A Preliminary Assessment of Game Event Detection in Emotional Mario Task at MediaEval 2021

Van-Tu Ninh[1], Tu-Khiem Le[1], Manh-Duy Nguyen[1],
Sinéad Smyth[2], Graham Healy[1], Cathal Gurrin[1]
[1]School of Computing, Dublin City University, Ireland
[2]School of Psychology, Dublin City University, Ireland
tu.ninhvan@adaptcentre.ie,tukhiem.le4@mail.dcu.ie,manh.nguyen5@mail.dcu.ie,sinead.smyth@dcu.ie,graham.healy@dcu.ie,cathal.gurrin@dcu.ie

ABSTRACT

The Emotional Mario task at MediaEval 2021 presents a new challenge of analysing the gameplay of ten participants on the well-known Super Mario Bros video game by detecting key events using facial and biometrics data. Our purpose in this work is to evaluate the application of emotion-related features in other domains of affective computing in game event detection. In this working notes paper, we present our work on in-game event detection using the conventional Random Forest model with a combination of Blood Volume Pulse and Electrodermal Activity statistical features with the facial expressions of the player as the input. In addition, we also investigate the evaluation of using the in-game visual features in another pipeline with the same Random Forest model to compare the efficiency of using in-game visual features in the model. The source code of our work can be found at https://github.com/nvtu/Emotional-Mario-Analysis.

1 INTRODUCTION

Being referred to as engines of experience, games act as a source of external stimuli that can trigger responses in human emotion (e.g., a person might feel intense stress when fighting against a boss in a game). However, the connection between games and human's emotions has not been comprehensively studied, which presents an open area of research. Therefore, the Emotional Mario Task was initiated to analyse this relationship [5]. The task employed 10 volunteers to play various stages in the Super Mario Bros video game and capture their reactions using a webcam and an E4 wristband. The ultimate goal is to (1) predict five key events in the game, and (2) summarise the gameplay by aggregating the best moments in the game. In this work, we focus mainly on the first task. Our aim is to analyse the contribution of facial expressions and physiological signals recorded from wearable devices to the detection and classification of five key events in the game.

2 APPROACH

2.1 Data Processing and Feature Extraction

2.1.1 Face, Game frame, and Sensor Synchronization and Processing: There are three types of data in the dataset captured using different devices with different sampling rates, which are: face video, in-game video and sensor data. Apart from the data-synchronisation codes given by the task organisers, we also modify the source code in the Github repository provided in [10] to extract all relevant frames corresponding to the actions in the game. For sensor data recorded from Empatica E4 device, the Blood Volume Pulse (BVP) and Accelerometer are pruned to 60 Hz from the original sampling rate of 64 Hz and 32 Hz respectively, to match the sampling rate of the video. For facial data, the Face Emotion Recognition (FER) features provided by the task organisers [5] extracted using the FER package [3] are inputted as a 7-dimensional vector into the model for training. Even though the use of in-game video is not recommended in this task, we also extract game-frame deep features from a ResNet-50 model pre-trained on the ImageNet dataset. These deep features are the same as the ones used in the preliminary work on the same dataset in [10], which is a 2048-dimensional vector.

2.1.2 Blood Volume Pulse (BVP). For Blood Volume Pulse (BVP) feature extraction, we extract statistical features commonly used for stress detection and emotion recognition using physiological signals. We use the Neurokit2[1] library, which employs the Elgandi processing pipleline to clean the photoplethysmogram (PPG) signal [6] and detect systolic peaks [2]. We then compute heart rate (HR), time-domain and frequency-domain of heart rate variability (HRV) using the extracted systolic peaks with a window-size of 60 seconds. For frequency-domain HRV features, the same parameters of low (LF: 0.04-0.15 Hz) and high (HF: 0.15-0.4 Hz) frequency bands as in [9] are used. Finally, the feature vector is standardised. This feature extraction process results in a 27-dimensional vector.

2.1.3 Electrodermal Activity (EDA):. We followed previous research [7] in stress detection analysis to extract statistical EDA features. Using Neurokit2 library, we extract components of the EDA signal that comprise Skin Conductance Response (SCR), Skin Conductance Level (SCL), SCR Peaks, SCR Onsets, and SCR Amplitude. Then, the statistical EDA features from the combination of four works [1, 4, 8, 9] are computed except for the slope of EDA signal along the time-axis, which results in a 35-dimensional vector. Finally, the feature vector is standardised.

2.2 Game Event Detection Models

In total, we develop two models whose names are A and B, respectively. Model A, which detects game events based on different combinations of emotion-related features, comprises of two stages.

As illustrated in Figure 1 (black arrow), the first stage of the model aims at detecting if a game event happens at a timestamp

Copyright 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
MediaEval'21, December 13-15 2021, Online

[1]https://github.com/neuropsychology/NeuroKit

CEUR-WS.org/Vol-3181/paper47.pdf

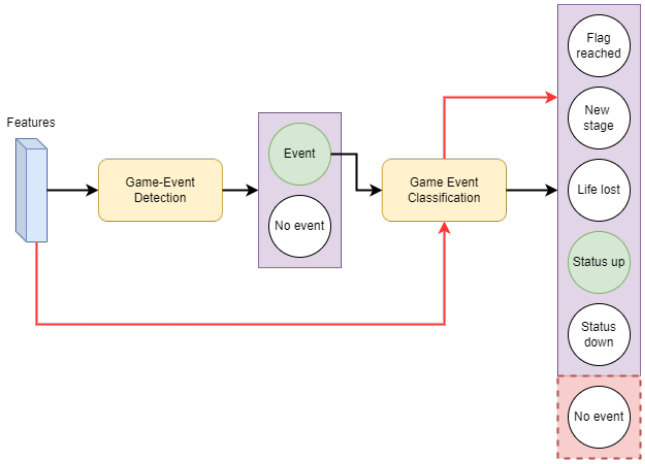

Figure 1: Overview of the game event detection model. The protocol of model A using emotion-related features for training is illustrated using black arrow. The protocol of model B using a combination of visual and biometrics features is demonstrated using red arrow. The classification result of model B consists of one additional no-event category shown in the red box.

while the second stage concentrates on classifying the corresponding game event (flag reached, life lost, status up, status down, new stage). Both stages employ a Random Forest model implemented in scikit-learn[2] and incremental trees[3] libraries with the same configuration of parameters. For the first stage training, as the number of samples of game-event/no-game-event is imbalanced which affects the learning process of the model, we shuffle the non-game-event samples, then divide them into batches whose size is equal to the one of game-event samples, and apply incremental training to the Random Forest model. The non-default parameter values that we employ in model A are shown in table 1.

Model B, which classifies game events using deep visual features extracted from game frames combined with BVP statistical features, is a simple incremental training Random Forest with the same parameter values as in Table 1 except for the number of estimators (100), minimum samples for splitting (default value), and maximum depth (default value).

Table 1: Non-default Parameter Values of Random Forest model in model A

Parameter	Value
Number of estimators	500
Minimum samples for splitting	4
Maximum depth	8
Best split max features	$\sqrt{\text{number of features}}$
Bootstrap samples	True
Out-of-bag samples	True
Class weight	balanced subsample

[2]https://scikit-learn.org
[3]https://github.com/garethjns/IncrementalTrees

Table 2: Evaluation results of our approaches compared to other teams for event timestamps in the range of +/- 5 seconds

Run	Precision	Recall	F1 score
Model A (ResNet50 + BVP)	0.3991	0.3001	0.3426
Model B (BVP + EDA)	0.0021	0.8903	0.0041
Model B (BVP + EDA + FER)	0.0019	0.7975	0.0039
GSE-AAU	0.0242	0.0812	0.0373
Random	0.2847	0.2847	0.2947

Table 3: Evaluation results of our approaches compared to other teams for matching events in the range of +/- 5 seconds

Run	Precision	Recall	F1 score
Model A (ResNet50 + BVP)	0.2068	0.1522	0.1753
Model B (BVP + EDA)	0.0014	0.5709	0.0028
Model B (BVP + EDA + FER)	0.0012	0.4998	0.0025
GSE-AAU	0.0112	0.0849	0.0197
Random	0.0667	0.0667	0.0667

3 RESULTS AND ANALYSIS

3.1 Evaluation Metrics

The organisers evaluate the runs based on exact event matching and event time-frame matching in a range of +/- one second and +/- five seconds using precision, recall, and f1 score. [5]. In our paper, we report the evaluation results of both exact event matching and time-frame matching in range of +/- five seconds.

3.2 Results

In total, we submitted three runs to the task. As described in section 2, model A is used with emotion-related features as input, while model B used additional ResNet-50 visual features of gameplay. In our prior experiment, we also tried using model B with emotion-related features as input to detect the event without success potentially due in part to the highly imbalanced nature of the dataset. The results in table 2 and 3 both show that there is a large gap in the precision of correct event detection between using emotion-related features extracted from physiological signals and using visual features from game-frame. This suggests that the game-frames contain a lot of information about the event compared to non-visual data. As demonstrated in table 2 and 3, the precision score of the model A is extremely low, while the recall score is considerably higher than other attempts in the task, which shows that the number of false positive predictions is significantly high. This means that a proper approach of event detection using emotion-related features has not been constructed successfully yet and further research on this task needs to be conducted.

ACKNOWLEDGMENTS

This publication is funded as part of Dublin City University's Research Committee and research grants from Science Foundation Ireland and co-funded by the European Regional Development Fund under grant numbers SFI/13/RC/2106, SFI/13/RC/2106_P2, SFI/12/RC/2289_P2, and 18/CRT/6223.

REFERENCES

[1] Jongyoon Choi, Beena Ahmed, and Ricardo Gutierrez-Osuna. 2011. Development and evaluation of an ambulatory stress monitor based on wearable sensors. *IEEE transactions on information technology in biomedicine* 16, 2 (2011), 279–286.

[2] Mohamed Elgendi, Ian Norton, Matt Brearley, Derek Abbott, and Dale Schuurmans. 2013. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PLoS One* 8, 10 (2013), e76585.

[3] Justin Shenk et al. 2021. Facial Expression Recognition with a deep neural network as a PyPI package. (2021). https://github.com/justinshenk/fer

[4] Jennifer Healey and Rosalind W. Picard. 2005. Detecting stress during real-world driving tasks using physiological sensors. *IEEE Transactions on Intelligent Transportation Systems* 6 (2005), 156–166.

[5] Mathias Lux, M. Riegler, Henrik Svoren, S. Hicks, Duc-Tien Dang-Nguyen, Kristine Jorgensen, Vajira Thambawita, and P. Halvorsen. 2021. Emotional Mario Task at MediaEval 2021. In *MediaEval*.

[6] Mohsen Nabian, Yu Yin, Jolie Wormwood, Karen S Quigley, Lisa F Barrett, and Sarah Ostadabbas. 2018. An open-source feature extraction tool for the analysis of peripheral physiological data. *IEEE journal of translational engineering in health and medicine* 6 (2018), 1–11.

[7] Van-Tu Ninh, Sinéad Smyth, Minh-Triet Tran, and Cathal Gurrin. 2021. Analysing the Performance of StressDetection Models on Consumer-Grade Wearable Devices. In *SoMeT*.

[8] Kizito Nkurikiyeyezu, Anna Yokokubo, and Guillaume Lopez. 2020. Effect of Person-Specific Biometrics in Improving Generic Stress Predictive Models. *Sensors and Materials* 32 (02 2020), 703–722. https://doi.org/10.18494/SAM.2020.2650

[9] Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. Introducing wesad, a multimodal dataset for wearable stress and affect detection. In *Proceedings of the 20th ACM international conference on multimodal interaction*. 400–408.

[10] Henrik Svoren, Vajira Thambawita, P. Halvorsen, Petter Jakobsen, Enrique Alejandro García Ceja, Farzan Majeed Noori, Hugo Lewi Hammer, Mathias Lux, M. Riegler, and S. Hicks. 2020. Toadstool: A Dataset for Training Emotional Intelligent Machines Playing Super Mario Bros.