Marios Fanourakis[1] and Guillaume Chanel[1]

[1]Affiliation not available

August 12, 2025

Abstract

Video games are a versatile and multi-faceted stimulus which can elicit complex player experiences. As a consequence, several datasets have been curated or created for studying human cognition, behaviours, and physiological responses where video games are the primary stimulus. Many of these datasets have a low number of participants or do not have a rich set of modalities and are always recorded in a laboratory setting. To address these issues, we have recorded 256 participants at LAN events while they played the first person shooter, Counter-Strike: Global Offensive. Our dataset consists of several complementary modalities: physiological signals (ECG, EDA, Respiration), behavioural signals (facial expressions, eyetracking, depth images, seat pressure), computer interaction (keyboard and mouse events, game actions), and stimulus information (gameplay video, game logs). We show that the number of participants in our dataset and the variety of modalities recorded is advantageous for training machine learning models.

AMuCS: Affective multimodal Counter-Strike video game dataset

Marios Fanourakis[1] **and Guillaume Chanel**[1]

[1]Social Intelligence and MultiSensing (SIMS) lab, University of Geneva, Geneva, Switzerland
*corresponding author(s): Marios Fanourakis (marios.fanourakis@unige.ch)

ABSTRACT

Video games are a versatile and multi-faceted stimulus which can elicit complex player experiences. As a consequence, several datasets have been curated or created for studying human cognition, behaviours, and physiological responses where video games are the primary stimulus. Many of these datasets have a low number of participants or do not have a rich set of modalities and are always recorded in a laboratory setting. To address these issues, we have recorded 256 participants at LAN events while they played the first person shooter, Counter-Strike: Global Offensive. Our dataset consists of several complementary modalities: physiological signals (ECG, EDA, Respiration), behavioural signals (facial expressions, eyetracking, depth images, seat pressure), computer interaction (keyboard and mouse events, game actions), and stimulus information (gameplay video, game logs). We show that the number of participants in our dataset and the variety of modalities recorded is advantageous for training machine learning models.

Background & Summary

Interactive media experiences such as video games are a versatile and multi-faceted stimulus. Video games are becoming increasingly more realistic with detailed graphics, accurate physics, convincing emotional characters, and immersive virtual reality. As such, video games elicit complex player experiences which makes them an attractive subject for the research community.

Over the past several years the collection of player data has become a serious consideration and necessity for both researchers and developers[1]. Player telemetry allows designers to observe and obtain an accurate representation of several in-game behaviours, which in turn can be used to make important game design decisions and modifications. Telemetry is also often used for matchmaking and ranking players in competitive multiplayer games[2] to provide more enjoyable experience for players at different skill ranges. However, telemetry often only measures in-game behaviours and thus fails to capture the player behaviours and reactions required for getting a full representation of the player state. Due to the ability of video games to immerse players and to elicit complex emotions, it has been widely studied in the domain of psychology and affective computing. In these domains, physiological activity, facial expression and body posture are often recorded in addition to in-game measures[3–8].

In affective computing, the focus is on the user experience with the goal of understanding user behaviours and emotions to make the playing experience more engaging. This could for instance be achieved by adapting the video game content and difficulty according to players' experience[8]. User emotions can be estimated continuously and in real-time by measuring physiological responses which are linked to the autonomous nervous system (ANS) and other behavioural information like facial expressions. A common finding is that it is necessary to perform multimodal recordings in order to capture as much as possible of the user experience[9,10].

Although the usage of physiology to study player behaviour is quite common in the literature, it is rarely done within a multiplayer setting where all players' data streams are collected synchronously. Furthermore, although multimodal data collection has been achieved, we are not aware of a study which has collected all the modalities presented in this paper in a multiplayer scenario.

To facilitate research in the fields of game analysis and psychology, several datasets have been made publicly available (see Table 1). The datasets utilize a wide variety of games as a stimulus, from 2D platformers like Super Mario Bros to 3D action adventure games like Assassin's Creed. Very few datasets use multiplayer games and even fewer collect data from all players concurrently. Multiplayer data could be useful for analyzing synchronous behaviours, interaction dynamics, and more. Typically, contextual data (e.g. game logs) is collected in a laboratory environment together with physiological signals such as electrocardiograms (ECG), electrodermal activity (EDA) and electroencephalograms (EEG).

Annotations vary between the datasets as well. They can be collected through questionnaires that are administered after the

experience and ask participants to summarize the experienced intensity of a sentiment. A common type of questionnaire for indicating emotions is the self-assessment manikin[11] (SAM) where participants are aided by a pictorial representation of the intensity of the emotions. Another type of annotation is based on a graphical element that represents a one or two-dimensional space of interest where participants can report their subjective feeling[12] (continuous-space annotation). The arousal-valence space is commonly used to annotate emotional experiences in a continuous-space. This may be accompanied by visual aids positioning categorical emotions in the space. Annotations can also be collected in a continuous-time manner for a specific sentiment[13,14]. Annotators are asked to watch a recording and at the same time use a software to indicate the absolute or relative intensity of a sentiment at each moment. These types of annotations can be used to compare the reactions to specific moments or events throughout the experience. The datasets listed in Table 1 are annotated in one or multiple of: fun, frustration, engagement, arousal, valence, and others.

Our study aimed at collecting a large affective multimodal video game dataset to identify robust emotional modalities and train emotion recognition machine learning models. The resulting dataset (AMuCS[15]) is the largest *multiplayer* dataset by an order of magnitude (245 vs 30 participants) and includes 11 recorded modalities compared to the average of 5 modalities in Table 1. It also remains competitive with single-player datasets like BIRAFFE, FUNii, and BIRAFFE2 which have a similar number of participants (206, 190, and 103 respectively). Due to its multimodal nature, this dataset can be used to analyze several facets of video gameplay including but not limited to emotional dynamics, player experience, player performance, player behaviours, and team dynamics. The AMuCS dataset[15] has previously been used to study pupil diameter[16].

Methods

We recorded a total of 256 participants playing video games in realistic conditions (7.66% female, 0.81% non-binary, 0.40% no answer). The participants mean age was 22.68 years old (standard deviation of 3.71, ranging from 18 to 36 years old). The languages spoken by the participants varied between Swiss German, French, and English. The data was collected in 71 experimental sessions where groups of 2 or 4 participants played the *Counter-Strike: Global Offensive (CS:GO)* first person shooter (FPS) video game on a computer. Several modalities were recorded during the game using custom data acquisition software modules:

All data was synchronized using the Lab Streaming Layer (LSL) software library[17]. More information about the data acquisition software and architecture can be found in our relevant publication[18] where we measured the synchronization delays to be within 50ms on average after offset corrections. Video frames were not sent over LSL to reduce the bandwith usage on the local network. Face and screen videos were saved directly on the participant's PC and video frame numbers and timestamps were sent to LSL for synchronisation.

The data was collected on-site at several video game LAN events in Switzerland over the course of two years: SwitzerLAN 2020 (`https://switzerlan.ch/`), SwitzerLAN 2021, PolyLAN 36 (`https://polylan.ch/`) , and PolyLAN 37. The experimental area was setup in an approximately 5 square meter area within the event. It included 4 gaming PCs, each equipped with the sensors mentioned earlier, and 1 server PC where the game server and LSL Lab Recorder were installed.

The SwitzerLAN 2020 event welcomed approximately 300 gamers in the BernExpo exposition area, a large open space in the city of Bern, Switzerland. This event took place in October of 2020 during the COVID 19 pandemic and various restrictions were in place such as wearing masks when not seated and temperature checks upon entering the LAN area. The physical setup of the experiment is illustrated in Figure 1a. The experiment was in a corner of the same area as the LAN, consequently, the environmental lighting and noise was not controlled. Furthermore, the participants were informed that they could keep wearing their mask for the experiment. 14 sessions totaling 41 participants were recorded at this event (sessions numbered 1 to 14).

The SwitzerLAN 2021 event welcomed approximately 1200 gamers in the BernExpo exposition area. This event took place in October 2021 and gamers had to show proof of vaccination, recovery, or a PCR test before entering the LAN area. The physical setup of the experiment was similar to the SwitzerLAN 2020 setup. Mask-wearing was not enforced in this event. 18 sessions totaling 67 participants were recorded at this event (sessions 15 to 33).

The PolyLAN 36 event welcomed approximately 200 gamers in a large auditorium of the EPFL campus in Ecublens, Switzerland. This event took place in November of 2021 and had similar restrictions as the SwitzerLAN 2021 event. The physical setup of the experiment was slightly different in this event and is illustrated in Figure 1b. The experiment was located at the last row of the auditorium. As with the other events, the environmental lighting and noise was not controlled. 7 sessions totaling 32 participants were recorded at this event (sessions 34 to 41).

The PolyLAN 37 event welcomed approximately 1200 gamers in the SwissTech Convention Center in the EPFL campus. This event took place in April 2022 and had similar restrictions as the previous two events. The physical setup is illustrated in Figure 1a. The experiment was located in an office adjacent to the LAN area, this resulted in much less environmental noise from the rest of the LAN as well as consistent lighting. 29 sessions totaling 116 participants were recorded at this event (sessions 42 to 71).

The ethics board of the University of Geneva (CUREG) approved the conduct of the study and sharing of the data under the project title "Emotionally intelligent peripherals for video game streamers and players - video annotations" (IRB committee number CUREG-2021-06-63) and conforms to all ethical guidelines set forth by the institution. All participants provided informed consent prior to their inclusion in the study. 245 participants (out of 256 recorded) accepted to share their *anonymized* data with other research institutions.

The *Counter-Strike: Global Offensive* game

In our experiments we utilized Valve's Counter-Strike: Global Offensive (CS:GO). It is a free and modable multiplayer first person shooter (FPS) developed in the Half-Life 2 game engine. It is also popular in the e-sport community. The game includes several game modes: demolition, hostage, deathmatch, and team deathmatch.

The experiment used the *team deathmatch* game mode where two teams try to eliminate each other. Players started with a 2 minute warmup round, where they could explore the game map and test their opponents without counting the score. After the warmup, players were respawned to random locations and frozen in place for 1 minute. This period could be used to establish a baseline of physiological activity. Once the freeze time was over, the main round round started and had a duration of 10 minutes. Each player started with 100 health points and 100 armor points and were randomly placed on the game map. They were equipped with a random set of weapons from an assortment of assault rifles, long range rifles, pistols, light and heavy machine guns, and a knife. The goal of the game was to kill the players in the enemy team as many times as they can while avoiding to get killed. A player was killed once their health points reached 0, and they were subsequently revived (respawned) at a random location in the map after 2 seconds. If a player managed to get 2 kills in a row without dying they were rewarded with an item (healthshot) which restored 50 health points when used.

The game data was recorded using our custom *sourcemod*(https://www.sourcemod.net/) plugin which enabled us to send the data on an LSL stream to be synchronized with the other experimental data. OBS was also used to record the screen and sound of both the game and the participant.

Experimental protocol

Groups of 2 or 4 participants were spontaneously recruited at the LAN events to play a single round of one versus one or two versus two team deathmatch. The participants first read and signed a consent form which describes the experiment and the data that will be recorded. Immediately after, they answered a questionnaire with demographic questions (age, gender, handedness) as well as questions about their experience playing various types of video games (Brain and Learning Lab Video Game questionnaire - https://www.unige.ch/fapse/brainlearning/vgq/, their fatigue level[19], and their closeness of relationship with the other participants in the experiment group[20].

Before attaching the ECG electrodes, the participants were given cotton soaked in alcohol in order to clean the electrode locations on their body. The 3-lead ECG sensor consisted of adhesive wet electrodes and was attached in the Einthoven triangle pattern. Specifically, the positive and negative leads were attached half way between the shoulder and the sternum on the left and right collar bone respectively, and the reference lead was attached on the right side just below the rib cage (similar to Figure 2a). The 2-lead EDA sensor consisted of dry electrodes secured by Velcro straps around the proximal phalanx of the index and middle finger. EDA sensors were attached to the hand which was used to manipulate the computer keyboard (similar to Figure 2b). The specific fingers and electrode locations were chosen to be the least obstructing for the participants while still having a signal of adequate quality. The "keyboard" hand does not move as much as the "mouse" hand and we found there were fewer movement artifacts at this location. The respiration belt (a stretch sensor) was attached over the shirt of the participant around their chest and over the diaphragm like in Figure 2c.

The participants were asked to sit approximately as they would be sitting during gameplay and then we used the Tobii eyetracker manager software (https://www.tobiipro.com/product-listing/eye-tracker-manager/) to adjust the screen distance and angle such that the eyetracker could reliably detect the eyes. The Tobii eyetracker was mounted on the bottom bezel of the screen. It was calibrated using a five point calibration procedure using the same software. The results were verified by asking the participant to look at various points on the screen. If there were significant errors in the gaze position, the calibration procedure was repeated.

Players adjusted their game settings to their own preference (ex. mouse sensitivity, UI elements, keyboard bindings, screen resolution and aspect ratio). They then joined the experiment's game server where they started with the 2 minute warmup round, and continued with the 1 minute baselining and then the 10 minutes main round.

After the game was finished, the sensors were removed and the players used the PAGAN tool[21] to self-annotate their own recorded gameplay video (i.e video of the screen). PAGAN is a web-based platform that can be used by researchers to easily crowdsource continuous annotations of videos. Players self-annotated their own gameplay according to the arousal **or** valence emotional dimensions using RankTrace[22], a relative and unbounded method for continuous annotations. The gameplay video served as a recall aid and we did not show the video of their own face so as not to bias the annotations towards facial expressions. Participants only annotated one of arousal or valence, not both. This was done to reduce the cognitive load of the annotation

process with the goal of producing higher quality annotations. We also chose not to have the participant annotate the video twice (once for each dimension) due to time constraints for the experiment.

All participants were compensated with 10 Swiss Francs for their participation. The player who scored the most points during the match received a gaming mouse (Logitech G305) as a prize.

Data Records

The AMuCS dataset[15] consists of 245 participants who agreed to share their data with other research institutions and is available upon request in the Yareta data archive platform. It is available to researchers for non-commercial applications under a data use agreement (DUA) with our research lab.

The dataset contains several modalities:

- Mouse/keyboard button presses - recorded at an irregular rate (as button presses occurred).
- Game data (health, armor, position, damage taken, damage received, etc.) - recorded at 64Hz using a custom game plugin.
- Gameplay video - recorded at 30Hz using *Open Broadcasting Studio (OBS)*.
- Color and depth video of the face - recorded at up to 30Hz using an *Intel RealSense D435* camera.
- Seat pressure - recorded at 10Hz with a *Sensing Tex* seat pressure mat.
- Physiological data (electrocardiogram, electrodermal activity, respiration) - recorded at 100Hz using a *Bitalino (r)evolution* device
- Eyetracker data (gaze, pupil diameter) - recorded at 60Hz using a *Tobii pro nano*.

A detailed table of the recorded data types is listed in Table 2. The gameplay video also includes the gameplay audio and the microphone recordings on the same audio track. In some instances, the microphone was not recorded due to technical issues.

Data pre-processing

The LSL LabRecorder software records data in the extensible data format (XDF). The files contain all local timestamps as well as timing information that facilitates the synchronization of the data streams. We used the Python xdf module (`https://github.com/xdf-modules/`) to read the files and automatically apply the timestamp synchronization. We then used the pandas package[23] in Python to format the data into easily queried data structures and to ultimately convert the xdf files to parquet (`https://parquet.apache.org/`) or comma separated value (CSV) files. Since the annotations of the gameplay were performed after data collection, they had to be aligned with the rest of the data. This was achieved by using the synchronized frame timing information of the gameplay that was recorded with LSL.

For convenience and ethical concerns, we derived some additional features that may be relevant and are either computationally complex or rely on data which will not be made public such as the face videos. These features include the luminance of the screen, in-game combat, a danger level indicator, and facial features like action unit (AU) activations.

Screen Luminance

To compute the perceived lightness (luminance) of each screen pixel the Lstar from CIELAB[24] was computed from the RGB values. First, the RGB values were converted from gamma encoding to linear encoding, then the standard coefficients for sRGB (0.2126, 0.7152, 0.0722 for R, G, and B respectively) were applied to compute RGB luminance. Finally, the RGB luminance was converted to the perceived lightness, Lstar, which closely matches human light perception. It is important to note that Lstar does not take the Helmholtz–Kohlrausch effect[25] into account wherein the intense saturation of spectral hue is perceived as part of the color's luminance.

Having the Lstar value for each pixel, we then averaged the Lstar pixel values within an 8 degree horizontal foveal area of the screen centered at the gaze target of the participant. We used a rectangular area instead of a circular area since it simplified our computations. This foveal area was approximately a 16cm by 9cm rectangular region on the screen with the same aspect ratio as the screen. We did this for each frame of the video recording, always centering on the gaze target at each frame using the eye-tracking data.

In the dataset, we provide the mean luminance of the entire screen, the mean luminance of the gaze region and the central region of the screen as well as the mean luminance of the screen excluding the gaze region and excluding the central region.

Note that the Lstar luminance measures the pixel activations and does not correspond to the absolute luminance as measured by a luminance meter. However, this information can still be useful when analyzing the pupil size and to attenuate the pupil light response from the pupil data as demonstrated in Fanourakis et al.[16].

Combat and other special game events

From the game event data we computed some special indicators/events such as the number of enemies that are: in the field of view, in close range ($<$ 500 game distance units), and in mid range ($<$ 1000 game distance units). We also computed a health danger indicator indicating if the health is below 70%, 50%, or 30%.

We labeled gameplay as combat if the player had received or dealt damage from/to another player within a 5 second window. Once there was a death event (either the player was killed or the enemy was killed), the combat state was reset even if it was within a 5 second window of the combat events previously mentioned. These game event combinations were selected among other recorded game events based on their relevance towards the game's two main goals: staying alive and killing the enemy. Combat is directly relevant to these two goals since the most common outcome of combat will either be a goal success (enemy killed) or a goal failure (player death).

The "health danger" indicator gives an indication of the probability of achieving or failing the goals. All else being equal, a player with lower health will be killed more quickly during combat. The "number of enemies in field of view" event puts the player in the position to seek goal resolution by taking action to stay alive and/or kill the enemy. The "number of enemies in close/mid range" can also give an indicator of the amount of danger that a player might be in. In conjunction with these combined events some other simple events can be of interest such as "reloading weapon" and "jumping". These events prevent the player from making a fight or flight decision: a player is not able to fire while reloading, and cannot take cover while jumping. This type of information can be useful when analyzing the game context and summarizing the various events into different phases of the game like in Weber et al.[26].

Face features

We applied Baltrusaitis' OpenFace[27] feature extraction on the color video of the face to extract the following features: gaze, facial landmarks, head pose, and continuous facial action unit activations. Although gaze is tracked by the Tobii eyetracker, this additional gaze estimate can be useful in the rare cases when the eyetracker failed to track the participant's eyes due to bad placement or movement outside of the eyetracker's operating range.

Technical Validation

In this section we will show the benefits of the large number of modalities and participants in our dataset. We will see how the number of modalities and number of participants in the training set influences machine learning prediction performance. This also gives a baseline of the predictive capacity of our dataset. Note that we did not aim to achieve competitive results and used classical machine learning methods such as gradient boosting. We used the f1 score and Cohen's Kappa to measure the performance of classifiers and the concordance correlation coefficient (CCC) for regressors.

Multimodal prediction of game events

FPS games are typically fast paced with several different game events happening in rapid succession and in bursts. These game events can elicit physiological and behavioural responses from the players. This relationship allows us to use our dataset to predict game events from physiological signals.

The main challenge is that the physiological responses are not at the same cadence as game events. The elicited fluctuations in physiology through the ANS are generally at much lower frequencies (generally below $0.5Hz$)[28,29] than game events from a fast paced game (bursts of events can be well above $2Hz$ in our dataset). To facilitate the process we may analyze the game events to derive general game states. Weber et al. did so by defining several game phases based on game micro-events for the game "Tactical Ops: Assault on Terror"[26]. They defined a total of 6 phases: use of in-game menu, safe, danger (enemy in field of view), combat (player uses weapon), under attack, ghost mode (player has died and is viewing the arena in 3rd person). They then defined events in the context of these phases and found significant differences of heart rate responses to these game events.

We will proceed in a similar strategy but define only two phases: safe and danger. We added some complexity to the detection of the danger phase of Weber et al. by not only taking into account the enemies in the field of view but also the distance to the enemy, the current health of the player, and if the player is reloading their weapon or jumping. We also merged combat, and under attack phases of Weber et al. with the danger phase. We discarded the ghost phase since this was not enabled in our game, and we also discarded the use of in-game menu phase since the menu was used relatively rarely. Details on how we computed combat and danger can be found in the previous Section.

We wanted to verify that there is a perceived difference of emotional arousal between the game phases so we compared the participants' arousal annotations during portions of the game when the players were safe versus when they were in danger according to the previously computed game phases. The results are shown in Figure 3 where the mean arousal (z-score) is -0.36 and 0.12 during the safe phase and danger phase respectively. A one-sided Mann-Whitney U test shows that the increase of the arousal annotation values during the danger phase is statistically significant (p-value smaller than 0.001).

The physiological modalities that we used for predicting the two phases were: the EDA, heart rate (HR), respiration, facial action units, on-screen gaze speed. The modalities were normalized per participant with a z-score. We used the signals of 121 participants who had usable data for these modalities.

A low pass filter with a cutoff of $5Hz$ was applied to the EDA signal to remove high frequency noise. The same filter was also used for the respiration signal. The instantaneous heart rate was computed from the ECG signal after a low pass filter of $45Hz$ (to remove high frequency noise) by applying Hamilton method of R-peak detection of the BioSPPy Python package[30], then computing the peak rates and smoothing them using a boxcar smoother of length 10. The on-screen gaze speed was computed from the gaze data by measuring the distance between consecutive gaze points on the screen. This feature gives an indication of saccade and fixation behaviours without information about where on the screen the player is looking at. We extracted several features from each of these signals by using a 15 second rolling window with a step size of 10 seconds. The features computed within each window are summarized in Table 3.

We similarly windowed the game phase signals and extracted only the maximum value in the windows. Due to the fast paced nature of the game the players find themselves more frequently in the danger phase than the safe phase, resulting in imbalanced classes. Across all the windowed game phase signals and participants we had a total of 1050 instances of the safe phase class and 6762 instances of the danger phase class.

We then used a gradient boost classifier from the scikit-learn Python package[31] with leave-one-participant-out cross validation to predict the game phase from the physiological modalities. We used the default parameters of the model: log loss, learning rate of 0.1, 100 estimators, Friedman MSE criterion, maximum depth of 3. The results are summarized in Table 4 where we report the mean f1 score and Cohen's kappa of the test sets. We performed one-sided paired Wilcoxon statistical tests to determine if the increase in performance (f1 score) of the models was statistically significant with p-value smaller than 0.01 (*) or 0.001 (**). The results of the statistical tests are summarized in Table 5.

On their own, EDA, gaze speed, and the facial action units have the best performance. In addition, their multimodal fusion lead to a significant increase of performance for game phase prediction. With all the modalities together we reach an f1 score of 0.75 (Cohen's Kappa of 0.52). The HR and respiration signals perform poorly on their own. But when they are combined with other signals the performance is not affected significantly. A more diverse set of features or a different machine learning model may be able to utilize these signals more effectively. Although there is extensive literature showing that heart rate is correlated with emotional arousal, our models were not able to distinguish between the safe phase and the danger phase despite that their arousal annotations had a statistical difference. One potential reason could be that the set of features we extracted from the heart rate were not appropriate. Indeed, heart rate variability (HRV) features are more often used in the literature and the lack of such features in our case could be the reason why our models performed poorly for this modality. Another potential reason could be that the heart rate fluctuations can be induced by both the sympathetic system and the parasympathetic system, thus making the heart rate response origin uncertain between emotional stimuli or other functions. In combination with the complexity of the game stimuli, it could very well be the case that heart rate on its own is not enough to determine the game phase. In the literature, experiments showing the effects of arousal on the heart rate typically use a single unambiguous emotional stimulus and enough data is recorded (more than 40s) to capture the low frequency fluctuations ($0.05Hz$ to $0.15Hz$) of the heart rate which are related to the sympathetic nervous system. On the other hand, the EDA is directly linked to the sympathetic nervous system and emotional stimuli have a more direct influence[32]. It is also important to note that inter-participant variability in the physiological response to stimuli can have a substantial negative impact on the classification performance of leave-one-participant-out cross validation. However, it is a useful technique to explore input features which generalize well to an unseen population.

Despite these shortcomings, it is evident from our results that there are statistically significant improvements in the performance of models when including multiple diverse modalities. The multimodal aspect of this dataset can therefore be a valuable attribute which can be utilized to train state of the art models. Potential avenues to improve the classification performance include using more complex machine learning models (ex. deep neural networks), and training more specialized models. The latter could be achieved, for example, by grouping participants according to their gaming experience and training models for each group.

Prediction of emotional arousal annotations

Machine learning models often generalize better when trained with larger datasets. AMuCS[15] is the largest multimodal dataset with continuous affect annotations and we will use it to show that model performance improves significantly as we increase the amount of data that is available for training. We will focus on emotional arousal as the target for our machine learning models to validate the continuous emotional annotations at the same time. Due to the high number of iterations necessary for statistical tests we must make sure the model is simple and will only use EDA and HR as input signals which further limits the target to arousal.

We fit a gradient boost regressor in Python (using the implementation in the scikit-learn package[31]) using an increasing amount of training data N, ranging from 5 training participants to 64 training participants in increments of 10. Note that we

have access to one more participant compared to what is published and summarized in Table 8 for the relevant modalities. This is because some participants did not give their consent to share their data with other research institutions and are not included in AMuCS[15]. We used the default parameters of the model: squared error loss, learning rate of 0.1, 100 estimators, Friedman MSE criterion, maximum depth of 3. We performed leave-one-out cross validation. For each left-out participant and for each total training data size N, we randomly selected N participants from the remaining 64 participants. We then fit the regressor using this random selection, and tested on the test participant. We repeated the random selection of N participants (with replacement) to fit and test new models until the mean of the CCC between all random trials was stable (i.e. the measured mean was within ± 0.005 of the real mean with 99% confidence) or a maximum of 1000 trials was reached.

We used the EDA and HR (computed from ECG as described earlier in this section) as input and the arousal annotation as target. The input and target modalities were normalized per participant (z-score). We used a window of size 7 seconds and step size 5 seconds for extracting the features.

In Table 6 and Figure 4 we report the mean CCC of the random trials between all left-out participants for each training data size N. We observe that as we increase N, the mean CCC is increased. We performed one-sided paired Wilcoxon statistical tests to confirm that the improvement of the mean CCC as we increase N is statistically significant with a p-value smaller than 0.001 (**) except between values of $N = 35$ and $N = 45$ where the p-value is smaller than 0.01 (*).

One-sided paired Wilcoxon statistical tests show that the performance improvement as we increase the number of participants in the training set is significant. The results are summarized in Table 7.

In Figure 4 we also observe an increase in the variance between the participants' results. The reason for this is that, although the CCC increases as we increase N, it does not increase equally across the participants. To verify this, we plot the change in CCC of each participant in Figure 5. We can clearly see that the participants have very different changes in performance. Most improve (green curves), some have no statistically significant change (blue curves), and for a few, the performance decreases (red curves), this explains the increase in variance that we saw in Figure 4 even though the results improve on average as we increase N. This trend of increased overall performance and simultaneous increase in the variance can also be observed in the boxplots of the performance difference from $N = 5$ in Figure 6.

In Figure 5, we observed that for some participants (8 participants out of 65), increasing N from 5 to 55 results in a statistically significant **decrease** in performance. Upon further investigation we have found that for 6 of those participants, the performance was generally bad (CCC below 0.1). This could be caused by poor quality annotations, for example if the participant did not understand the task. The other 2 participants had an above average CCC and we have not determined the precise cause of this performance decrease.

Effect of time lag between the target and input window

In the previous section, we modified the number of participants in the training set while keeping the input features and targets time-aligned (i.e. no offset). A similar methodology was applied to compare how the performance of the gradient boost regressor changes when applying different offsets between the input and target feature windows. In this analysis the number of participants in the training set is constant ($N = 64$). As previously, we used a 7 second window to compute the EDA and HR features as well as the mean of the arousal annotations. We applied a series of offsets in the range of -5 to +7 seconds. That is, the start of the input feature window was 5 seconds before the start of the target window and up to 7 seconds after the start of the target window.

The results are illustrated in Figure 7 where we observe that the performance with the different offsets was not significantly different (one-sided paired Wilcoxon tests) from the time-aligned performance (offset = 0 seconds). A potential explanation is that the length of the feature window (7 seconds) for computing input features and the arousal mean is sufficient to include, at least in part, any delayed physiological responses to arousal changes induced by game events. Therefore applying a global offset between the windows does not have any significant effect on the performance of the regressor.

On the other hand, it may be necessary to account for the inter-participant variability of the physiological response delay by computing optimal window delays for each participant before training the regressor. This could be achieved using the cross correlation measure to determine a time-invariant delay for each participant. It may also be the case that the response delay is time-varying. Regularizing the data to account for this time-varying delay (for example, with dynamic time-warping) may lead to an improved performance of the regressor.

Data Quality

All data modalities were visually inspected for determining their quality. We grouped data into four categories: not usable, partial, usable, and good. The first level of inspection was simply to determine how much of the data was missing or outside of an acceptable signal to noise ratio (SNR) during the main round of the game (10 minutes of data). The SNR was mainly a concern with the EDA, ECG, and respiration signals. For the EDA signal, we accounted for the presence and visibility of phasic responses versus other signal artifacts. For the ECG signal, we accounted for the visibility of the QRS complex or R-peaks versus other signal artifacts. The signals was tentatively labeled as:

- good - if less than 10% of the signal was missing or outside of SNR range
- usable - if less than 33% of the signal was missing or outside of SNR range
- partial - if less than 50% of the signal was missing or outside of SNR range
- not usable - if more than 50% of the signal was missing or outside of SNR range

Next we looked more closely at the quality of some of the signals. For EDA, we accounted for the resolution of the signal. Participants' base skin conductivity varied significantly and several had intrinsically low skin conductivity. This meant that their phasic responses had very low amplitude and could be more challenging to analyze but still usable, hence we labeled such EDA signals as usable. In several cases the resolution was too low to discern any phasic responses and these were labeled as not usable. At the other extreme, there were several participants whose skin perspiration was too high leading to saturation of the EDA signals. These were also labeled as not usable. If minor artifacts were visible, such as those that may occur during keystrokes, the signal was labeled as usable.

For ECG, we accounted for the visibility of the QRS complex. If all parts of the QRS complex were visible, then we labeled the data as good, if only the R-peaks were visible we labeled the data as usable, otherwise we labeled it as not usable or partial.

For respiration, we mainly focused on artifacts from the heart rate. If these artifacts had amplitude larger than 10% of the amplitude changes caused by breathing then the signal was labeled as usable, otherwise as good.

For the face video, if the participant's face was fully visible throughout the session it was labeled as good, if the face was partially covered (ex. wearing a mask) or otherwise not fully visible it was labeled as partial.

For the arousal and valence annotations we merely indicated if the corresponding annotation fulfilled the base criteria mentioned earlier (the proportion of the data missing). To avoid introducing any bias, we did not impose any of our preconceptions on what a good annotation should look like for the corresponding gameplay. However, a visual analysis of the annotations led to the conclusion that participants labeled their gameplay with dissimilar patterns.

In Table 8 we summarize the number of usable data for some combinations of modalities.

Despite several participants having modalities of low quality or entirely missing, there remains enough data to maintain this dataset among the current largest in size. It is feasible that state of the art methods can also be employed to overcome the issues of missing modalities[33,34].

Usage Notes

The dataset can be accessed by following the DOI link
https://doi.org/10.26037/yareta:gvvoc4wfsfhupm4ygge26wupnm.
Interested parties can request access to the data archive on the Yareta platform where they will be asked to complete a data use agreement (DUA), sign it, and submit it along with the request. The request can then be processed for approval. At the time of writing, the Yareta platform requires a Switch edu-ID account (https://www.switch.ch/en/edu-id) for authentication, however, other authentication methods may be available at a future time. Although the Switch edu-ID is primarily used by Swiss universities, an account can be created by anyone irrespective of their affiliation to Swiss academic institutions (https://eduid.ch/registration). The full public dataset, which includes video screen captures and depth videos, amounts to a total of 955GB of data and can only be downloaded in full.

Python version 3.10 was used to process and analyze the data. We made use of the following Python packages: NumPy[35], Pandas[23], scikit-learn[31], SciPy[36], BioSPPy[30], NeuroKit2[37], and PyTorch[38].

In the documentation directory of the dataset archive we provide a file containing detailed descriptions of each gamedata stream. In the same directory, we also provide a file indicating the quality of each modality for every participant, we recommend that it is taken into consideration when analyzing the data.

In the Python script directory of the dataset archive we provide a script which can be used to merge all the data modalities into a single pandas dataframe synchronized to the timestamps of a modality of choice. It may need to be adapted to meet individual needs but can be used as a starting point. We also provide a Python script example for reading video frames from the gameplay videos or the depth videos.

The units of the recorded EDA data cannot be converted to micro Siemens (μS). Due to the EDA sensor's limited sensing range we made a hardware modification to the sensor which enabled us to manually adjust the amplification gain as needed and measure conductivity values that were beyond the design specifications of the sensor albeit at a reduced precision. The gain parameters are thus different for each participant and do not match the manufacturer specifications. The analysis of EDA data should focus on its dynamics rather than its absolute values and normalization of the data for each individual participant is recommended.

Caution should be used when analyzing the pupil diameter since the environmental and screen luminance were not controlled. Certain game events which may induce psychosensory pupil responses also produce distinctive luminance changes on the screen. Thus, the pupil light response is the main driver of pupil diameter changes in our experiment. Attempts to attenuate the pupil light response from the data were somewhat successful but luminance artifacts remained[16].

Code availability

The code related to the experimental setup can be found in https://gitlab.unige.ch/sims/esports-data-platform. The individual data acquisition modules can be found in https://gitlab.unige.ch/sims/lsl-modules. Various scripts and packages for analyzing the data can be found in https://gitlab.unige.ch/Guillaume.Chanel/e-sport-ml and access is available upon request. Some basic scripts for reading the data are already included in the dataset archive.

References

1. Drachen, A., Canossa, A. & Yannakakis, G. N. Player modeling using self-organization in Tomb Raider: Underworld. In *2009 IEEE Symposium on Computational Intelligence and Games*, 1–8, 10.1109/CIG.2009.5286500 (IEEE, 2009).

2. Delalleau, O. *et al.* Beyond Skill Rating: Advanced Matchmaking in Ghost Recon Online. *IEEE Transactions on Comput. Intell. AI Games* **4**, 167–177, 10.1109/TCIAIG.2012.2188833 (2012).

3. Kivikangas, J. M. *et al.* A review of the use of psychophysiological methods in game research. *J. Gaming & Virtual Worlds* **3**, 181–199, 10.1386/jgvw.3.3.181_1 (2011).

4. Christy, T. & Kuncheva, L. I. Technological Advancements in Affective Gaming: A Historical Survey. *GSTF J. on Comput. (JoC)* **3**, 38, 10.7603/s40601-013-0038-5 (2014).

5. Yannakakis, G. N., Martinez, H. P. & Garbarino, M. Psychophysiology in Games. In *Emotion in Games*, 119–137, 10.1007/978-3-319-41316-7_7 (Springer, 2016).

6. Clerico, A. *et al.* Biometrics and classifier fusion to predict the fun-factor in video gaming. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, 1–8, 10.1109/CIG.2016.7860418 (IEEE, 2016).

7. Aranha, R. V., Correa, C. G. & Nunes, F. L. S. Adapting software with Affective Computing: a systematic review. *IEEE Transactions on Affect. Comput.* **3045**, 1–1, 10.1109/TAFFC.2019.2902379 (2019).

8. Chanel, G. & Lopes, P. User Evaluation of Affective Dynamic Difficulty Adjustment Based on Physiological Deep Learning. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10.1007/978-3-030-50353-6_1 (2020).

9. Gilroy, S. W., Cavazza, M. O. & Vervondel, V. Evaluating multimodal affective fusion using physiological signals. In *Proceedings of the 16th International Conference on Intelligent User Interfaces*, IUI '11, 53–62, 10.1145/1943403.1943413 (Association for Computing Machinery, New York, NY, USA, 2011).

10. Baig, M. Z. & Kavakli, M. A survey on psycho-physiological analysis & measurement methods in multimodal systems. *Multimodal Technol. Interact.* **3**, 10.3390/mti3020037 (2019).

11. Bradley, M. M. & Lang, P. J. Measuring emotion: The self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. Psychiatry* **25**, 49–59, 10.1016/0005-7916(94)90063-9 (1994).

12. Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178, 10.1037/h0077714 (1980).

13. Cowie, R., Douglas-Cowie, E., Savvidou, S., Sawey, E. M. M. & Schroeder, M. FEELTRACE: AN INSTRUMENT FOR RECORDING PERCEIVED EMOTION IN REAL TIME. *Proc. ISCA Tutor. Res. Work. on Speech Emot.* (2000).

14. Gunes, H. & Schuller, B. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image Vis. Comput.* **31**, 120–136, 10.1016/j.imavis.2012.06.016 (2013).

15. Fanourakis, M. A. & Chanel, G. Affective multimodal counter-strike video game dataset (amucs) - public, 10.26037/yareta:gvvoc4wfsfhupm4ygge26wupnm (2022).

16. Fanourakis, M. & Chanel, G. Attenuation of the dynamic pupil light response during screen viewing for arousal assessment. *Front. Virtual Real.* **3**, 10.3389/frvir.2022.971613 (2022).

17. The lab streaming layer development team. sccn/labstreaminglayer: Lab streaming layer. https://github.com/sccn/labstreaminglayer.

18. Fanourakis, M., Lopes, P. & Chanel, G. Remote Multi-Player Synchronization using the Labstreaming Layer System. In *Foundations of Digital Games Demos* (Malta, 2020).

19. Greenberg, S., Aislinn, P. & Kirsten, D. Development and Validation of the Fatigue State Questionnaire: Preliminary Findings. *The Open Psychol. J.* **9**, 50–65, 10.2174/1874350101609010050 (2016).

20. Gächter, S., Starmer, C. & Tufano, F. Measuring the Closeness of Relationships: A Comprehensive Evaluation of the 'Inclusion of the Other in the Self' Scale. *PLOS ONE* **10**, e0129478, 10.1371/journal.pone.0129478 (2015).

21. Melhart, D., Liapis, A. & Yannakakis, G. N. PAGAN: Video Affect Annotation Made Easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 130–136, 10.1109/ACII.2019.8925434 (IEEE, 2019). 1907.01008.

22. Lopes, P., Yannakakis, G. N. & Liapis, A. RankTrace: Relative and unbounded affect annotation. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, vol. 2018-Janua, 158–163, 10.1109/ACII.2017.8273594 (IEEE, 2017).

23. The pandas development team. pandas-dev/pandas: Pandas. https://github.com/pandas-dev/pandas, 10.5281/zenodo.3509134.

24. Robertson, A. R. The CIE 1976 Color-Difference Formulae. *Color. Res. & Appl.* **2**, 7–11, 10.1002/j.1520-6378.1977.tb00104.x (1977).

25. Donofrio, R. L. Review Paper: The Helmholtz-Kohlrausch effect. *J. Soc. for Inf. Disp.* **19**, 658, 10.1889/JSID19.10.658 (2011).

26. Weber, R., Behr, K.-M., Tamborini, R., Ritterfeld, U. & Mathiak, K. What Do We Really Know About First-Person-Shooter Games? An Event-Related, High-Resolution Content Analysis. *J. Comput. Commun.* **14**, 1016–1037, 10.1111/j.1083-6101.2009.01479.x (2009).

27. Baltrusaitis, T., Zadeh, A., Lim, Y. C. & Morency, L.-P. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 59–66, 10.1109/FG.2018.00019 (IEEE, 2018).

28. BERNTSON, G. G. *et al.* Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology* **34**, 623–648, 10.1111/j.1469-8986.1997.tb02140.x (1997). https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.1997.tb02140.x.

29. Freeman, R. & Chapleau, M. W. Testing the autonomic nervous system. In Said, G. & Krarup, C. (eds.) *Peripheral Nerve Disorders*, vol. 115 of *Handbook of Clinical Neurology*, 115–136, 10.1016/B978-0-444-52902-2.00007-2 (Elsevier, 2013).

30. Carreiras, C. *et al.* BioSPPy: Biosignal processing in Python. https://github.com/PIA-Group/BioSPPy/ (2015–).

31. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

32. Dawson, M. E., Schell, A. M. & Filion, D. L. The Electrodermal System. In Cacioppo, J. T., Tassinary, L. G. & Berntson, G. (eds.) *Handbook of Psychophysiology*, 217–243, 10.1017/CBO9780511546396.007 (Cambridge University Press, Cambridge, 2017).

33. Jaques, N., Taylor, S., Sano, A. & Picard, R. Multimodal autoencoder: A deep learning approach to filling in missing sensor data and enabling better mood prediction. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 202–208, 10.1109/ACII.2017.8273601 (2017).

34. Zuo, H., Liu, R., Zhao, J., Gao, G. & Li, H. Exploiting modality-invariant feature for robust multimodal emotion recognition with missing modalities. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5, 10.1109/ICASSP49357.2023.10095836 (2023).

35. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362, 10.1038/s41586-020-2649-2 (2020).

36. Virtanen, P. *et al.* SciPy 1.0: Fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272, 10.1038/s41592-019-0686-2 (2020).

37. Makowski, D. *et al.* Neurokit2: A python toolbox for neurophysiological signal processing. *Behav. Res. Methods* **53**, 1689–1696, 10.3758/s13428-020-01516-y (2021).

38. Ansel, J. *et al.* Pytorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*, 10.1145/3620665.3640366 (ACM, 2024).

39. Yannakakis, G. N., Martínez, H. P. & Jhala, A. Towards affective camera control in games. *User Model. User-Adapted Interact.* **20**, 313–340, 10.1007/s11257-010-9078-0 (2010).

40. Karpouzis, K., Yannakakis, G. N., Shaker, N. & Asteriadis, S. The platformer experience dataset. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 712–718, 10.1109/ACII.2015.7344647 (IEEE, 2015).

41. Alchalabi, A. E., Shirmohammadi, S., Eddin, A. N. & Elsharnouby, M. FOCUS: Detecting ADHD Patients by an EEG-Based Serious Game. *IEEE Transactions on Instrumentation Meas.* **67**, 1512–1520, 10.1109/TIM.2018.2838158 (2018).

42. Song, M. *et al.* Audiovisual Analysis for Recognising Frustration during Game-Play: Introducing the Multimodal Game Frustration Database. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 517–523, 10.1109/ACII.2019.8925464 (IEEE, 2019).

43. Blom, P. M., Bakkes, S. & Spronck, P. Towards Multi-modal Stress Response Modelling in Competitive League of Legends. In *2019 IEEE Conference on Games (CoG)*, vol. 2019-Augus, 1–4, 10.1109/CIG.2019.8848004 (IEEE, 2019).

44. Kutt, K. *et al.* BIRAFFE : bio-reactions and faces for emotion-based personalization. In *3rd Workshop on Affective Computing and Context Awareness in Ambient Intelligence (AfCAI 2019)* (Cartagena, Spain, 2019).

45. Beaudoin-Gagnon, N. *et al.* The FUNii Database: A Physiological, Behavioral, Demographic and Subjective Video Game Database for Affective Gaming and Player Experience Research. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, 1–7, 10.1109/ACII.2019.8925502 (IEEE, 2019).

46. Granato, M., Gadia, D., Maggiorini, D. & Ripamonti, L. A. An empirical study of players' emotions in VR racing games based on a dataset of physiological data. *Multimed. Tools Appl.* **79**, 33657–33686, 10.1007/s11042-019-08585-y (2020).

47. Smerdov, A., Zhou, B., Lukowicz, P. & Somov, A. Collection and validation of psychophysiological data from professional and amateur players: a multimodal esports dataset. *arXiv* **XX**, 1–12 (2020). 2011.00958.

48. Svoren, H. *et al.* Toadstool: A Dataset for Training Emotional Intelligent Machines Playing Super Mario Bros. In *Proceedings of the 11th ACM Multimedia Systems Conference*, 309–314, 10.1145/3339825.3394939 (ACM, New York, NY, USA, 2020).

49. Alakus, T. B., Gonen, M. & Turkoglu, I. Database for an emotion recognition system based on EEG signals and various computer games – GAMEEMO. *Biomed. Signal Process. Control.* **60**, 101951, 10.1016/j.bspc.2020.101951 (2020).

50. Melhart, D., Liapis, A. & Yannakakis, G. N. The Affect Game AnnotatIoN (AGAIN) Dataset. *arXiv* (2021). 2104.02643.

51. Kutt, K. *et al.* BIRAFFE2, a multimodal dataset for emotion-based personalization in rich affective game environments. *Sci. Data* **9**, 274, 10.1038/s41597-022-01402-6 (2022).

52. Dresvyanskiy, D. *et al.* Dycoda: A multi-modal data collection of multi-user remote survival game recordings. In Prasanna, S. R. M., Karpov, A., Samudravijaya, K. & Agrawal, S. S. (eds.) *Speech and Computer*, 163–177, 10.1007/978-3-031-20980-2_15 (Springer International Publishing, Cham, 2022).

Acknowledgements

This work was funded by Innosuisse with grant number 34316.1 IP.ICT.
The authors would like to thank Logitech S.A. for their vital collaboration in this study and for providing computer peripherals and the prizes for the winners of each round. We also thank the organizers of SwitzerLAN and PolyLAN for accommodating our study in their events and our student assistants for the long days and nights helping us collect the data.

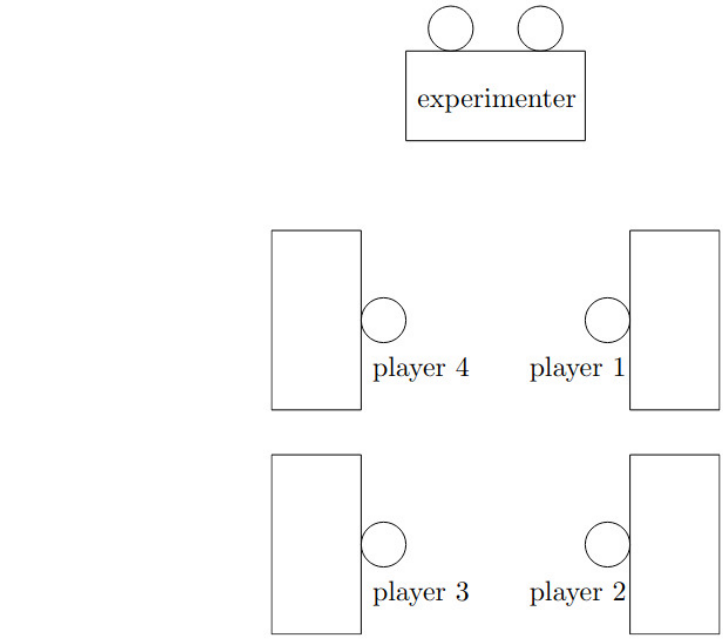
Author contributions statement

G.C. conceived the study. All authors conducted the experiments, analyzed the results, and reviewed the manuscript.

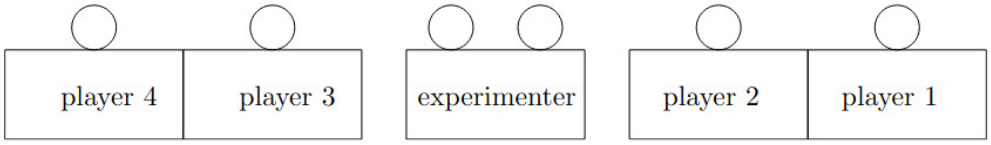
Competing interests

This work was done in collaboration with Logitech S.A but funded by Innosuisse. The authors declare no competing interests (for instance financial, profession or personal) with logitech S.A or any other party.

Figures & Tables

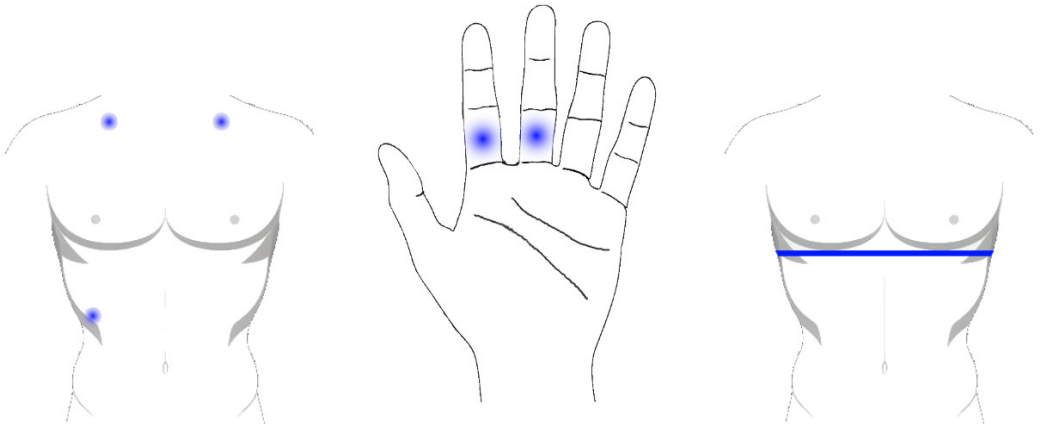


(a) SwitzerLAN 2020/2021 and PolyLAN 37



(b) PolyLAN 36

Figure 1. Physical experimental setup.



(a) Placement of the three ECG leads. **(b)** Placement of the two EDA leads. **(c)** Placement of the respiration sensor.

Figure 2. Bitalino sensor placements highlighted in blue.

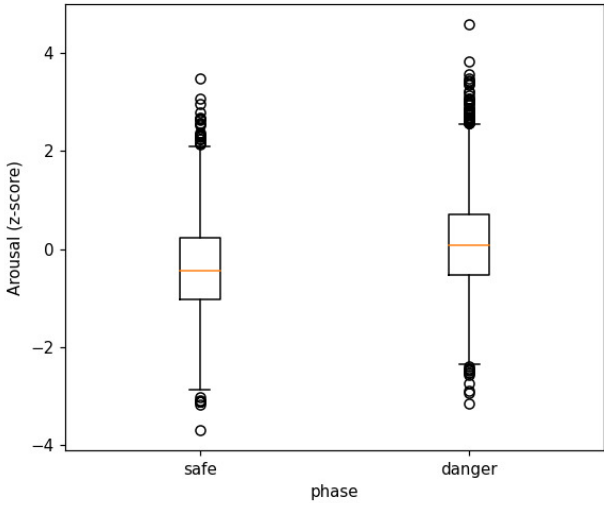


Figure 3. Participant annotations during each of the game phases. Arousal annotations during the *danger phase* tend to be higher compared to the *safe phase*. Orange markings indicate the median value.

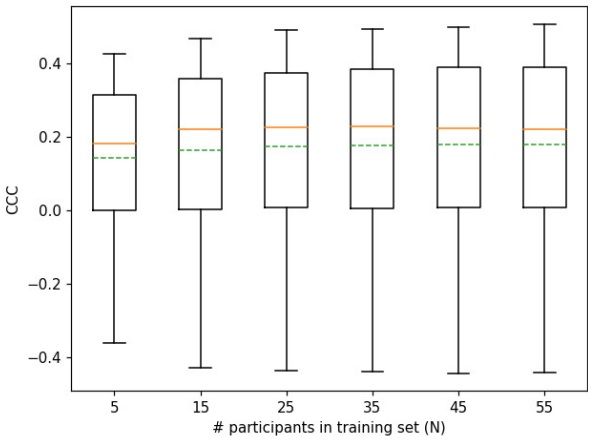


Figure 4. Boxplots of participants' mean CCC for each *N*-number of participants in training set. Orange markings indicate the median value, green markings indicate the mean value.

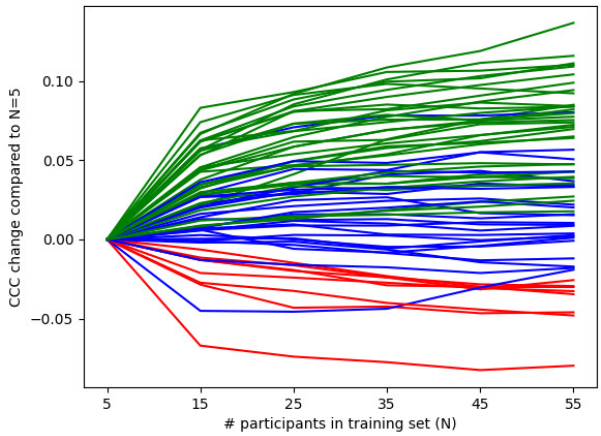


Figure 5. Change in CCC for each participant vs N compared to $N = 5$. Green curves show statistically significant increase at $N = 55$ (33 participants), red curves show statistically significant decrease at $N = 55$ (8 participants), blue curves do not have statistically significant changes between $N = 5$ and $N = 55$ (24 participants).

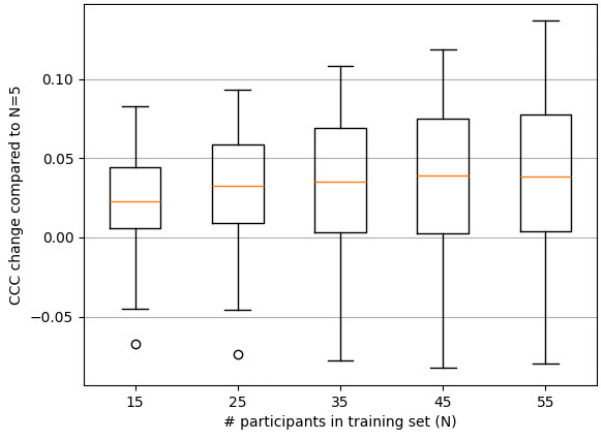


Figure 6. Boxplots of the change in CCC of participants at different N compared to $N = 5$.

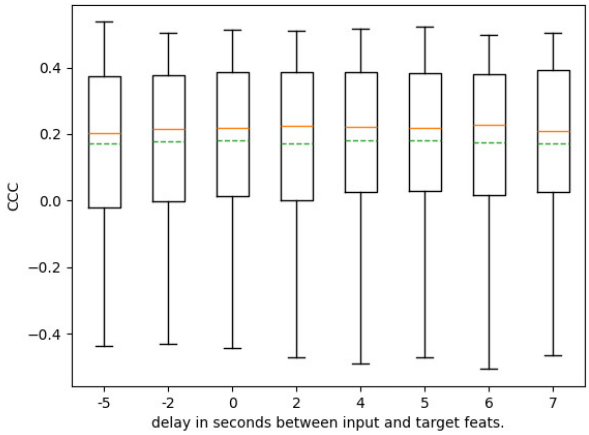


Figure 7. Boxplots of participants' mean CCC for each offset between the input and target window. Orange markings indicate the median value, green markings indicate the mean value.

Dataset	#Part.	Game	Modalities	Annotations	Notes
Maze-Ball[39] 2013	36	Maze-Ball (3D prey/predator)	BVP, SC	Pairwise, comparing conditions more/less anxious, exciting, frustrating, fun and relaxing	Environment: in lab; Conditions: control style, camera angle (8x conditions); Gameplay duration: 90s per condition (12min total)
PED[40] 2015	58	Infinite Mario Bros	Face video, game logs	Self-reported level of engagement, frustration and challenge (scale 0-4)	Environment: in lab; Conditions: level "A" and level "B"; Gameplay duration average: 1min per level (6h for 380 game sessions)
FOCUS[41] 2018	9	FOCUS (custom)	EEG (Emotiv Epoc+)	-	Environment: in lab; Conditions: control character via keyboard vs EEG, ADHD vs non-ADHD subjects; Gameplay duration average: 1min (keyboard), 2.5min (EEG, non-ADHD), 4min (EEG, ADHD)
MGFD[42] 2019	67	Crazy Trophy (custom voice controlled game)	Face video, speech	Self-reported frustration 4 point Likert scale; 4 external continuous annotations	Environment: in lab; Conditions: game control and feedback inconsistency; Gameplay duration average: 45s per level (4.5min total for 6 levels)
Blom et al.[43] 2019	8	League of Legends (multiplayer)	Keyboard, mouse, HR, GSR, PPG, game logs	-	Environment: in lab
BIRAFFE[44] 2019	206	Affective SpaceShooter 2, Freud me out 2	ECG, GSR, face video, joystick input, game logs	Self-reported valence and arousal	Environment: in lab; Conditions: the different games
FUNii[45] 2019	190	Assassin's Creed: Unity, Assassin's Creed: Syndicate	ECG, EDA, Respiration, EMG, eyetracking, face video, controller inputs	Continuous self-reported fun	Environment: in lab; Gameplay duration max: 35min
RAGA[46] 2020	36	Project Cars, RedOut	ECG, fEMG, GSR, Respiration	Continuous self-reported arousal and valence	Environment: in lab; Conditions: VR vs standard monitor
Multimodal eSports[47] 2020	10 (2x5)	League of Legends (multiplayer)	HR, IMU, EMG, GSR, eyetracker, EEG (Emotiv), pulse oximeter, keyboard, mouse, infrared (FLIR), game logs, environmental: temperature, humidity, and CO_2	Self-evaluation of own performance and teammates performance	Environment: in lab; Conditions: professional vs amateur team, vs bots or humans, with vs without team communication
Toadstool[48] 2020	10	Super Mario Bros	Accelerometer, temperature, EDA, BVP, IBI, HR, face video, gameplay video	Questionnaire	Environment: in lab; Gameplay duration: 35min
GAMEEMO[49] 2020	28	3x games: Train Sim World, Unravel, Slender - The Arrival, Goat Simulator	EEG (Emotiv)	SAM (arousal and valence)	Environment: in lab; Conditions: the different game types (boring, calm, horror, funny); gameplay duration: 5min per game (20min total)
AGAIN[50] 2021	124	Custom games: 3x racing, 3x shooters, 3x platformers	Gameplay video, game logs	Continuous and unbounded self-reported arousal	Environment: in lab; Conditions: the different games; Gameplay duration: 2min per game (18min total)
BIRAFFE2[51] 2022	103	3x 2D games: Room of the Ghosts, Jump!, Labyrinth	ECG, GSR, face video, gamepad input, game logs	Self-reported valence and arousal	Environment: in lab; Conditions: the different games; Gameplay duration max: 5min per game (15min total)
DyCoDa[52] 2022	30	Survival Game (multiplayer)	Infrared video, depth video, gameplay video and audio, speech audio	Self-evaluation questionnaires	Environment: in lab; total 10h of recorded data; Collaborative problem solving
AMuCS (ours)[15]	245*	Counter-Strike: Global Offensive (multiplayer)	Keyboard, mouse, ECG, EDA, Respiration, eyetracker, face features, depth video, seat pressure, gameplay video and audio, game logs	Continuous and unbounded self-reported arousal or valence	Environment: in-the-wild; Gameplay duration: 10min

Table 1. Open video game datasets with physiological or affective modalities.

signal	sensor	Hz	notes	raw file	processed file	published
ECG	Bitalino (r)evolution BT	100	10-bit resolution	xdf	csv	Y
EDA	Bitalino (r)evolution BT	100	10-bit resolution	xdf	csv	Y
Respiration	Bitalino (r)evolution BT	100	10-bit resolution	xdf	csv	Y
Left eye x gaze	Tobii pro nano	60		xdf	csv	Y
Left eye y gaze	Tobii pro nano	60		xdf	csv	Y
Left eye pupil diameter	Tobii pro nano	60		xdf	csv	Y
Right eye x gaze	Tobii pro nano	60		xdf	csv	Y
Right eye y gaze	Tobii pro nano	60		xdf	csv	Y
Right eye pupil diameter	Tobii pro nano	60		xdf	csv	Y
Seat pressure	Sensing Tex seat pressure mat	10	10-bit resolution; 16x16 sensor grid	xdf	csv	Y
Face color video	Intel RealSense D435	15/30	RGB24; 640x480 pixels	rosbag + xdf	mkv + csv	N
Face depth video	Intel RealSense D435	30	gray16le; 640x480 pixels	rosbag + xdf	mkv + csv	Y
openFace features	openFace[27]	15/30		N/A	csv	Y
Gameplay video	OBS	30	Same as game settings resolution	mp4 + xdf	mp4 + csv	Y
Keyboard buttons	Keyboard	N/A		xdf	csv	Y
Mouse buttons	Mouse	N/A		xdf	csv	Y
Game - isDucking	CS:GO plugin	64		xdf	csv	Y
Game - isJumping	CS:GO plugin	64		xdf	csv	Y
Game - isReloading	CS:GO plugin	64		xdf	csv	Y
Game - health	CS:GO plugin	64		xdf	csv	Y
Game - armor	CS:GO plugin	64		xdf	csv	Y
Game - bulletShots	CS:GO plugin	64		xdf	csv	Y
Game - damageToEnemy	CS:GO plugin	64		xdf	csv	Y
Game - damageFromEnemy	CS:GO plugin	64		xdf	csv	Y
Game - inFOV1/2/3/4	CS:GO plugin	64		xdf	csv	Y
Game - aimTarget	CS:GO plugin	64		xdf	csv	Y
Game - position X/Y/Z	CS:GO plugin	64		xdf	csv	Y
Game - velocity X/Y/Z	CS:GO plugin	64		xdf	csv	Y
Game - eyeVector X/Y/Z	CS:GO plugin	64		xdf	csv	Y
Valence/arousal annotations	PAGAN (gameplay vid.)	N/A		csv	csv	Y

Table 2. List of recorded data. Only a subset of the *game data* is listed (CS:GO plugin). Note that for the *face color video*, the first 108 participants (33 sessions) have a lower sampling rate of approximately 15Hz due to technical issues. This also affected the *openFace features*.

Feature	Description
Mean	The mean value of the signal
Variance	The variance of the signal
Range	The difference between the maximum and minimum values of the signal
Minimum	The minimum value of the signal
Maximum	The maximum value of the signal
First value	The first value of the signal
Last value	The last value of the signal
Centroid	1D center of mass of the signal, indicative of the slope of the signal but less computationally expensive
Number of peaks	The number of peaks in the signal detected from the inflection points of the derivative
Peaks mean	The mean of the peak amplitudes
Peaks variance	The variance of the peak amplitudes
Number of valleys	The number of valleys in the signal detected from the inflection points of the derivative
Valleys mean	The mean of the valley amplitudes
Valleys variance	The variance of the valley amplitudes

Table 3. Features derived for each windowed signal.

group	modalities	f1 score	Cohen's kappa
A	HR	0.50	0.06
B	EDA	0.68	0.38
C	Respiration	0.47	0.02
D	On-screen gaze speed	0.63	0.30
E	Facial action units	0.67	0.36
F	HR, EDA	0.69	0.40
G	HR, EDA, respiration	0.69	0.41
H	HR, EDA, respiration, gaze speed	0.72	0.45
I	HR, EDA, respiration, facial action units	0.73	0.48
J	HR, EDA, respiration, facial action units, on-screen gaze speed	0.75	0.52

Table 4. Game phase prediction results, leave-one-participant-out cross validation from a pool of 121 participants. All the participants in the pool had usable data for modalities group J.

↓better than→	A	B	C	D	E	F	G	H	I	J
A	-									
B	**	-	**	*						
C			-							
D	**		**	-						
E	**		**	*	-					
F	**		**	**		-				
G	**		**	**	*		-			
H	**	**	**	**	**	*	*	-		
I	**	**	**	**	**	*	*		-	
J	**	**	**	**	**	**	**	*	*	-

Table 5. One-sided paired Wilcoxon statistical test of game phase model performance with the different training modalities defined in Table 4. ** p-value smaller than 0.001, * p-value smaller than .01.

N	CCC mean
5 participants	0.141
15 participants	0.164
25 participants	0.172
35 participants	0.175
45 participants	0.177
55 participants	0.180
64 participants	0.181

Table 6. Arousal prediction results, leave-one-participant-out cross validation.

↓better than→	5	15	25	35	45	55
5	-					
15	**	-				
25	**	**	-			
35	**	**	**	-		
45	**	**	**	*	-	
55	**	**	**	**	**	-

Table 7. One-sided paired Wilcoxon statistical test of arousal model performance with different training set size N. ** p-value smaller than 0.001, * p-value smaller than .01.

signals	#sessions	#participants
ECG	67 (94%)	192 (78%)
EDA	65 (92%)	144 (58%)
Respiration	69 (97%)	210 (85%)
Eyetracker	70 (98%)	233 (95%)
Face video	64 (91%)	218 (89%)
Gameplay video	69 (97%)	227 (92%)
Keyboard buttons	67 (94%)	216 (88%)
Mouse buttons	67 (94%)	221 (90%)
Game logs	68 (95%)	234 (95%)
Valence annotations	65 (91%)	111 (45%)
Arousal annotations	68 (95%)	117 (47%)
Face video + valence annotations	59 (84%)	102 (41%)
Face video + arousal annotations	64 (90%)	108 (44%)
ECG + EDA + game logs	62 (87%)	132 (53%)
ECG + EDA + game logs + arousal annotations	50 (70%)	64 (26%)
ECG + EDA + game logs + face video + eyetracker	60 (84%)	126 (51%)
Eyetracker + gameplay video + game logs	66 (92%)	218 (88%)
Eyetracker + gameplay video + game logs + arousal annotations	63 (88%)	109 (44%)
Eyetracker + gameplay video + game logs + ECG + EDA	59 (83%)	123 (50%)

Table 8. Number of sessions (out of 71) and participants (out of 245) with usable data in the AMuCS dataset.