Is it Really Fun? Detecting Low Engagement Events in Video Games

Emanuela Guglielmi*, Gabriele Bavota†, Nicole Novielli‡, Rocco Oliveto*, and Simone Scalabrino*
*DEVISER @ University of Molise, Italy
†SEART @ Software Institute, Università della Svizzera italiana, Switzerland
‡COLLAB @ Universiy of Bari

***Abstract*—The gaming industry has witnessed remarkable growth in recent years, attracting millions of people who engage in its products both as a hobby and for professional purposes (*e.g.,* e-sports). Video games are software products that have a unique and fundamental requirement: They must be engaging. Previous research introduced approaches aimed at measuring engagement, some of which specifically designed for video games. Such approaches could be useful for practitioners since they can be adopted on the large collection of gameplay videos daily published on platforms such as Twitch and YouTube to allow developers to monitor players' engagement and detect areas in which it is low. Such specialized approaches have been evaluated on datasets in which the engagement was manually assessed by *external* evaluators based on the face of the player (we call it *perceived* engagement). We still do not know whether such approaches can capture the *real* engagement of players. Also, it is unclear to what extent practitioners would be willing to adopt such approaches in practice. In this paper, we provide two contributions. First, we ran an experiment with human 40 players aimed at defining a dataset of gameplay sessions in which participants self-reported their *real* engagement after every minute. We captured both their face and the gameplay. Based on this data, we compared state-of-the-art machine learning-based approaches to detect lowly engaging sessions. Our results show that the best model correctly classifies engagement in 74.7% of the cases and ranks video games in terms of their *real* engagement very similarly to how players would rank them (Spearman ρ = 0.833). Second, to assess the practicality of adopting such approaches in an industrial setting, we conducted two semi-structured interviews with senior game developers, who provided generally positive feedback and interesting insights for future developments.**

***Index Terms*—Gameplay videos, Video Games Quality, Mining software repositories**

I. Introduction

In Today's culture, video games have emerged as an increasingly important medium of expression. Their pervasive popularity, particularly among younger generations, has contributed to the growth of the gaming industry, which is the most profitable entertainment industry in the world [1], [2]. Besides, playing video games has evolved into a profession for many players (*e.g.,* e-sports and speed-running players and audience). As compared to other types of software, video games have a peculiar non-functional requirement: They need to entertain the users and be fun to play. In other words, they must be engaging. Politowski *et al.* [3] report the "lack of fun" is indeed one of the most common issues leading to the failure of video game projects.

Assessing the extent to which a game is engaging requires playtesting, a specific kind of end-to-end testing [4] in which players are asked to play with different parts (or even variants) of a game and provide their feedback about several aspects, including the game engagement. While other quality attributes of the game can be objectively and (in some cases) automatically assessed (*e.g.,* Frames Per Second as a proxy for performance), engagement is inherently subjective and hard to assess without asking the opinion of players. The high subjectivity implies that what is observed during playtesting in terms of user engagement might not generalize to the greater public. Indeed, playtesting is difficult to scale up in terms of involved players. The difficulties experienced by practitioners in assessing the level of entertainment of their products have been also documented in the literature [5]. In addition, due to the limited availability of automated approaches to quality control in game development [6], many games are released with unresolved problems that become apparent only when customers start playing the game [7]. Given the large amount of game content daily published by streamers,[1] it is very likely that some of these problems were encountered during their sessions.

Developers could rely on such a goldmine of information from streaming platforms such as YouTube [9] or Twitch [10] to find areas of a video game that are not engaging enough for most of the player, *i.e.,* that are affected from *objective* issues in terms of entertainment. Also, continuously mining the players' engagement in gameplay videos would allow developers to monitor how the engagement is impacted by the game updates introduced after the release or how to plan future releases. The facial expression in the recorded face of the streamers can be exploited to achieve this goal. Indeed, previous work showed that facial expressions can be leveraged to assess engagement [11] and/or emotions triggered by advertisements or movies [12]. Recent studies have shown that these solutions can effectively evaluate engagement in video games [13], [14]. However, the latter are different from other types of media because they are interactive. Therefore, specific facial expressions related to engagement could be radically different [11], [15].

While such specialized approaches show promising results, their capabilities have only been assessed on *perceived* engagement.

[1] For example, Fortnite has been the main subject of over 1.1 Billion hours of streaming videos published on Twitch in 2019 [8].

In such a scenario, players' facial expressions have been labeled by external evaluators, who defined the ground truth (*i.e.,* "engaged" or "not engaged") on which the approaches have been evaluated. In other words, the such ground truth is based on a subjective interpretation of the video content rather than on the direct feedback from the players themselves. It is still unclear to what extent such approaches are able to capture the *direct* engagement of the players, or whether this is possible in the first place. Also, it is unclear whether such approaches would be used by practitioners and how.

In this paper, we aim to fill this gap. First, we ran an experiment with players aimed a collecting a new dataset that provides the *direct* level of engagement experienced by the players during gameplay sessions. We asked 40 participants to play eight video games while we filmed their facial expressions and simultaneously recorded their gaming sessions. We paused the games at regular intervals of a minute and asked them to reporrt their level of engagement. In total, we collected 1,130 labeled pairs of 1-minute videos of facial expressions and engagement labels. Based on such a dataset containing *direct* engagement data, we compare two state-of-the-art models (Affectiva [16] and K [14]) and an additional ML-based approach that relies on state-of-the-art features (we conveniently call it FFBD– Facial Features-Based Detection– in the remainder of the paper). We show that the best model (*i.e.,* FFBD) achieves a good level of precision in identifying low-engagement events (74.7%). We also test whether using the best model would allow practitioners to find out which games suffer more of low-engagement issues. To this aim, we ranked the eight video games used in our study in terms of number of low-engagement events both by using (i) the predictions performed by the approach, and (ii) the actual low-engagement events declared by the players. The results show that the two rankings are very similar (Spearman ρ = 0.833).

As a second contribution, we explored the feasibility of applying the proposed approach in an industrial context. We conducted semi-structured interviews with two senior video game developers. Overall, they mostly provided positive feedback: They both acknowledged the utility of an approach for detecting engagement and its applicability in the game developement workflow. However, they highlighted that such approaches might not identify low engagement events if, during the gameplay the streamer is influenced by external factors (*e.g.,* chat interaction). Despite this limitation, developers noted that engagement estimation approaches could still offer valuable guidance, helping practitioners identify areas of concern, before and after the game release.

Our findings have important implications for researchers and practitioners. Researchers can rely on our dataset to define new approaches for engagement estimation in video games. Practitioners could integrate such approaches both (i) in playtesting activities to detect low-engagement levels at a finer grain, and (ii) in their after-release monitoring of the entertainment of their product, to timely detect when their products needs updates. We release a prototype tool implementing our approach which can be directly used on online gameplay videos to find low engagement events.

II. Background: Predicting Player Engagement

We present the problem of predicting player engagement during the game session and the most relevant approaches in the literature for achieving this goal.

A. Engagement in Video Games

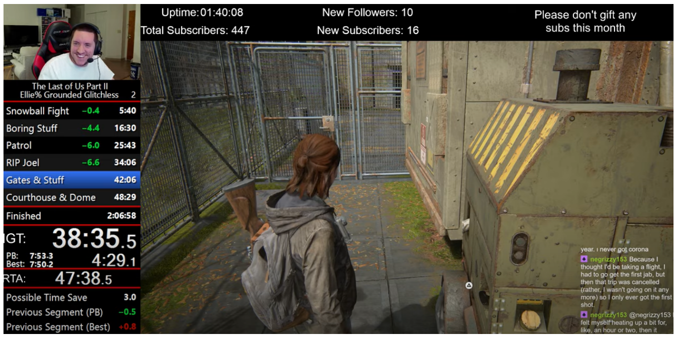
Engagement is a complex phenomenon. There is no agreement in previous research on the precise definition of engagement as a measurable psychological state, and many models for characterizing it have been proposed [17]. Engagement is often linked to the concepts of *immersion* and *involvement* in the context of video games [18]–[20]. In our work, we adopt the following definition of engagement: *Engagement is the degree to which the player is involved in the video game and, thus, the extent to which they are willing to continue playing.*

While playing video games, we assume that engagement depends on (i) the players' attitude and preferences in terms of the video game at hand, (ii) incidental factors, such as the tiredness or the willingness of playing at a given moment, and (iii) the engaging nature and quality of the *gameplay* of the video game. While the first two factors are out of the control of video game developers, the latter is what constitutes a non-functional requirement of every video game. The *gameplay* can be defined as the set of "*all actions performed by the player, influencing negatively or positively the outcome of the uncertain game situation in which they are engaged in*" [21]. More specifically, gameplay is usually implemented in a *gameplay loop*, *i.e.,* a cyclic sequence of phases that ultimately rewards the player and aims at increasing their willingness of playing (thus, engagement). A simplified example of *gameplay loop* for an RPG game is the following: The player buys weapons and armors to fight enemies, then they engage in fights and, finally, they are rewarded with more items or gold through which they can buy better weapons and armors, and the cycle restarts.

Engagement in a video game can drop when this cycle breaks. This happens, for example, when the game is too hard and, thus, the reward is never achieved (in the example, the enemies are too hard to defeat). Nevertheless, it is not true that the lower the difficulty the higher the engagement either: If the reward can be obtained without a challenge (in the example, the enemies are too easy to defeat) the player might experience boredom and thus being not engaged. As described by Ermi *et al.* [19], engagement can result in "*challenge-based immersion*", in relation to the mental skills "*such as strategic thinking or logical problem solving*". In short, the game should be aimed at being balanced [22]. Finding the "sweet spot" for obtaining engagement is a matter *game design* and *level design*.

B. Engagement Detection Approaches and Their Evaluation

Previous work defined several approaches for automatically measure engagement in video games [13], [14]. While each approach has its own peculiarities (we report them below), they are all based on the same premise. A player plays a video game and he/she is engaged or not engaged while doing so.



Snowball Fight −0.4 5:40
Boring Stuff −4.4 16:30
Patrol −6.0 25:43
RIP Joel −6.6 34:06
Gates & Stuff 42:06
Courthouse & Dome 48:29
Finished 2:06:58
IGT: 38:35.5
PB: 7:53.3
Best: 7:50.2
4:29.1
RTA: 47:38.5
Possible Time Save 3.0
Previous Segment (PB) −0.5
Previous Segment (Best) +0.8

year. i never got corona
negrizzy153 Because I thought I'd be taking a flight, I had to go get the first jab, but then that trip was cancelled (rather, I wasn't going on it any more) so I only ever got the first shot.
negrizzy153 @negrizzy153 I felt myself heating up a bit for, like, an hour or two, then it

Fig. 1: Example of low-engagement

As a result, the players feel emotions (*e.g.,* happiness) and reacts accordingly with specific facial expressions (*e.g.,* raises the eyebrows or moves the head in certain ways). Some features are extracted from the recorded facial expression and adopted to train a model to predict engagement. Since the training of such models is supervised, it is required to annotate the facial expressions with the ground truth. There are two ways in which such models are evaluated and, thus, the ground-truth is defined, *i.e.,* through *perceived* engagement and *real* engagement.

In the former, the models are evaluated in assessing engagement *as a human would* [13]. External evaluators (different from the players) observe the facial expression of players and manually assess whether they are engaged or not. This annotation constitutes the ground truth. This type of evaluation allows to understand to what extent the models are as good as humans in assessing the engagement. Achieving the maximum accuracy in this type of evaluation means that the model exactly mimics what a human would do.

In reality, however, human might fail at assessing the engagement of other humans by only relying on their facial expression. For example, a person with furrowed eyebrows and a neutral expression might can be interpreted as symptoms of engagement, while the same expression might indicate tiredness, anger, and thus low engagement. Consider, for example, the gameplay video at https://youtu.be/iAdcCjrL_6M?t=347 (one of the frames is also reported in Fig. 1). The streamer laughs, which could be interpreted as "fun." However, he is actually frustrated as he is blocked in a game mechanic that is not working, as it becomes clear in the continuation of the video. A human who only watches that specific segment would think that the player is engaged.

Thus, it appears clear that evaluating engagement prediction approaches on the *perceived* engagement might provide an over-estimation of their actual capabilities. Using the *real* engagement as the ground-truth, however, requires a direct feedback by the players, which should be collected in a timely manner (*e.g.,* right after the gaming session).

III. A Dataset of Real Engagement

In this section, we describe how we defined a dataset that contains evaluations of the *direct* engagement of the players. Specifically, to do this, we conducted a human study with players.

TABLE I: Video Games used for the empirical evaluation and their characteristics.

Video Game	Description
Space Invaders	Destroy the space invaders by firing your laser
Amidar	Visit all places on the grid while avoiding enemies
Gopher	Protect a crop of three carrots from a gopher.
Rayman	Platforming adventure
Snake	A modified version of Snake
Lonely	Reckless descent through pristine mountains
Golf Assassin	White pixel minigolf platformer

Context. The experimental context consists of *subjects* and *objects*. The subjects are 40 video game players, *i.e.,* people that regularly play video game and that played at least once any video game in the last year. To recruit them, we used convenience sampling. More specifically, we involved (i) students at the University of Molise, (ii) personal contacts of the authors, (iii) other people we reached out by locally publicizing such an activity.

The objects are 8 video games reported in Table I. We selected them aiming at covering different genres. Specifically, four of them were ATARI games (Space Invaders, Qbert, Amidar, Gopher), *i.e.,* classic 2D games, while four were free-to-play games available on the Steam Platform (Rayman, Snake, Golf Assassin, and Lonely). On the one hand, ATARI games offer simple game cycles compared to other commercial games considered. Such games reduce the possibility that non-game factors (such as narrative) influence observed engagement. The impact of such factors on player engagement is interesting, but not relevant to our work, which focuses on software-related aspects. On the other hand, modern games available on the Steam platform provide more complex mechanics. Again, we avoided games with a narrative to reduce their influence on observed engagement. As for the ATARI games, we used the versions available in the Gym Python library [23]. Gym offers a set of Atari 2600 environment simulated through the Arcade Learning Environment. In our study we use the default environment of each game. As for the games taken from the Steam Platform, we simply downloaded and installed them.

Protocol. Before starting the experiment, we asked all participants to sign an informed consent form which explains the purpose of the study and the treatment of the data acquired during the experiment. We also carefully explained the study protocol and what we would have asked them to report (*i.e.,* their level of involvement, as we explain later).

The experiment consisted in several subsequent *gaming sessions*. In each gaming session, the participants were asked to play a game and report their engagement at different times. More specifically, each gaming session started with a description of the game and the commands to be used. When the participant started the game session, we started acquiring their face through a webcam. Then we let the player play for 3 minutes. After each minute of gameplay, the game was paused and a small window was shown, asking the participants' level of involvement ("What is your level of involvement?") on a Likert scale from 1 (very low) to 5 (very high). We use this as an operationalization of the psychological state of engagement of the player, as previously done in the literature [24].

We divided the study in two parts: a mandatory first part and an optional second part. In the first part, we ran eight gaming session, one for each of the eight games in Table I. Then, the participants were asked whether they wanted to continue playing. If the answer was yes, the second part started, in which we ran additional gaming sessions (one at a time) with randomly chosen games and then asked whether they wanted to continue playing. The second part continued until the answer to the last question was negative. We did this to make sure that the player was not annoyed by the experiment (which could reflect in their facial expression and affect their engagement). In total, we collected approximately 19 hours of recordings.

Each participant was involved for ∼30 minutes, on average. The study was conducted in a controlled setting. Specifically, it has been executed in a laboratory with the same equipment in order to have consistent data acquisition across participants. One of the authors prepared the environment and supervised all the participants with the aim of monitoring the correct execution and intervening if they needed clarifications. To avoid biases due to the interactions among participants, we involved one participant at a time and we made sure that no other person was present during the experiment, except for the supervisor (one of the authors) who, however, did not interact with participants and was out of their sight. This allowed us to check that the change in the participant's emotions is due to the game played at that specific time alone.

The order in which the first 8 games were played was the same for each participant, while, as specified above, the execution of the other games is random. To guide the execution of the experiment and to automatically measure the variables we were interested in, we implemented a script that automated the execution of the whole games, the acquisition of the participants' face, and their level of involvement.

Before running the experiment, we ran a small pilot study with four additional participants (not involved in the main study), in order to test the framework and the protocol and to spot any possible problem before starting the study.

Dataset Characteristics. After collecting all the raw recordings of the participants' faces during the game sessions, we extracted the frames from each recording (10 frames per second), totaling 600 frames for each 1-minute gaming session. We labeled each of such sequences in a binary way: we used the *low engagement* class when participants reported 1 or 2 as level of involvement, and the *non-low engagement* class for the other evaluations (3 to 5). In the end, we collected a total of 1,130 data points, each one characterized by the frames extracted from the face recording associated with a binary label concerning the engagement (low engagement or non-low engagement).

IV. Study I: Predicting Real Engagement

The *goal* of our study is to understand if it is possible to use direct level of engagement to detect poorly engaging parts in video games, both for individual players and globally.

In particular, we aim to answer the following research questions (RQs):

- **RQ$_1$**: *To what extent is it possible to detect the direct engagement in a gaming session with state-of-the-art approaches?*
- **RQ$_2$**: *To what extent can state-of-the-art approaches rank the video game parts similarly to how humans would in terms of direct engagement?*

To answer both our RQs, we rely on the dataset collected in the human study presented in Section III.

A. Engagement Detection Approaches

We compare three engagement detection approaches: *Affectiva*, the approach by Killedar *et al.* [14], and an additional ML-based approach based on state-of-the-art features. We could not include FaceEngage [13] in our experiment because the tool implementing the approach is not publicly available. We describe such approaches below.

1) Affectiva: Affectiva is a leader in media analytics, in the field of emotion analysis. Their commercial solution relies on a large emotion database, consisting of data from 10 million consumer responses to over 53,000 advertisements in 90 countries. Affectiva measures the level of engagement through the weighted sum of a set of action units computed by looking at the facial muscles activation. In particular, it provides an engagement score between 0 and 100. The engagement score is calculated as the weighted sum of the following facial expressions: Inner and outer brow raise, Brow furrow, Cheek raise, Nose wrinkle, Lip corner depressor, Chin raise, Lip press, Mouth open, Lip suck, Smile. We used the free version of Affectiva available through its JavaScript APIs [25] in the IMotion platform [26]. Affectiva computes and reports an engagement score for a single frame, with values ranging between 0 and 100. For each game session recording, we aggregated such frame-level scores by computing the average engagement.

2) K+*:* Killedar *et al.* [14] introduced an approach aimed at assessing players' engagement via facial expressions. Such an approach focuses on facial emotions rather than expressions, meaning that the only features used to decide if the player is lowly engaged are seven emotions: neutral, sad, angry, happy, surprise, disgusted, and fearful. The authors process the detected emotions and use fuzzy logic to obtain an overall engagement score. This score is derived by evaluating the intensity of facial emotions and combining them to produce an engagement index, which reflects the experience of the player. We could not exactly replicate the approach by Killedar *et al.* since (i) no replication package is available, and (ii) at several points, the description of the approach is ambiguous. However, one rule is very clear: When the player prevalently shows neutrality, the engagement is very low. Thus, we decided to implement an optimistic version of the approach by Killedar *et al.* (K+) based only on this rule. In our implementation, if the precondition is met (*i.e.,* if the prevalent emotion detected in most frames is "neutral"), we assigned the label *low engagement*; otherwise, we assumed that the remainder of the approach is correct, *i.e.,* we use as predicted label the actual label.

TABLE II: Biometric Data Acquired and their Relationship with Engagement.

Name	Description	Relationship with Engagement
Pitch	Rotation around the side-to-side axis	↓ Frustration or disapproval
Roll	Rotation around the front-to-back axis	↑ Interest in something
Yaw	Rotation around the vertical axis	↓ Distraction
AU01	Inner Brow Raiser	↑ Involvement with the content
AU02	Outer Brow Raiser	↑ Involvement with the content
AU04	Brow Lowerer	↑ Involvement with the content
AU05	Upper Lid Raiser	↓ Frustration
AU06	Cheek Raiser	↑ Involvement with the content
AU07	Lid Tightener	↓ Frustration
AU09	Nose Wrinkler	↑ Involvement with the content
AU10	Upper Lip Raiser	↓ Frustration
AU11	Nasolabial Deepener	↑ Involvement with the content
AU12	Lip Corner Puller	↑ Involvement with the content
AU14	Dimpler	↓ Lack of Involvement
AU15	Lip Corner Depressor	↑ Involvement with the content
AU17	Chin Raiser	↑ Involvement with the content
AU20	Lip Stretcher	↑ Involvement with the content
AU23	Lip Tightener	↓ Frustration
AU24	Lip Pressor	↓ Frustration
AU25	Lip Part	↑ Involvement with the content
AU26	Jaw Drop	↑ Involvement with the content
AU28	Lip Suck	↑ Involvement with the content
AU43	Eyes Closed	↑ Involvement with the content
Anger	Intense emotion of displeasure, frustration, and hostility	↓ Frustration
Disgust	Revulsion and aversion towards something unpleasant or offensive	↑ Involvement with the content
Fear	Perception of possible imminent danger or threat.	↑ Involvement with the content
Happiness	Positive and joyful emotion characterized by contentment and satisfaction.	↑ Involvement with the content
Sadness	Deep and intense emotion of sorrow and unhappiness.	↑ Involvement with the content
Surprise	Sudden and unexpected emotional reaction.	↑ Involvement with the content
Neutrality	Absence of clear emotional response or indifference towards a situation.	↓ Lack of involvement

Note that, in our implementation, we use the emotion detection provided by Py-Feat instead of the custom one used in the paper since the training dataset was not available and, again, we could not replicate that part of the approach. The set of emotions detected by both techniques is the same.

3) FFBD: We considered an additional approach — we conveniently call it FFBD, *i.e.,* Facial Features-Based Detection. FFBD is based on a complete set of state-of-the-art facial features and relies on a classic ML algorithm (*i.e.,* Random Forest[2]), which requires to be trained to classify a given instance in a binary way (*low* or *non-low* engagement). We describe in Section IV-B how we trained the model for our experiment. FFBD relies on three categories of features: emotion, expression, and behavior features. In Table II, we describe such categories and explain how they might relate to engagement based on empirical evidence from previous work.

Emotion Features. Positive emotions, such as joy, excitement, or satisfaction, are related with positive engagement in the game [27], [28]. Such emotions can be triggered, for example, by success in the game, achieving goals, overcoming challenges. Whereas, negative emotions, such as frustration, anger, or boredom, may indicate a negative level of engagement. We use Py-Feat [28], a publicly available tool that leverages consolidated approaches for emotion recognition through facial expression analysis to determine emotions and other characteristics from static images that include the face of a human (in our case, the player) [29]. More specifically, given an image, the tool reports a probability that the person is experiencing the emotion reported in Table II. To compute video-level features starting from the emotions measured on each frame composing the video, we calculate for each of them, the first, second, and third quartiles (Q1, Q2, and Q3), and their standard deviation (SD). This means that for a video including 600 frames, we will have 600 values for each emotion (*i.e.,* a probability that the user is experiencing that emotion) that represent a distribution on which the above-listed statistics are computed. In addition, for each emotion e we compute its total duration (TD_e) as the percentage of frames in the video for which it is the prevalent emotion (*i.e.,* the one with the highest probability). For example, if there are five frames, and in three of them the player has a prevalent *happy* emotion, the $TD_{happiness}$ feature is 60%. We also compute the longest sequence of frames in which an emotion is prevalent (LD_e). To do this, we first assign the prevalent emotion to each frame. Then, for each e, we compute the longest sequence of consecutive frames marked with e and calculate its percentage with respect to the total duration of the video. For example, if there are, in total, ten frames, and there are at most 3 consecutive frames in which the player is *happy*, the $LD_{happiness}$ is 30%. Finally, we compute features related to the emotions of the player shown at the very last frame of the recording. The assumption is that such emotions shown by the player just before the time in which we are interested in capturing the engagement might be important in assessing it. We consider both the probabilities reported by Py-Feat to each emotion (ranging from 0 to 1) and the binary value (0 or 1), where 1 is assigned to the prevalent emotion in the last frame, while 0 to the other ones. Similar measures have be used in previous work [30].

Expression Features. While emotions help determine the level of engagement of a player, a finer-grained level of detail (*i.e.,* the raw action units) might help detecting expressions possibly related to low engagement that are not related to specific emotions. Again, we use Py-Feat [28], [29] to compute the probability that the player is performing all the action units we consider (see Table II). As we do for the emotion features, we aggregate each expression feature previously listed (*i.e.,* the probabilities reported by Py-Feat for each action unit) by computing the first, second, and third quartiles, and their standard deviation, for all frames in the video. Also in this case, we add features related to the last frame (one for each action unit). Such features have as values the probabilities that the player performed the related action unit in the last frame.

Behavior Features. The behavior of a video game player can be used to assess their engagement by determining the player's level of immersion in the game. For example, the physical proximity of the player to the screen can indicate a high level of engagement. Since we assume we only have information about the face, we only focus on the behavior exhibited through the head. Specifically, we consider as features the pitch, roll and yaw of the head. We use Py-Feat to compute the three previously-mentioned features for each frame. Such features range between ±90 degrees (roll), ±75 degrees (yaw) and ±60 degrees (pitch) [29].

[2] We used the implementation with the default configuration provided in Weka — http://www.cs.waikato.ac.nz/ml/weka/.

As we do for both the previously-reported categories of features, we aggregate each behavior feature by computing the first, second, and third quartiles, and their standard deviation, of the measurements performed for all frames of the video. We also use the values of Pitch, Roll, and Yaw observed in the very last frame, as done for the other categories of features.

B. Experimental Design

To answer RQ$_1$, we compare the three previously-described approaches on our dataset containing *direct* engagement evaluations. Both Affectiva and K+ do not require training. So, we simply use them to predict the engagement of all the data points in our dataset. Note that Affectiva returns a continuous score rather than a binary classification. To classify the engagement level as *low* or *non-low*, we used two thresholds. The first one is $k = \frac{2}{5} \times 100 = 40$. We did this because, as explained, we discretized engagement as *low* when the engagement level self-reported by the users is lower than $\frac{2}{5}$. The second threshold is the one that allows to achieve the best F1 score. We report the results achieved by using threshold levels between 10 and 100 at steps of 10 to have a complete view of the performance of Affectiva. As for FFBD, instead, we use a 10-fold cross validation to evaluate the model, which consists in dividing the dataset in ten equally-sized folds and using 9 of them for training (1,017 instances) and one for testing (113 instances). Note that, as a result, we obtain the predicted engagement for all the instances, thus making the results comparable to the ones obtained with the other approaches. We noticed that Affectiva fails to compute the engagement scores on some frames (*e.g.,* frames with a light rotation of the face, whereas with PyFeat we are able to do this). To perform a fair comparison, we re-trained/tested FFBD also on a subset of instances from which excluded such frames.

We computed and report the recall, precision, and F1-score (the harmonic mean of precision and recall) for both the classes (*i.e., low engagement* and *non-low engagement*). We also report the AUC (Area Under the ROC curve [31]). An AUC $\simeq$ 0.5 indicates a model having the same prediction accuracy of a random classifier. A perfect model (*i.e.,* zero false positives and zero false negatives) has AUC = 1.0.

To answer RQ$_2$, we rank the game areas considered in our study in two different ways. First, we do that based on the best prediction model resulting from RQ$_1$ and, specifically, on the total number of predicted *low-engagement* events. Second, we rank the areas based on the number of self-reported *low-engagement* events. Ideally, a perfect alignment of the ranking means that the predicted engagement could be used instead of the *direct* one to detect areas that are lowly engaging according to several players without explicitly asking them. Note that since in our experimental design each user plays each game for at most 3 minutes and they play at most a level of each game, we can safely say that all the recordings correspond to a single game section.

To compare the two rankings, we computed the Spearman's rank correlation coefficient (ρ) [32], which computes the statistical dependence between the rankings of two variables.

TABLE III: RQ$_1$: Affectiva threshold evaluation. The metrics are referred to the *low engagement* class.

k	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**	**F1**
10	174	239	504	213	0.42	0.45	0.44
20	214	351	392	173	0.38	0.55	0.45
30	261	433	310	126	0.38	0.67	0.48
40	**302**	**522**	**221**	**85**	**0.37**	**0.78**	**0.50**
50	326	583	160	61	0.36	0.84	0.50
60	354	632	111	33	0.36	0.92	0.52
70	367	668	75	20	0.35	0.95	0.52
80	381	705	38	6	0.35	0.99	0.52
90	**386**	**717**	**26**	**1**	**0.35**	**0.99**	**0.52**
100	387	743	0	0	0.34	1.0	0.51

A high correlation coefficient indicates that the two rankings are very similar. We also reported the p-value of the correlation, which indicates the probability that the correlation is different from 0.

C. Results

RQ$_1$: Individual Level of Engagement. Table III reports the results obtained using the engagement score provided by Affectiva. The best results in terms of F1-score can be achieved with $k = 90$. Thus, we use Affectiva $_{k=90}$ and Affectiva $_{k=40}$ (for the previously-explained reasons) as representatives of a binary classifier based on Affectiva.

We show in Table IV the comparison among the three approaches. The top part of the table reports the results achieved on the whole dataset, while the bottom part provides the results achieved on the subset of frames that can be treated by Affectiva. The effectiveness of FFBD is consistently higher than the other two approaches. However, such a approach confuses several low engagement events for non-low engagement events (the recall is only 0.41 for the *low engagement* class). This is probably due to the unbalance between the two classes: 35% of the instances are low engagement, while the other 65% are non-low engagement. The AUC of FFBD is 0.79, which shows that the state-of-the-art features have a high capability of distinguishing the two classes.

Both the Affectiva binary classification versions we considered have a low power of distinguishing the two classes. This is clear when looking at the AUC, which is only 0.58 for such a the other approach, while it is 0.79 for our evalaution.[3] It is worth noting that Affectiva achieves much better results in the classification of engagement in other contexts [12]. This suggests that assessing engagement in video games is a different problem. Retraining Affectiva would probably allow to achieve results comparable to the one achieved by FFBD. We could not do that because it is a closed-source tool. K+ achieves slightly better results than Affectiva but, again, it is an optimistic implementation of the direct approach, K, that would very unlikely be able to achieve such results in its intended implementation.

In an attempt to characterize engagement in video games, we measured the importance of the state-of-the-art features used in FFBD through the Info Gain algorithm [33].

[3] Note that AUC is independent from the threshold, so the AUC for Affectiva is the same for both the thresholds we considered.

TABLE IV: Comparison between FFBD and the other two approaches.

Tools	Precision	Recall	F1	AUC
FFBD $_{bas}$	0.75	0.75	0.75	**0.79**
FFBD $_{comp}$	0.73	0.73	0.73	**0.79**
Affectiva $_{k=40}$	0.37	0.78	0.50	0.58
Affectiva $_{k=90}$	0.35	0.99	0.52	0.58
K+	0.27	1.00	0.53	0.42

TABLE V: Feature importance for emotion- (♥) and expression- 😐) features. LF indicates the features computed in the last frame (either *cont*inuous or *bin*ary).

Rank	Category	Feature	Importance
1	♥	Happiness (Q1)	0.032
2	♥	Happiness (LF-cont)	0.031
3	♥	Surprise (Q2)	0.029
4	♥	Surprise (Q3)	0.027
5	♥	Happiness (SD)	0.026
6	♥	Fear (Q1)	0.022
7	♥	Surprise (Q1)	0.022
8	😐	AU01 (Q2)	0.021
9	😐	AU17 (Q1)	0.020
10	♥	Fear (Q2)	0.020

Then, we ranked them based on their prediction power. We report the top ten features in Table V. It can be noticed that at least a feature from two of the three categories of state-of-the-art features (expression and emotion) appears in the ranking, while the first one from behavior features (first quartile of Yaw) appears in the 15th position. Most of the important features are emotion-related, which supports the focus given in the literature to such an aspect [17], [34]. More specifically, it appears that the ones related to happiness and surprise are the most important ones (top five positions). Among the several action unit we considered in our expression-related features, only two are among the most important ones, *i.e.,* AU01 (inner brow raiser) and AU17 (chin raiser).

Answer to RQ$_1$. State-of-the-art approaches are highly effective in predicting the *direct* engagement of players (best F1: 0.79). However, the recall for the *low engagement* class is lower than 50% for the best model.

RQ$_2$: Real Engagement in Practice. Table VI shows the rankings of video games obtained by using the self-reported number of low engagement events and the predicted one. To this aim, we relied on FFBD, which achieves the best results. The number of low engagement events (actual and predicted by the model) at game level are significantly and strongly correlated (Spearman ρ = 0.83, with a p-value of 0.015).

It can be observed that the first three positions in the ranking (*i.e.,* the three games for which the players reported to be less engaged with) are the same. In other words, a prediction model can correctly rank them in terms of engagement. On the other hand, it struggles when the number of low engagement events is generally lower (*i.e.,* in the lower part of the ranking), probably because of its low recall. The error, in general, is at most of two ranking positions.

TABLE VI: RQ$_2$: Video Games rankings (direct and predicted), with the number of low engagement events for both the scenarios and the rank difference between the two.

#	Real (SR)		Prediction (SA)		Diff
1	Amidar	87	Amidar	44	0
2	Qbert	68	Qbert	36	0
3	Space Invader	63	Space Invader	32	0
4	Lonely	46	Gopher	31	+1
5	Gopher	39	Snake	24	+2
6	Golf Assassin	33	Lonely	23	-2
7	Snake	30	Rayman	18	+1
8	Rayman	19	Golf Assassin	16	-2

Answer to RQ$_2$. An approach based on state-of-the-art features — FFBD — allows to rank game sections based on the presence of low-engagement events as players would do.

V. Study II: Industrial Applicability

We further studied the applicability of state-of-the-art engagement detection approaches with industrial practitioners. We report below the design and results of such a study.

A. Study Design

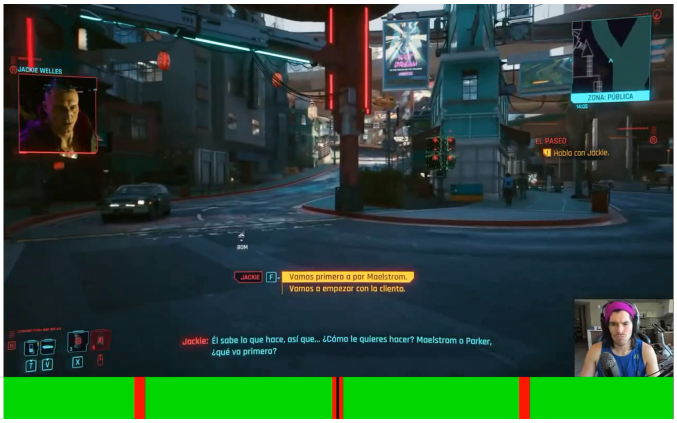
The *goal* of this second study is to evaluate the practical applicability of an engagement detection tool in an industrial context.

This study was steered by the following research question:

- **RQ$_3$**: *Would an engagement detection tool be industrially relevant?*

1) Context: The context of the study is composed of two subjects (participants) and an object (gameplay video). As subjects, we involved two *game developer*. We selected the participants using convenience sampling (both of them are former students at the University of Molise). Specifically, such participants are (i) Lorenzo Valente, Lead Developer in Tiny Bull Studios (Italy), and (ii) Jonathan Simeone, Full Stack Developer in Datasound (Italy), both with more than 7 years of experience in game development.

As for the object, we used as a gameplay video of the *Cyberpunk 2077* game. We chose such a video game because it has received several negative reactions when it was released [35]–[37], thus making it more likely to find videos in which streamers displayed low engagement. To select the specific gameplay video, we iteratively ran FFBD on the first results related to such a video game from YouTube until we found one for which it detected at least 5 low engagement events. To showcase the capability of our study, we implemented a prototype tool that, given a gameplay video and the location of the streamer's face as input, adds the predicted engagement-related information on the video (*i.e.,* indicates in which parts the player is poorly engaged). More specifically, we ran FFBD (the best-performing one) to predict the streamer's engagement for every minute of the video, and we added a bar at the bottom of the video in which we mark in red the potential low engagement events, while in green we indicate the non-low engagement events (see Fig. 2).



Vamos primero a por Maelstrom.
Vamos a empezar con la clienta.

Jackie: Él sabe lo que hace, así que... ¿Cómo le quieres hacer? Maelstrom o Parker, ¿qué va primero?

Fig. 2: An example of what we showed to participants. The bar below shows engagement in time (red → low engagement).

2) Experimental Procedure: To answer RQ$_3$, we conducted semi-structured interviews. Before each interview, one of the authors explained the objective of the study. Each interview lasted about 30 minutes and was conducted by one of the authors, who recorded and transcribed what the participants said for the following analyses. The interviews were based on a reflective strategy: We encouraged participants to share their experiences, thoughts and insights in an introspective way. To do this, we mostly asked open-ended and exploratory questions. Specifically, the interview conductor asked the participants to focus on the parts of gameplay video at which the approach detected possible low engagement events (one at a time), but they could freely navigate the video to get more context. After they analyzed each of them, we asked for feedback aimed at understanding whether (i) the trigger for the identified low engagement event is actually due to a problem in the game, (ii) the information about the low engagement event is sufficient to reproduce the problem (*i.e.,* recreate the game conditions that led to the observed low engagement event), and (iii) witch other information allows to identify low engagement events.

After the evaluation of the five events, we asked questions aimed at getting feedback on an hypothetical engagement detection tool that works as the prototype tool we showed them. Specifically, we asked whether they would use this tool to identify low engagement events: (i) in the playtesting phase, before a release, and (ii) in the testing phase after the game release. Also, we asked them whether they would use it in combination with other tools (and, if so, which ones). Finally, participants were asked to indicate on a Likert scale from 1 to 5 (the higher the better) the usefulness of an engagement detection tool in identifying potential low engagement events, and whether they had possible suggestions on how the detection accuracy could be improved. For all questions, participants were asked to motivate their answers. We report the complete list of questions we asked in our replication package [38].

B. Results

Identifying Low Engagement Events. Lorenzo confirmed that three of the five parts of the video we made them watch contained low engagement events. Lorenzo confirmed that the streamer is not having fun in the first low engagement event observed. He states: "The streamer was at the end of a scripted sequence where the player is in a piloted vehicle and where he shots the last enemies, and from that point he just had to wait for the mission to end. Scenes like these generally represent parts of the game with a lower level of engagement." In the analysis of another low engagement event, Lorenzo claims that the streamer is in an area where he is looking for a mission or clue from the map that he cannot find. He adds: "It is a problem of the game because it requires the player to find the goal in order to move forward in the game. However, if they receive too little information, players have trouble moving forward, and the engagement is lowered." An additional low engagement event is identified in the cutscenes where the streamer is in a narrative phase of the game where they have to listen to parts of dialogue. On the other hand, the two instances classified by Lorenzo as non-low engagement contain (i) a phase of the game where the streamer is exploring, and (ii) a phase of the game where the player was being very careful (*e.g.,* area full of enemies). In these two events, Lorenzo reported that it is very likely that the player was entertained in those situations.

Jonathan, instead, reported that all of the five parts of the video we made them watch contained low engagement events. However, he claimed that two of the events (the same ones that Lorenzo reported as false positives) were harder to assess. To claim this, Jonathan examined the context related to the part of the gameplay that preceded the reported low-involvement event. He observed an actual change in engagement by the streamer where in both cases there was a shift from a more dynamic to a more static gameplay phase.

Both Lorenzo and Jonathan reported that all events identified by the tool provide sufficient information to reproduce the conditions that caused of the issue. The game parts are clear and provide sufficient context.

Additional Information to identify Low Engagement Events. Lorenzo claims that a piece of information that could be useful in analyzing these events is whether the streamer is interacting with the chat. The level of engagement could be affected by the fact that the streamer is distracted reading or responding to comments. Therefore, it would be useful to integrate information about the presence of external elements that influence the streamer (*e.g.,* whether the streamer is looking toward the chat or they are happy because they received a donation).

Jonathan claims that it would be interesting to analyze the streamer's game actions (*i.e.,* how they interact with game elements). He states: "When my engagement level drops, I start exploring the game scene and the menus. If they have too much information or contain distracting elements, I do not read the descriptions carefully, so I have to go back and forth and I have even lower engagement.

For example, it could be helpful to consider also (i) the time spent in a specific game scene, (ii) the number of times the player dies, and (iii) whether they interact with the available game elements (*e.g.,* weapons). Many elements could be confusing, and if the player does not use them this could make the game more difficult and consequently lower the level of engagement."

Both Lorenzo and Jonathan claim that valuable information could be gained from analyzing the streamer's audio. By combining it with the available video information, one could understand how the streamer interacts with the game or live broadcast. For example, the streamer could express amusement while commenting during the live stream, even when immersed in a gameplay scene with minimal engagement.

Practical Application. In relation to the possibility of using a tool for engagement detection *before the release* of the game, Lorenzo stated that it could be very useful for game designers during beta-testing: "A game designer might find it useful to identify parts of the game that are particularly boring to the player. Such information would give them the ability to identify parts of the game that could be changed (*e.g.,* excluded)." Jonathan finds in the such tools a support in beta-testing as well. "I would use such a tool because it would be a great asset [...]. A low engagement event would allow me to identify a game design problem. When the player is not engaged it means that there is a part of design where I did something wrong." In relation to the use of an engagement detection tool *after the release* of the game, both Lorenzo and Jonathan were positive. In particular, they recognized the potential of the large and growing amount of information now available through the streaming platforms. Lorenzo states, "This tool could be used at scale, on a much larger pool of streamers. By analyzing their gameplays, a game developer can get a lot of information to combine to get a complete overview. In this way, one could identify a possible part of the game where players have a low level of engagement." Again, Jonathan shows strong confidence: "Absolutely yes, the game continues to be tested by end buyers and I want to keep monitoring it. Having such information would give me the ability to automatically detect problems even after the game is released." Both participants would use an engagement detection tool in combination with other tools. In details, Lorenzo states, "Yes, I would use it in combination with other tools (if they are available) which automatically allow me to identify things that were missed during the testing phase." Jonathan points out, "I do not know of any other tools that allow you to obtain information on the level of engagement, but I would use any tool similar to this one if it supports me in identifying issues missed in testing."

Usefulness of Detection Approaches and Suggestions. In terms of usefulness, Lorenzo gives a rating of 5 out of 5. In relation to the possibility of improving the detection accuracy, he suggests taking information from the streamer's audio and any textual information captured from the chat interaction.

In relation to the same aspect, Jonathan gives a rating of 4 out of 5. "I found consistency in what I saw. However, in some cases, I had to force myself looking at context before identifying the issue." He also suggests that to improve the accuracy it would be interesting to allow developers to give feedback on identified low engagement events in order to recognize false positives. Therefore, a continuous learning model could be useful to improve the detection accuracy through feedback from those who use it. Similarly to Lorenzo, Jonathan highlights the possibility of obtaining information through analyzing the streamer's audio and textual information from the chat.

Answer to RQ$_3$. The participants provided positive feedback on using an engagement detection tool, suggesting improvements like incorporating audio, chat, and game context to better detect low engagement events. They emphasized that the tool's usefulness depends on its integration into game developers' workflows, with potential use both before and after a game's release.

VI. Threats to Validity

Threats to construct validity are related to possible inaccuracies in the *real* engagement assessment we performed to define our dataset. Our interruptions might have influenced the engagement of the participants. To understand if this has been an issue, we asked a subset of participants to re-watch all their eight gameplay sessions, including both their face and the game recording, and re-evaluate their level of engagement (Likert scale from 1 to 5) without knowing their previous evaluation. We asked them to try to remember whether they were enjoying playing the video game or not. We involved 10 randomly selected participants out of the 40 we involved in the study. We used the same methodology we used in our experiment to transform the evaluations from 1 to 5 into the two classes "low engagement"/"non-low engagement"). Finally, we compared the new binary labels with the ones in our dataset. We observed an agreement rate of 88.75%. More specifically, we found only 9 cases of disagreement out of a total of 80 observations. Out of these, 6 observations transitioned from low to non-low engagement, while 3 observations changed from non-low to low engagement. This result suggests that our procedure for labeling the instances is sufficiently accurate and, most importantly, it is unlikely that our procedure artificially increased the number of "low engagement" events. However, this analysis further reveals the dynamic nature of engagement and the influence of other external and internal factors.

Threats to internal validity concern the design choices that we made that could affect the results of the study. The main threat is related to the implementation of the approaches we compared in the first study. First, we could not fully replicate one of them (K) because of the small amount of details available. To mitigate this threat, we compared our approach with an *optimistic* version of K, K+, which achieves better results than K by construction. Similarly, we did not have access to the commercial version of Affectiva, and thus we used the free implementation available online. We acknowledge that such a version might achieve worse results than the commercial one.

Finally, we did not include in the comparison the approach by Chen *et al.* [13] because the replication package was not available and we lacked sufficient details to re-implement the approach from scratch. Another possible limitation is that we did not perform hyperparameter tuning for Random Forest. We decided not to run this step due to the very limited amount of data points we had: We preferred to avoid dedicating some of them for such a step rather than for training/testing. Also, our results are basically a lower bound of what can be obtained with hyperparameters tuning.

Threats to external validity concern the generalizability of our results. Our test set consists of 40 players selected from a population with specific characteristics (*i.e.,* mostly young University students). Such a sample may not be representative of the entire population. Further replications of our study and extensions of our dataset are necessary to improve such an aspect. As for our semi-structured interviews, we only involved two developers that share part of the educational background (Bachelor degree at the same University) and that work in small companies. The results we obtained might not be generalizable to the whole population of game developers, and other developers might find our approach not useful.

VII. Related Work

We discuss the related literature focusing on (i) techniques assessing quality aspects of video games which are related to the users' engagement, and (ii) more general engagement measurement (outside the video games context).

Identifying quality issues in video games is a relevant issue previously explored in the literature [5], [39]. Video games can suffer from a wide range of problems. Truelove *et al.* [7] define a taxonomy in which they classify the types of problems reported in video games. These include problems related to game balance. Guglielmi *et al.* [6] introduced an approach to identify segments of videos in which streamers highlight anomalies and categorizes them accordingly to their type. Among the categories, they also considered balancing issues, which are conceptually related to engagement. They found that balancing issues in video games are very difficult to identify through the analysis of game contents and streamers' comments from gameplay videos.

Pfau *et al.* [40] applied Deep Player Behavior Modeling to generate models that could reproduce the playtesting strategy. The main goal of the tests was automatic balancing of game difficulty. However, trained AI agents, which emulate the actions of individual players, can also be used for game exploration and detection of bugs and problems within the game. Also, Pfau *et al.* [41] present a fine-grained study of Guild Wars 2 community attitudes about balancing factors. They introduce a player-driven quantitative tool to approximate the closest balancing configurations that could optimize player experience and satisfaction. Based on previous work, Pfau *et al.* [42] claim that conceptions about the definition of balancing often diverge between industry, academia, and gamers, and different balance design can lead to gamers' experiences that are worse than the actual imbalances. In the study, the authors collect game balancing concepts from industry and academia and introduce a player-driven approach to optimizing player experience and satisfaction. Politwoski *et al.* [43] defined an approach to integrate game testing to balance video games with autonomous agents. They propose a systematic way to assess whether a game is balanced by (i) comparing difficulty levels between game versions and game design problems and (ii) skill or luck demands.

Several studies targeted the assessment of engagement in video games [44]–[53]. These studies differ for the methodologies employed in the measurement and for the insights they seek to obtain about the players' experience.

The detection of facial expressions recognition is an alternative way of assessing the experience of the player, since it requires a much less invasive equipment (a camera) and, as previously explained, could be applied to mine information from gameplay videos. Moniaga *et al.* [54] present a dynamic game balancing system that adjusts game difficulty according to the players' facial expressions, enhancing the gaming experience. Differently from our work, their focus is to correlate four emotions/expressions (*i.e.,* angry, frustration, smirk, and smile) to the difficulty level of the game, increasing/decreasing it consequently. Engagement, on which we focus, is a wider concept, since a player may be not engaged while being frustrated for too difficult parts or bored for too easy ones. Kwon *et al.* [55] proposed a framework for automatically assessing the emotions of players during gameplay. They exploit facial expressions to identify users' emotions, including happiness, surprise, sadness, anger, disgust, and fear. Again, while emotions can be correlated to engagement (and, indeed, we use them as some of the features in our approach), they only tell part of the story, since users may experience some of these emotions (*e.g.,* sadness, anger) both while being engaged or not.

Chen *et al.* [13] presented an approach for estimating user engagement during game play based on facial features extracted from YouTube gameplay videos. Their approach is trained and evaluated on a dataset containing *perceived* engagement (the authors evaluate the engagement of the players). The FaceEngage dataset includes 783 clips of game videos from 25 players. The authors use a traditional machine learning techniques and deep learning models. In the first case, they use facial motion features, such as blink frequency, gaze, and head movements. These features are processed with traditional ML classifiers such as AdaBoost, SVM, k-NN and Random Forests. On the other hand, the deep learning approach uses a pre-trained convolutional neural network (CNN) to automatically extract face features. These features are then passed through a recurrent neural network (RNN) with an attention mechanism to learn temporal dependencies and predict engagement levels. The authors show that the second method outperforms the first one. In addition, the authors conducted experiments that demonstrate the robustness of their model against variations in video duration, game genres, and users. Specifically, these evaluations achieve 83.8% of accuracy for estimating engagement. Based on the dataset introduced by [13], Pan *et al.* [56] propose a multimodal deep learning model.

Their approach uses non-intrusive and non-restrictive multimodal data (facial, pixel, and sound modalities) to automatically estimate the engagement of game streamers. A limitation of the dataset used to train FaceEngage is that it relies on *perceived* engagement, manually assessed by external persons, rather than on the *direct* engagement experienced by the players.

VIII. Conclusion and Future work

We presented two studies. In the first, we compared three state-of-the-art approaches on a dataset we collected containing the labels representing the *direct* engagement of players. In the second, we interviewed two senior game developers to assess the industrial applicability of an engagement-detection approach. The results show that some models are highly accurate in the identification of *direct* low engagement events. Besides, the senior game developers we interviewed showed interest in the adoption of engagement detection approaches. Future research should be aimed at further improving the detection of low engagement by also considering contextual aspects of gameplay videos (such as the chat).

IX. Data Available

Our replication package [38] includes the implementation of the script we used to run the experiment, the dataset (not including the original recordings for privacy issues, but only the extracted features), and the scripts for analyzing data.

Acknowledgments

This work has been partially supported by the European Union - NextGenerationEU through the Italian Ministry of University and Research, Projects PRIN 2022 "DevProDev: Profiling Software Developers for Developer-Centered Recommender Systems", grant n. 2022S49T4W, CUP: H53D23003610001.

References

[1] Satista, "Global video game market value," https://www.statista.com/statistics/292056/video-game-market-value-worldwide/., 2023, [Online].

[2] ——, "Video gaming worldwide," https://www.statista.com/topics/1680/gaming/., 2022, [Online].

[3] C. Politowski, F. Petrillo, G. C. Ullmann, and Y.-G. Guéhéneuc, "Game industry problems: An extensive analysis of the gray literature," *Information and Software Technology*, vol. 134, p. 106538, 2021.

[4] EA, "Electronic arts playtesting," https://www.ea.com/playtesting/about., 2023, [Online].

[5] R. E. Santos, C. V. Magalhães, L. F. Capretz, J. S. Correia-Neto, F. Q. da Silva, and A. Saher, "Computer games are serious business and so is their quality: particularities of software testing in game development from the perspective of practitioners," in *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, 2018, pp. 1–10.

[6] E. Guglielmi, S. Scalabrino, G. Bavota, and R. Oliveto, "Using gameplay videos for detecting issues in video games," *Empirical Software Engineering (EMSE)*, p. To appear, 2023.

[7] A. Truelove, E. S. de Almeida, and I. Ahmed, "We'll fix it in post: what do bug fixes in video game update notes tell us?" in *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)*. IEEE, 2021, pp. 736–747.

[8] G. C. Team, "Best games to stream," https://ganknow.com/blog/best-games-to-stream/, 2023.

[9] "Youtube," https://www.youtube.com/, 2023.

[10] "Twitch," https://www.twitch.tv/, 2023.

[11] J. Whitehill, Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan, "The faces of engagement: Automatic recognition of student engagementfrom facial expressions," *IEEE Transactions on Affective Computing*, vol. 5, no. 1, pp. 86–98, 2014.

[12] Affectiva, "Affectiva Media Analytics for Entertainment Content Testing," https://tinyurl.com/47w2ezwb, 2023.

[13] X. Chen, L. Niu, A. Veeraraghavan, and A. Sabharwal, "Faceengage: Robust estimation of gameplay engagement from user-contributed (youtube) videos," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 651–665, 2019.

[14] T. Killedar, G. Suriya, P. Sharma, M. Rathor, and A. Gupta, "Fuzzy logic for video game engagement analysis using facial emotion recognition," in *2021 8th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2021, pp. 481–485.

[15] J. Shen, H. Yang, J. Li, and Z. Cheng, "Assessing learning engagement based on facial expression recognition in mooc's scenario," *Multimedia Systems*, pp. 1–10, 2022.

[16] Affectiva, "Affectiva," https://www.affectiva.com/, 2023.

[17] K. Doherty and G. Doherty, "Engagement in hci: conception, theory and measurement," *ACM Computing Surveys (CSUR)*, vol. 51, no. 5, pp. 1–39, 2018.

[18] E. Brown and P. Cairns, "A grounded investigation of game immersion," in *CHI'04 extended abstracts on Human factors in computing systems*, 2004, pp. 1297–1300.

[19] L. Ermi and F. Mäyrä, "Fundamental components of the gameplay experience: Analysing immersion." in *DiGRA Conference*. Citeseer, 2005.

[20] E. Adams and A. Rollings, *Fundamentals of game design (game design and development series)*. Prentice-Hall, Inc., 2006.

[21] E. Guardiola, "The gameplay loop: a player activity model for game design and analysis," in *Proceedings of the 13th International Conference on Advances in Computer Entertainment Technology*, 2016, pp. 1–7.

[22] I. Games, "Game Balance: A Pivotal Issue in Game Design," https://www.innovecsgames.com/blog/game-balance-a-critical-issue-in-designing-top-titles/, 2023.

[23] OpenAI, "Gym Documentation," https://www.gymlibrary.dev/, 2022.

[24] E. N. Wiebe, A. Lamb, M. Hardy, and D. Sharek, "Measuring engagement in video game-based environments: Investigation of the user engagement scale," *Computers in Human Behavior*, vol. 32, pp. 123–132, 2014.

[25] J. H. Cheong, "Affectiva-API-APP," https://github.com/cosanlab/affectiva-api-app, 2018.

[26] IMotion, "IMotion," https://imotions.com/, 2022.

[27] Affectiva, "Affectiva's Emotion Metrics," https://tinyurl.com/43f3avkm, 2023.

[28] J. H. Cheong, "Py-Feat," https://py-feat.org/pages/intro.html, 2022.

[29] J. H. Cheong, E. Jolly, T. Xie, S. Byrne, M. Kenney, and L. J. Chang, "Py-feat: Python facial expression analysis toolbox," *Affective Science*, vol. 4, no. 4, pp. 781–796, 2023.

[30] G. Laudato, S. Scalabrino, N. Novielli, F. Lanubile, and R. Oliveto, "Predicting bugs by monitoring developers during task execution," in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*. IEEE, 2023, pp. 1–13.

[31] A. P. Bradley, "The use of the area under the roc curve in the evaluation of machine learning algorithms," *Pattern recognition*, vol. 30, no. 7, pp. 1145–1159, 1997.

[32] L. Myers and M. J. Sirois, "Spearman correlation coefficients, differences between," *Encyclopedia of statistical sciences*, vol. 12, 2004.

[33] S. Gnanambal, M. Thangaraj, V. Meenatchi, and V. Gayathri, "Classification algorithms with attribute selection: an evaluation study using weka," *International Journal of Advanced Networking and Applications*, vol. 9, no. 6, pp. 3640–3644, 2018.

[34] D. Girardi, A. Ferrari, N. Novielli, P. Spoletini, D. Fucci, and T. Huichapa, "The way it makes you feel predicting users' engagement during interviews with biofeedback and supervised learning," in *2020 IEEE 28th International Requirements Engineering Conference (RE)*, 2020, pp. 32–43.

[35] ""cyberpunk 2077 - steam"," https://steamcommunity.com/app/1091500/discussions/7/4029094770944301584/, 2023.

[36] "Cyberpunk 2077 has flatlined-crashing explained," https://www.gamepressure.com/newsroom/2023-cyberpunk-2077-has-flatlined-20-crashing-explained/z9601b, 2023.

[37] "Cyberpunk 2077 three years later," https://www.grimdarkmagazine.com/review-cyberpunk-2077-three-years-later/, 2023.

[38] E. Guglielmi, G. Bavota, N. Novielli, R. Oliveto, and S. Scalabrino, “Replication package of ”automatically detecting low engagement events in video games”,” https://doi.org/10.6084/m9.figshare.23814783, 2023.

[39] J. Banyte and A. Gadeikiene, “The effect of consumer motivation to play games on video game-playing engagement,” *Procedia economics and finance*, vol. 26, pp. 505–514, 2015.

[40] J. Pfau, A. Liapis, G. N. Yannakakis, and R. Malaka, “Dungeons & replicants ii: automated game balancing across multiple difficulty dimensions via deep player behavior modeling,” *IEEE Transactions on Games*, vol. 15, no. 2, pp. 217–227, 2022.

[41] J. Pfau and M. Seif El-Nasr, “Balancing video games: A player-driven instrument,” in *Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, 2023, pp. 187–195.

[42] ——, “On video game balancing: Joining player-and data-driven analytics,” *ACM Games: Research and Practice*, vol. 2, no. 3, pp. 1–30, 2024.

[43] C. Politowski, F. Petrillo, G. ElBoussaidi, G. C. Ullmann, and Y.-G. Guéhéneuc, “Assessing video game balance using autonomous agents,” in *2023 IEEE/ACM 7th International Workshop on Games and Software Engineering (GAS)*. IEEE, 2023, pp. 25–32.

[44] E. N. Wiebe, A. Lamb, M. Hardy, and D. Sharek, “Measuring engagement in video game-based environments: Investigation of the user engagement scale,” *Computers in Human Behavior*, vol. 32, pp. 123–132, 2014.

[45] Y. Cao, “Understanding emotional experience in video games: A psychophysiological investigation,” in *Extended Abstracts of the 2022 Annual Symposium on Computer-Human Interaction in Play*. ACM, 2022.

[46] D. K. Mayes and J. E. Cotton, “Measuring engagement in video games: A questionnaire,” in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 45, no. 7. SAGE Publications Sage CA: Los Angeles, CA, 2001, pp. 692–696.

[47] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart, and J. N. Pidruzny, “The development of the game engagement questionnaire: A measure of engagement in video game-playing,” *Journal of experimental social psychology*, vol. 45, no. 4, pp. 624–634, 2009.

[48] R. Fridlund and E. Gustafsson, “Engagement in video games: A comparison between a linear and a branching narrative,” 2023.

[49] O. AlZoubi, B. AlMakhadmeh, M. Bani Yassein, and W. Mardini, “Detecting naturalistic expression of emotions using physiological signals while playing video games,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 2, pp. 1133–1146, 2023.

[50] J. Juvrud, G. Ansgariusson, P. Selleby, and M. Johansson, “Game or watch: The effect of interactivity on arousal and engagement in video game media,” *IEEE Transactions on Games*, vol. 14, no. 2, pp. 308–317, 2021.

[51] M. Granato, D. Gadia, D. Maggiorini, and L. A. Ripamonti, “Feature extraction and selection for real-time emotion recognition in video games players,” in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2018, pp. 717–724.

[52] W. Yang, M. Rifqi, C. Marsala, and A. Pinna, “Physiological-based emotion detection and recognition in a video game context,” in *2018 International joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[53] H. Schønau-Fog and T. Bjørner, “”sure, i would like to continue” a method for mapping the experience of engagement in video games,” *Bulletin of Science, Technology & Society*, vol. 32, no. 5, pp. 405–412, 2012.

[54] J. V. Moniaga, A. Chowanda, A. Prima, M. D. T. Rizqi *et al.*, “Facial expression recognition as dynamic game balancing system,” *Procedia Computer Science*, vol. 135, pp. 361–368, 2018.

[55] S. Kwon, J. Ahn, H. Choi, J. Jeon, D. Kim, H. Kim, and S. Kang, “Analytical framework for facial expression on game experience test,” *IEEE Access*, vol. 10, pp. 104 486–104 497, 2022.

[56] S. Pan, G. J. Xu, K. Guo, S. H. Park, and H. Ding, “Video-based engagement estimation of game streamers: An interpretable multimodal neural network approach,” *IEEE Transactions on Games*, 2023.