Review of Is it really fun? Detecting low engagement events in video games

Zhiao Wei
University of Waterloo
Master of Computer Science in Crysp Lab
z23wei@uwaterloo.ca

Review of Is it really fun? Detecting low engagement events in video games

Emanuela Guglielmi, Gabriele Bavota, Nicole Novielli, Rocco Oliveto, Simone Scalabrino, Proc. of the 22nd Intl. Conf. on Mining Software Repositories (MSR), Ottawa, April 2025.

Problem Being Solved

Video games need to be fun and engaging, but measuring engagement is hard. Previous research created automated tools that analyze facial expressions in streaming videos to detect when players aren't engaged. Previous tools were only tested on "perceived engagement" (what observers think players feel), not "real engagement" (what players actually feel).This matters because facial expressions can be misleading because players' laughing might be frustrated, not happy; a neutral face might mean focus, not boredom. Because games are interactive (unlike movies), engagement shows differently on faces. This paper tests whether these tools can detect real player-reported engagement and whether developers would actually use them.

New Idea

The key innovation is collecting the first real engagement dataset where 40 players self-reported their actual feelings (1-5 scale) every minute while playing 8 games, rather than having external observers guess engagement from facial expressions with producing 1,130 data points with ground truth labels. The authors compare three approaches and propose FFBD (Facial Features-Based Detection), which combines emotion features, facial action units (AUs), and head pose using Random Forest, achieving the best performance (F1=0.75, AUC=0.79) and demonstrating industrial viability through developer interviews.

Positive Points

1. Practical Application for game designer. Game developers will prepublish the game demo. 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.
2. Post-Release Monitoring. The tool allows developers to automatically detect problems even after the game is released. Developers can continuously monitor how engagement is impacted by game updates or plan future releases based on real player data
3. Fine-Grained Real-Time Detection. The tool provides minute-by-minute engagement assessment, allowing developers to pinpoint exact moments when engagement drops. This level of granularity is impossible with traditional playtesting methods that only collect feedback after entire gaming sessions.

Negative Points

1. Low Recall Rate. The model only achieves 41% recall for the low engagement class, meaning it misses more than half of the actual low engagement events. This could cause developers to overlook important game design problems.
2. Cannot Distinguish External Factors like chat interaction, donations. The approach might not identify low engagement events if the streamer is influenced by external factors (e.g., chat interaction, donations). The tool only analyzes facial expressions but cannot tell if low engagement is caused by the game itself or external distractions.
3. Limited Information Dimensions. Maybe add streamer's audio commentary, in-game actions, and chat interactions. The approach only uses facial features (emotions, expressions, head movements) but lacks other critical information such as streamer's audio commentary, in-game actions (e.g., death count, menu navigation time), and chat interactions, which could significantly improve detection accuracy. Narrative games.

Future Work

1. Multimodal Data Integration. Incorporate additional data modalities to improve detection accuracy. Future work should incorporate additional data modalities to improve detection accuracy. As suggested by the interviewed developers, integrating streamer audio commentary, chat interaction data, and in-game behavioral metrics (e.g., death count, time spent in menus) could significantly enhance the model's ability to distinguish game-related engagement issues from external distractions.
2. Expanding Dataset Scale. Collect data from a more diverse population. The current dataset includes only 40 players and 8 games. Future work should collect data from a more diverse population across different demographics, gaming expertise levels, and a wider variety of game genres (e.g., narrative-heavy games, multiplayer games) to improve generalizability.

Rating

Rating:(3/5): This is a high-quality empirical paper that pioneers real engagement detection with promising results (83.3% ranking correlation) and positive industry feedback, but is held back from top-tier status by low recall (41%), limited dataset scale (40 players, 8 games), and lack of causal insights into why engagement drops.

Summary of Discussion Points on class

(1) Is low engagement always a design flaw, or could it be player factors, external distractions, or intentional design (e.g., tension in horror games)?
I think low engagement does not always indicate a design flaw—some genres like horror games or hard-mode challenges intentionally create frustration or tension as part of meaningful engagement. The class argued that the paper's definition of "real engagement" remains vague and needs stronger psychological foundations to properly define and measure it. My proposed solution is to integrate multimodal data (audio, chat, game logs) to distinguish intentional design choices and external factors from genuine design flaws, and use aggregated patterns across multiple players rather than individual instances to identify systematic issues.
The class discussed some ways to improve accuracy through multi-modal integration, such as adding audio cues (streamer commentary), chat logs (viewer interaction), and in-game behavioral data (deaths, menu time, skip rates). Some participants suggested expanding the dataset to include more players, diverse demographics, and narrative-heavy games that the paper deliberately avoided. New insights highlighted that future research should combine psychological definitions, physiological signals, and multimodal data for more robust engagement assessment. The intersection of AI, psychology, and game design was identified as the most promising direction for future engagement-measurement research.

(2) Does "real engagement" measured with constant interruptions actually reflect uninterrupted gameplay engagement? Can a model trained on short, fragmented lab sessions generalize to long natural streaming sessions?
The training data comes from a highly artificial lab setting (40 students, 3-minute sessions, interrupted every minute for ratings, controlled lighting/equipment), but the tool is intended for analyzing natural streaming videos (continuous multi-hour gameplay, variable quality, diverse environments). I think interrupting players every minute to collect ratings may distort natural immersion and authenticity of engagement because deep engagement requires uninterrupted immersion—constant interruptions break flow state.
And Class students suggested expanding the dataset to include more players with diverse demographics, longer continuous gaming sessions, and different game genres (especially narrative-heavy games that were excluded). They also said that non-intrusive, continuous data collection methods could better capture genuine gameplay immersion without breaking flow state. Potential solutions include retrospective video review (players watch their own footage after and rate engagement), physiological signals (heart rate, eye tracking) that don't require interruptions, and semi-supervised learning using unlabeled streaming data to bridge the lab-to-reality gap.

(3) Should we prioritize recall (catching all issues, currently only 41%) or precision (reducing false alarms, currently 74.7%) for practical use? In industrial settings, which metric is more critical?
I pointed out that the model's recall for low-engagement events (41%) was relatively low, meaning many true low-engagement moments were missed—the best model achieves 74.7% precision but misses 59% of actual problems. The class discussed the trade-off between precision and recall, arguing that different jobs need different metrics: testing needs to catch everything (high recall), while monitoring needs accuracy (high precision). However, 41% recall is problematic for both use cases. The class suggested that companies might need to adjust thresholds depending on business goals and that the tool should offer adjustable modes allowing developers to tune sensitivity. The class also mentioned that game companies typically measure engagement through behavioral metrics rather than relying solely on facial analysis, suggesting this tool should be complementary rather than standalone.
And I like the professor's idea that elaborated on how adjusting model thresholds shifts the balance between recall (catching all low-engagement cases, important for pre-release testing) and precision (avoiding false alarms, important for post-release monitoring at scale), depending on the company's priorities and the cost of missing issues versus investigating false positives. New insights highlighted that the tool should offer adjustable sensitivity settings rather than a one-size-fits-all approach. Companies need human baseline comparisons—if human testers also only catch 40-50% of issues, then automated 41% recall might still be valuable as a complementary tool. Future work could implement human-in-the-loop active learning where developers provide feedback on detected events to continuously improve the model.