification workflow and their effect explored on the overall performance of the predictive task. Finally, increasing the dataset will be necessary for future work to yield better results.

ACKNOWLEDGMENTS

This work is supported by Fundação para a Ciência e Tecnologia (FCT), under HEI-Lab R&D Unit (UIDB/05380/2020) and Project PlayersAll: media agency and empowerment (EXPL/COM-OUT/0882/2021).

REFERENCES

[1] Justin T. Alexander, John Sear, and Andreas Oikonomou. 2013. An investigation of the effects of game difficulty on player enjoyment. *Entertainment Computing* 4, 1, 53–62.

[2] Maria-Virginia Aponte et al. 2009. Scaling the Level of Difficulty in Single Player Video Games. In *Entertainment Computing – ICEC 2009*. Springer Berlin Heidelberg, Berlin, Heidelberg, 24–35.

[3] James A Arnett and Seth S Labovitz. 1995. Effect of physical layout in performance of the Trail Making Test. , 220–221 pages.

[4] Lawrence A Beck. 1992. Csikszentmihalyi, Mihaly. (1990). Flow: The psychology of optimal experience. , 93–94 pages.

[5] R Bellman. 1966. Dynamic programming. , 34–37 pages.

[6] Doug A. Bowman and Ryan P. McMahan. 2007. Virtual Reality: How Much Immersion Is Enough? *Computer* 40, 7 (2007), 36–43.

[7] M M Bradley and P J Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential". , 49–59 pages.

[8] Jeanne H. Brockmyer, Christine M. Fox, Kathleen A. Curtiss, Evan McBroom, Kimberly M. Burkhart, and Jacquelyn N. Pidruzny. 2009. The development of the Game Engagement Questionnaire: A measure of engagement in video game-playing. *Journal of Experimental Social Psychology* 45, 4 (2009), 624–634.

[9] Guillaume Chanel and Phil Lopes. 2020. User Evaluation of Affective Dynamic Difficulty Adjustment Based on Physiological Deep Learning. , 3–23 pages.

[10] G Chanel, C Rebetez, M Bétrancourt, and T Pun. 2011. Emotion assessment from physiological signals for adaptation of game difficulty. , 1052–1063 pages.

[11] Olive Jean Dunn. 1961. Multiple comparisons among means. , 52–64 pages.

[12] D Micklewright, A St Clair Gibson, V Gladwell, and A Al Salman. 2017. Development and validity of the rating-of-fatigue scale. , 2375–2393 pages.

[13] Rosalind W. Picard. 1997. *Affective Computing*. MIT Press, Cambridge, MA.

[14] Lorcan Reidy, Dennis Chan, Charles Nduka, and Hatice Gunes. 2020. Facial electromyography-based adaptive virtual reality gaming for cognitive training. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. ACM, New York, NY, USA.

[15] J D Rodriguez, A Perez, and J A Lozano. 2010. Sensitivity analysis of k-fold cross validation in prediction error estimation. , 569–575 pages.

[16] Georgios N. Yannakakis and Julian Togelius. 2015. Experience-driven procedural content generation. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 519–525.