The Scream Stream: Multimodal Affect Analysis of Horror Game Spaces

Emmanouil Xylakis
Institute of Digital Games
University of Malta
Msida, Malta
emmanouil.xylakis@um.edu.mt

Antonios Liapis
Institute of Digital Games
University of Malta
Msida, Malta
antonios.liapis@um.edu.mt

Georgios N. Yannakakis
Institute of Digital Games
University of Malta
Msida, Malta
georgios.yannakakis@um.edu.mt

***Abstract*—Virtual environments allow us to study the impact of space on the emotional patterns of a user as they navigate through it. Similarly, digital games are capable of eliciting intense emotional responses from their players; moreso when the game is explicitly designed to do this, as in the Horror game genre. A growing body of literature has already explored the relationship between varying virtual space contexts and user emotion manifestation in horror games, often relying on physiological data or self-reports. In this paper, instead, we study players' emotion manifestations within this game genre. Specifically, we analyse facial expressions, voice signals, and verbal narration of YouTube streamers while playing the Horror game *Outlast*. We document the collection of the *Outlast Asylum Affect corpus* from in-the-wild videos, and its analysis into three different affect streams based on the streamer's speech and face camera data. These affect streams are juxtaposed with manually labelled gameplay and spatial transitions during the streamer's exploration of the virtual space of the *Asylum* map of the *Outlast* game. Results in terms of linear and non-linear relationships between captured emotions and the labelled features demonstrate the importance of a gameplay context when matching affect to level design parameters. This study is the first to leverage state-of-the-art pre-trained models to derive affect from streamers' facial expressions, voice levels, and utterances and opens up exciting avenues for future applications that treat streamers' affect manifestations as *in-the-wild* affect corpora.**

***Index Terms*—Video games, Survival horror, Level design, Virtual environments, Facial Expressions Analysis, Speech Analysis, Arousal, Surprise, Fear**

I. Introduction

The complex relationship between human emotions and physical space has become a compelling case for academic research, especially with recent advances in affective computing, interactive media, and virtual environments. Many projects have studied the affect-eliciting properties of space, using both interactive [1], [2] and non-interactive stimuli [3]–[5]. Video games offer an ideal medium for such studies [6] because they involve complex spatial navigation tasks while also being explicitly designed to trigger players' emotional responses—especially in horror games.

This paper introduces a new way of capturing emotion of players in video games, through their affect manifestations when streaming "let's play" gameplay footage on the YouTube video platform. "Let's play" videos are popular forms of crowdsourced content where a streamer plays the game while narrating their playthrough and interacting in real-time with an audience. We treat such "let's play" videos by popular YouTube streamers as *in-the-wild* affect data [7], and process them in terms of three moment-to-moment affect expression modalities: *facial* expression affect labels, *utterance* affect labels (what has been said) and *para-linguistic* affect ratings (how it has been said). The raw video data and processed affect streams compose the *Outlast Asylum Affect corpus* introduced in this paper which features 16 popular YouTube streamers playing the first map of the *Outlast* (Red Barrels, 2013) horror game. We juxtapose these affect streams with moment-to-moment changes in gameplay as well as spatial transitions during navigation of the game's virtual space. Our goal is to identify relationships between game design features and affect manifestations. We do this via linear relationships as agreements between affect changes and spatial or gameplay changes during room transitions in the game map, and via a Random Forest model that predicts affect changes. Results indicate that certain aspects of the gameplay context are important factors for emotion manifestation.

This study seeks to establish a more nuanced understanding of how game spaces can evoke specific emotional experiences, thereby contributing to our understanding of the design of spaces, experiences and games more broadly.

Fig. 1. One of the many jump scares of the *Outlast Asylum Affect Corpus*, depicting in-game visuals and face cam of YouTube user AnidaGaming. Image used with the streamer's permission.

This project has received funding from the European Union's Horizon 2020 programme (grant agreement No 951911).

979-8-3315-1645-1/24/$31.00 ©2024 IEEE

II. Impact of Video Game Spaces on Affect

Affect within video games is an extensively researched topic [6]: both game developers and researchers actively solicit (and analyze) player feedback regarding the various comprised facets [8] that make up a video game title. The emotional resonance of spaces in video games is studied mainly through the facet of level design [9], but ambience is also impacted by the game's sounds, visuals and gameplay [10].

Joosten *et al.* [11] explored the role of ambient illumination and player performance on eliciting different affective states in *Neverwinter Nights* (Bioware 2002). In their study, sixty participants first solved puzzles in virtual rooms with different lighting color settings, and then reported their arousal and valence. Results showed that red illumination led to higher arousal ratings, while yellow illumination led to higher valence ratings. In *The Underwood project*, McCall *et al.* [12] studied spatial uncertainty by employing an interactive virtual environment designed to elicit a predefined tension arc by manipulating light levels, scale of different 3D models, audio and interactivity. Participants' affect was captured during gameplay via physiological data (heart rate) and post-experience ratings of tension, i.e. whether each room was "frightening", "creepy" and "unpredictable". Results indicated that tension was primarily influenced by the presence of hiding places, nearby hostile agents, blocked paths, and darkness.

Horror games are especially powerful stimuli for studying affect, as they are explicitly designed to elicit players' (negative) emotions. Already in 2009, Niedenthal [13] highlighted the importance of light and darkness, vision obscurity and building scale in creating tension and uncertainty for the player in different horror games. To study the impact of Virtual Reality interactions on anxiety, Ferreira *et al.* [14] developed a custom virtual environment with multiple scenarios for different tension levels, following horror game design principles. Players' tension levels were captured during gameplay via three different types of biomarkers (electrocardiogram signals, electrodermal activity and respiration recordings), followed by a post-game questionnaire. Results showed that the scenario with the highest player tension had reduced lighting levels, flickering lights, and introduced the player to hostile agents. This study highlighted that environment parameters play an important role in manipulating player tension, but these parameters have to be viewed within the context of the game. Similarly, Graja *et al.* [15] tasked participants to play the popular horror game *P.T.* (Konami, 2014), capturing in-game recordings of galvanic skin response and post-game self-reports of tension. Game features that were investigated were changes in lighting, player actions and sound. Their results suggested that, despite few cases where in-game biomarkers matched self-reports, there were promising tendencies between events related to lighting and sound changes. The authors highlighted that the order in which effects (e.g. sound changes) were arranged had a strong impact on emotional responses.

The above works measure affect via both objective approaches (e.g. biomarkers) and subjective approaches (self-reports) in order to reveal the impact of level design on players' emotional experiences. In this study, instead, we rely solely on emotion manifestations (rather than intrusive sensors) and emotion labels derived from pre-trained models to capture affect. This may introduce unexpected biases compared to a ground truth derived from annotations, but is more ecologically valid and scales as in-the-wild data acquisition [7]. We follow the literature and explore both gameplay and spatial features to assess their impact on affect manifestations captured in the composite "let's play" video streams.

TABLE I
Outlast Asylum Affect Corpus properties

Outlast Assylum Affect Corpus	
Corpus video duration	$\approx$ 8.5 hours
Nr. of streamers	16
Nr. of videos	26
Nr. of labelled events	3125
Nr. of frames (1 sec sampling)	35283
Nr. of frames (avg. per Streamer)	2205
Extracted Affect (raw video data)	
Nr. of frames with Facial label (30 per sec)	962220
Nr. of snippets with a Vocal Arousal value (1 per sec)	32074
Nr. of phrases with Utterance emotion label	9878

III. Outlast Asylum Affect Corpus

The *Outlast Asylum Affect Corpus* contains approximately 8.5 hours of YouTube streams of different playthroughs on the same level (*Asylum*) of the horror game *Outlast*. We discuss the game and the level in Section III-A. All streams consist of a full-screen view of the game (containing its own audiovisual content) and a face camera overlay of the streamer along their own vocal narration of the playthrough (see Fig. 1). We use the streamer's face and audio to derive affective states (see Section III-D). while the first author manually labelled and tagged core gameplay and spatial features on each video (see Section III-C). Table I summarizes the properties of the *Outlast Asylum Affect Corpus*. The processed corpus, with derived data (affect signals and expert annotations) is made available[1] with links to the original YouTube videos.

A. Summary of the Outlast game and the Asylum level

Outlast is a horror game developed by Red Barrels and released for PC in 2013 and game consoles in 2014. In the game, the player takes the role of an investigative journalist and navigates a dilapidated psychiatric hospital overrun by homicidal patients. The game is played via a first-person camera perspective, and the player can move, jump, climb, crouch but not defend against or attack enemies. While encounters with enemies are generally sparse, the player can only outrun or hide from enemies. If a player dies, they respawn at the most recent checkpoint; checkpoints are hard-coded by the level design. The plot of *Outlast* is mostly conveyed through dialogue with the few sane non-player characters (NPCs) remaining in the hospital, from notes strewn around the hospital, or from

[1] https://osf.io/5jtx3/

the player's own camcorder recordings which trigger "self-reflection" as voice lines of the player's character. The virtual environment tends to be fairly dark, prompting the player to make use of the night vision capabilities of their camcorder, which has limited power and must be recharged with batteries. Overall, *Outlast* tends to rely on jump scares and audio cues to trigger strong, visceral reactions which do not last very long. However, the background audio and environment design (with blood and gore) is likely to keep players alert and aroused. Jump scares are usually interactive (i.e. the player can move away from them) or non-interactive cutscenes (where the game controls the player's actions for a short duration).

Asylum is the first level of *Outlast* and thus sets the mood and dangers of the game. The level includes 41 rooms across three floors. The player's trajectory, at least in the first half of the playthrough, is predetermined through locked doors and barricaded hallways. This design pattern is common in tutorial levels, allowing the game designers to control which parts of the mechanics or story are shown to the player and in which order. This is also convenient for our affect corpus, as the order of streamers' reactions is expected to match. Each floor contains diverse rooms, but the basement especially features many narrow and dark spaces. The *Asylum* level features two hostile NPCs, while many NPCs are neutral bystanders but contribute to the eeriness via audio cues and jump scares.

B. Properties of the Raw Video Data

The video data was collected via YouTube, in the form of 16 complete runs of the *Asylum* level of *Outlast* from 16 YouTube streamers. Streamers were chosen for their popularity: each channel has thousands of subscribers. Some runs through the *Asylum* level were split into separate videos (up to 4) which were merged in terms of their labelling (see Sections III-C and III-D) during data processing (see Section IV). Whether combining multiple videos or processing one video, segments introducing the game and the stream were removed. All processed data have a full-screen view of the game (containing its own audiovisual content) and a face camera overlay of the streamer along their own vocal narration of the playthrough (see Fig. 1). Complete playthroughs in the dataset lasted between 22 and 48 minutes.

C. Labeling of Gameplay and Spatial Features

Each video in the *Outlast Asylum Affect Corpus* was labelled by the first author in terms of timings when events occurred. Since this study is mostly interested in the impact of virtual space on affect, each playthrough is split according to the room the player is in (i.e. based on timings when the player entered and exited the room). Rooms are assigned eight *spatial* descriptors (see Table II) that, for the most part, describe the room size, its navigability (blocked paths), its contents (e.g. empty, neatly arranged furniture, or chaotically strewn debris and corpses), and the level and color of the illumination. Moreover, rooms may contain gameplay elements, or trigger new interactions. These *game* features are also labelled, as they affect the gameplay affordances of each room. Seven game features are labelled (see Table II), including the presence of hiding places, batteries, triggers that allow the player to continue their level traversal (e.g. keys, levers), notes, and events (e.g. interactive jump scares or audio cues) or non-interactive cutscenes in the room. We treat each room visit as one set of such values for the entirety of the room visit. We note that players may revisit rooms they were in before: if the conditions are different (e.g. lights that were previously on are now off, or the player already picked up the battery in this room) the features are labelled accordingly.

TABLE II
THE 15 FEATURES DESCRIBING SPATIAL AND GAME PROPERTIES AND THEIR VALUES. NUMBERS IN PARENTHESES ARE THE VALUES USED TO MEASURE DIFFERENCES BETWEEN ADJACENT ROOMS (SEE SECTION IV).

	Feature	Values
Spatial	Area size	Small (0), Medium (1), Large (2)
	Ceiling height	Low (0), Medium (1), High (2)
	Light contrast	None (0), Uneven (1), Even (2)
	Light levels	Dark (0), Dimly Lit (1), Bright (2)
	Light (color) temperature	Warm (0), Cold (1)
	Blocked path	False (0), True (1)
	Empty room	False (0), True (1)
	Interior arrangement	Chaotic (0), Ordered (1)
Game	Hiding place	False (0), True (1)
	Triggers present	False (0), True (1)
	Battery present	False (0), True (1)
	Note present	False (0), True (1)
	Cutscene	False (0), True (1)
	Event	False (0), True (1)

D. Generating Multimodal Labels of Affect

The available modalities in the *Outlast Asylum Affect Corpus* that capture the streamer's affective state are the streamer's facial expressions and their voice. We leverage established pre-trained models for capturing affect via facial expressions, and process both the utterances (i.e. what is said) and audio information of the voice (i.e. how it is said) for affect recognition. The resulting three data streams (see Fig. 2) are sampled at 1 Hz throughout the entire video, and capture specific emotional dimensions relevant to horror gameplay. Specifically, we are interested in *arousal* levels and, from categorical emotions [16], we are interested in *fear* and *surprise* as the most targeted by this type of game: fear from disturbing imagery such as blood and gore and surprise from jump scares. The raw data points with extracted affect per modality are found in Table I.

1) Facial Expression: For this modelling task, the streamer's face region on the video is cropped (see Fig. 1). Categorical emotions are assigned via Google's Mediapipe[2], using the 'efficient face' pre-trained model of Zhao *et al.* [17]. This model outputs probabilities within $[0, 1]$ for seven labels: anger, disgust, fear, joy, sadness, surprise, neutral. Since the model takes frames as input (with 30Hz sampling rate), we average the probabilities per second (1 Hz sampling rate) and retain probabilities for fear (F_f) and surprise (F_s).

2) Streamer's Voice (Para-linguistic Data): To analyze affect in the player's speech, we first isolate the streamers'

[2]https://developers.google.com/mediapipe

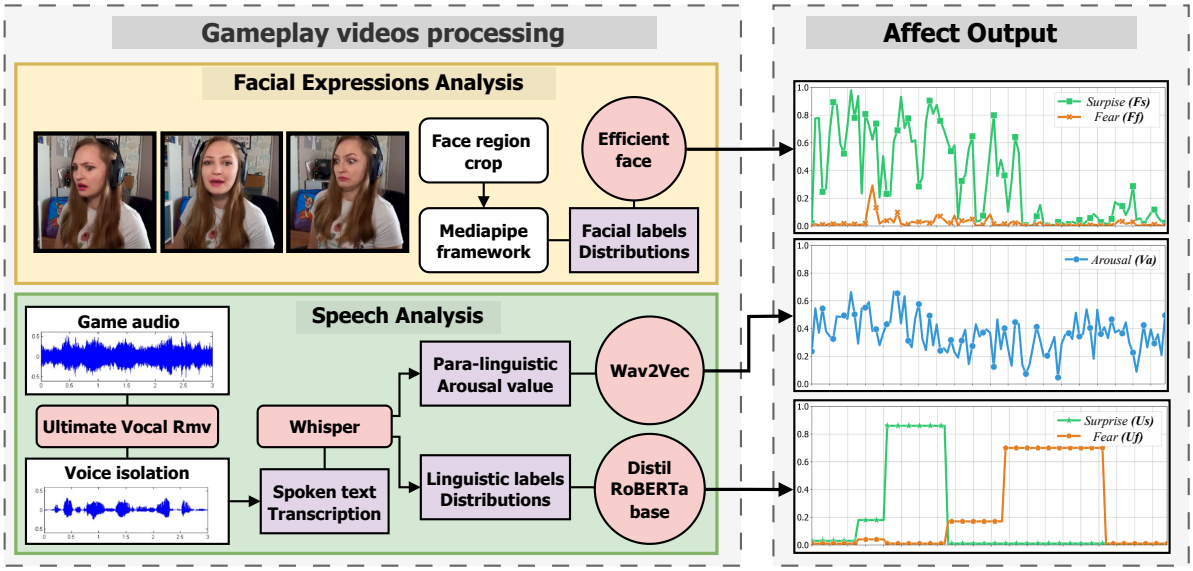

Fig. 2. Affect labels from different modalities of the streamer's face camera and audio, derived through pre-trained models. Face camera depictions are from YouTube user AnidaGaming; image used with the streamer's permission.

vocal cues from the game audio with the 'Ultimate Vocal Remover'[3] application [18]. We then apply Speech Emotion Recognition on the isolated vocal cues using Audeering's fine-tuned version of the 'wav2vec 2.0' model [19], [20]. This model outputs arousal, valence, and dominance values at 1 Hz sampling rate. We retain only the arousal values (V_a).

3) Streamer's Utterances: Unlike traditional players, streamers constantly converse with their audience. Thus, streamers' utterances can be information-rich, although the context of these utterances may differ from the current game context (such as discussing the streamer's day or commenting on something an audience member said). We use the isolated vocal cues from Section III-D2 and transcribe them as text using OpenAI's Whisper [21]. Then, the transcription is processed via the 'Emotion English DistilRoBERTa-base'[4] model which returns probabilities within $[0, 1]$ for seven labels (anger, disgust, fear, joy, sadness, surprise, neutral). As in Section III-D1, we aggregate the probabilities at 1 Hz sampling rate and retain only probabilities for fear (U_f) and surprise (U_s).

IV. Data Processing

In this paper, we focus on the impact that different rooms (and their properties) have on the player's affect. Due to this choice, we treat the duration where a player is in a specific room of the game level as a single time window and measure the affect manifestations within it. We view affect in two ways: in terms of its *mean value* while inside the room, and in terms of its *amplitude* (i.e. the maximum value of this affect signal within the room, minus its minimum value). Mean affect is an *absolute* measure of affect while amplitude is a *relative* measure of affect which is expected to be less prone to biases and, in our case, averaging artifacts due to differences in room visit durations [22]. In each visit to a specific room, we measure these two affect metrics for each of the five affect signals of different manifestations: fear and surprise of facial expressions, arousal of voice, fear and surprise of utterances. If the player re-visits the same room later, we create a new time window and calculate a new mean and amplitude per affect signal. We wish to match these affect metrics with properties of the room. Thus, we track the game and spatial properties of the room at the moment the player entered it. While some properties may change when the player enters the room (e.g. a cutscene may play, or the player may pick up the note), we only change these properties if the player enters the room again (if the room does not have a note present anymore).

With the above processing steps, we have a room's conditions (considered unchanging during the player's current visit) and an affect metric (mean or amplitude) for each affect signal. We follow an ordinal view of emotion and its triggers [23], and thus compare the affect between consecutive room visits in the playthrough. We measure whether the affect metric (mean or amplitude) changes in one room compared to the previous room. We classify each room transition as an *increase* in the affect metric (e.g. the mean affect in one room increases compared to the mean affect in the previous room) or as a *decrease*; we discard transitions where there is no discernible change in the affect metric. Following the literature [3], we only consider changes in the affect metric (increasing or decreasing) if their absolute difference (between rooms) is above a threshold ϵ. Based on best practices in the literature [24], [25], we use a threshold $\epsilon = 0.05$ for all affect metrics in every signal. With this threshold, the number of changes per signal are similar for both mean and amplitude.

We aim to match the characteristics of the two rooms with

[3]https://github.com/Anjok07/ultimatevocalremovergui

[4]https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/

increases or decreases in the affect metric. Since all properties in Table II are scalar, we note whether each property increases or decreases during a room transition. As an example, if a player moves from a brightly lit room with a note to a dark room with an event, the properties "light levels" and "note present" decrease while "event" increases during the room transition. We match these changes to affect in Section V.

V. Results

As discussed in Section III, we have collected a dataset of 16 playthroughs of the *Asylum* level from the *Outlast* game. This study investigates how room transitions impact the manifestations of emotion captured in para-linguistic, facial, and utterance data (see Section III-D). We follow two methods: one assuming a linear relationship between affect changes and changes in each room property (in Section V-A) and one assuming that all properties combined can predict changes in affect by training a machine learning model (in Section V-B).

A. Linear Relationships between Affect and Design Changes

The simplest way of finding a relationship between a room property changing and an affect metric changing is to see whether these changes *coincide* and are *in agreement*. To do this, we observe cases where both the affect metric increases and a room property increases, and mark it as an agreement; if the affect metric increases and the room property decreases (or vice versa) we mark it as a disagreement. In cases where only one of the two (affect or property) changes, we ignore this room transition altogether. We measure and report agreement ratio as the number of agreements between affect metric and room property aggregated across all 16 playthroughs, divided by the number of agreements and disagreements (i.e. when there was a definitive shift in both gameplay property and affect manifestation). Significance (at $p < 0.05$) is calculated with binomial testing [26] on this agreement ratio, assuming a 50% chance of agreement or disagreement.

Table III shows the agreement ratio for both changes in affect mean and changes in affect amplitude, aggregated from all 16 playthroughs in the dataset. We observe that changes in affect amplitude between rooms more often match changes in room features when transitioning from one room to the next. Specifically, when measuring changes in affect mean there are in total 10 significant agreements and 5 significant disagreements while for changes in affect amplitude there are 44 significant agreements and 17 significant disagreements. This deviation is surprising given the fact that the number of changes for mean and amplitude of the same signal are approximately the same. For mean affect, we observe that the presence of an event in a room coincides with increased arousal in the streamer's voice and fear in the streamer's utterances. This is not surprising, as these events are often jump scares. Interestingly, cutscenes (which are often also designed to be more elaborate jump scares) have a strong effect on mean affect changes but coincide with a drop in mean fear, likely due to the long duration of cutscene segments. Regarding spatial features and mean affect changes, we observe that more even illumination (light contrast) coincides with increased fear in streamers' utterances, but the reverse is true as lights get colder (increased color temperature). Warmer (yellow-ish) colors are often used for more dimly and sparsely lit rooms, which often include dark corners and jump scares. Bright cool light is often reserved for rooms where players must interact with the environment, e.g. read a note or use keys.

TABLE III
Agreement ratio between in-game properties of the game level and affect means or affect amplitude for different affect manifestations. We mark agreements as △ and disagreements as ▽ when statistically significant (above chance).

Changes in Affect Mean					
Feature	F_f	F_s	V_a	U_f	U_s
Area size	50%	56%△	50%	54%	51%
Ceiling height	53%	54%	50%	51%	54%
Light contrast	54%	50%	50%	62%△	50%
Light levels	56%△	46%	49%	49%	52%
Light temperature	50%	56%	46%	44%▽	52%
Blocked path	41%▽	52%	50%	52%	48%
Empty room	52%	58%△	48%	50%	47%
Interior arrangement	53%	55%	44%▽	59%△	50%
Hiding place	51%	55%	55%△	52%	47%
Triggers present	47%	51%	54%	45%	50%
Battery present	51%	53%	50%	51%	50%
Note present	49%	56%	56%△	47%	54%
Cutscene	43%▽	55%△	49%	45%▽	50%
Event	52%	51%	56%△	61%△	51%
Changes in Affect Amplitude					
Feature	F_f	F_s	V_a	U_f	U_s
Area size	54%△	51%	52%	52%	53%△
Ceiling height	53%	57%	57%	54%	58%△
Light contrast	54%△	52%	55%	50%	54%
Light levels	40%▽	40%▽	39%▽	38%▽	36%▽
Light temperature	69%△	77%△	72%△	69%△	71%△
Blocked path	48%	45%▽	43%▽	43%▽	45%▽
Empty room	34%▽	28%▽	33%▽	37%▽	29%▽
Interior arrangement	45%	42%▽	49%	50%	43%▽
Hiding place	74%△	80%△	74%△	72%△	73%△
Triggers present	47%	50%	44%▽	49%	47%
Battery present	64%△	70%△	67%△	64%△	68%△
Note present	54%△	57%△	60%△	59%△	60%△
Cutscene	95%△	91%△	77%△	90%△	87%△
Event	74%△	79%△	69%△	81%△	73%△

Changes in affect amplitude, as noted earlier, coincide more often with changes between adjacent rooms' features. This is in part due to the way amplitude is computed. Looking at the distance between lowest and highest state within a room, a jump scare would result in a high amplitude as the streamer would momentarily cry out and promptly return to a more neutral state; if traversing to the next room takes more than a few seconds, this temporary increase would not be very pronounced for mean affect (e.g. for V_a) but would be evident in affect amplitude. With this in mind, it is not surprising that both events (usually jump scares) and cutscenes result in increased amplitude for all affect signals. A more interesting finding is that the presence of notes or batteries also triggers increased amplitude across all signals. This is not surprising considering the "let's play" live commentary. When a note is found, it is immediately read out loud by the streamer: the text of the note is often creepy, which is then captured

in the streamers' utterances as they read it, but also on their expressions. Moreover, reading a note offers the streamer a pause from the game (as usually no-one is chasing them during those times) and allows the streamer to engage with the audience and discuss the game, leading to more cadence in voice and expressions. Furthermore, batteries and hiding places, which are essential to the player's in-game survival, are often placed near hostile NPCs. Therefore, finding a battery or a hiding place may signal to the streamer that danger is nearby, increasing anticipation registered via affect manifestations.

Far more spatial features seem to have an impact on affect amplitude than on affect mean, most notably the presence of an empty room or a blocked path and increased light levels (leading to decreased affect amplitude for all signals, except F_f for blocked path) and an increase in light color temperature (from warm to cold lights) which leads to increased amplitude on all affect signals. While these are interesting observations, and in general match expectations from the literature regarding e.g. room illumination [11], [15], it is difficult to estimate why such spatial features have such profound impact without considering other factors such as the co-occurrence of in-game events: the non-linear models of Section V-B which combine all features can perhaps address this limitation.

B. Training Random Forest Models to Predict Affect Changes

To assess how the combination of changes in room properties impact affect state transition as a whole, we train Random Forest (RF) classifiers with all property changes (degree of change) as input and the affect measure change (increase or decrease) as output. Given the small dataset for affect changes (between 679 and 1377 data points), we leverage RFs for their more robust performance on small datasets and on similar experiments for predicting affect change [27]. The input is the difference between properties of the final room versus those in the previous room (for the 15 features in Table II), while the desired output is whether there is an increase or a decrease in the affect metric in the final room compared to the previous room. We only use data for which there is a clear increase or decrease in the affect metric (with the threshold of $\epsilon = 0.05$), and ignore any room transitions where there is no change in the affect metric. Moreover, to balance the data we mirror the ordinal relationships: i.e. for each room transition, we produce two data points, one with the difference of the first room minus the second room and one with the difference of the second room minus the first room. We thus attain a baseline accuracy of 50% for every fold every fold (training, testing and hyperparameter tuning).

A leave-one-subject out cross validation protocol is followed for training and testing the RF classifier. Since we have 16 playthroughs of the *Asylum* level by 16 streamers, we reserve a single streamer's data for hyperparameter tuning[5] per affect signal, while the remaining 15 streamers are used for training and testing. Of these 15 streamers, 14 streamers are used for training while the remaining streamer is used for testing, repeatedly choosing a new streamer for testing (15 repetitions). This ensures that the data of this streamer (both the playthrough stimulus and the emotion manifestations) is unseen by the trained models. We repeat the process 15 times, selecting a new streamer for the test set each time. As RFs are stochastic, we repeat training per fold 5 times and the average RF statistics on the test set from all 75 trials (15 test sets $\times$ 5 repetitions) are shown in Table IV.

[5]The tuned hyperparmeters are: the number of trees used, maximum tree depth, minimum number of samples per leaf node and the minimum number of samples required to split a tree node.

TABLE IV
RANDOM FOREST ACCURACY ON THE TEST SET, AVERAGED FROM 75 TRIALS.

Changes in Affect Mean	
Affect	Accuracy
Fear of facial expressions (F_f)	65%
Surprise of facial expressions (F_s)	59%
Arousal of voice (V_a)	57%
Fear of utterances (U_f)	55%
Surprise of utterances (U_s)	55%
Changes in Affect Amplitude	
Affect	Accuracy
Fear of facial expressions (F_f)	77%
Surprise of facial expressions (F_s)	74%
Arousal of voice (V_a)	71%
Fear of utterances (U_f)	71%
Surprise of utterances (U_s)	67%

Table IV shows the accuracy of the trained RFs on the test set (the unseen playthrough of an unseen streamer), averaged from 5 repetitions of the leave-one-subject out cross validation protocol. As expected from observations in Section V-A, predicting increase or decrease in affect mean is more challenging (with best accuracies at 65% for F_f) compared to predicting changes in affect amplitude. This is largely due to the way the metric is computed, and the game genre which relies on short bursts of emotion via jump scares, than an issue of the training protocol. We note that overall, it is easier to predict emotion in facial expressions from the spatial and gameplay inputs we use. Best accuracies are achieved when predicting this modality, whether considering mean affect or amplitude. The most challenging to predict are the emotions (at least fear and surprise) of utterances. Overall, we observe that even with the simple room properties labelled by experts (which do not include, for example, visual decor or audio information) we can reach accuracies over 70% for several affect signals when considering their highs and lows (as amplitude) instead of, for example, the mean values throughout a room traversal. However, we note that impurity-based feature importance metrics indicate that the most dominant predictor for affect amplitude across signals is the presence of events—validating findings of Section V-A.

VI. DISCUSSION

This paper focused on a real-world case of emotional activities (gameplay) based on a commercial, popular horror game and the affect manifestations of professional YouTube streamers. Results indicate that, when viewing the traversal of the game's level architecture (and embedded narrative events)

there is a relationship between each room's properties and the players' emotion manifestations. Comparing the mean arousal, fear, or surprise between two consequent rooms makes for a challenging affect modelling task. On the other hand, the genre of horror games leads to many short bursts of emotion and thus comparing those bursts (via affect amplitude) between rooms leads to more accurate models with average test accuracies as high as 77% for fear in facial expressions (see Table IV). In terms of the insights such models offer, the choice of using amplitude as the affect metric expectedly offers limited design insights: for the most part, jump scares (tagged as "events") are the most efficient triggers for increased affect amplitude.

While results are promising given the challenging task of in-the-wild multimodal affect analysis, it is worth considering the complex nature of the stimulus and the limitations of our approach for parsing it. Leveraging expert labels of custom design features (see Table II) allows us to track properties that are deemed important in the literature, but also specific to the game (e.g. the presence of batteries). However, the fact that the annotation of design features was carried out by one individual (the first author) may be limiting. On the one hand, the criteria for this annotation are relatively objective (especially for some features such as "battery present") and thus inter-annotator disagreements are unlikely; validating our hypothesis through another annotator could enhance this corpus regardless. On the other hand, the long duration of gameplay videos (almost 9 hours in total) may cause annotation errors, especially when features change multiple times during a playthrough (e.g. due to the player respawning). In addition, some important features cannot be easily labelled manually: visual effects, background music, audio cues, and creepy iconography may impact emotion but are difficult to capture in timed labels. For the most part, audio cues coincided with jump scares and are included in the all-encompassing "event" tag. However, the nature or context of such audio cues is not fully captured. Future work could explore expanding the manual labels with outputs of pre-trained models for visuals, e.g. a vision transformer [28], or audio, e.g. the BEATS model [29]. Such inputs would likely hinder the explainability of the model, which was already problematic due to the composite nature of the stimulus [10].

We also note that this paper viewed only a subset of possible affect manifestations. While observing face cameras, voice, and utterances provides a holistic view of the streamer's affect state, we only processed fear, surprise and arousal predictions of pre-trained models in these modalities. Observing other dimensions (e.g. sadness or joy) in more experiments would dilute the findings of this study. Similarly, other ways of processing the affect signals beyond changes in mean and changes in amplitude—such as the gradient of affect within a time window [22]—did not result in very accurate models and offered limited insights. Future work could aggregate the affect data from different modalities into more concise metrics, such as fusing them into a singular—truly multimodal—affect construct (e.g. transforming categorical labels to dimensional affect data) rather than predict each signal separately. However, this would require a ground truth via third-person affect annotations which would likely also add reporting biases due to complex stimuli and an extensive corpus.

This paper proposes a way of mapping game design properties to affect manifestations of professional players and YouTube streamers. The benefit of this approach is that this data exists in the wild, and is easy to acquire. Even though the streamer community is not necessarily the most diverse, it is also easy to acquire a balanced corpus with sufficient search. The emotion manifestations are also of fairly high quality, given the fact that streamers learn to talk through their gameplay and emote (especially in horror games). There may be noise within such corpora: streamers may over-emote or trigger specific events in order to increase audience engagement and viewership. Such noise is expected and perhaps even more evident in other in-the-wild affect datasets [7]. Future work should further explore the potential of such high-quality stimuli (commercial games) and real-time emotions from players. On the one hand, a validation study regarding the output of pre-trained models compared to expert annotations of affect would assess the validity of our approach. On the other hand, testing the method proposed in this paper on other game genres beyond horror could also gauge its generalizability. Other games may trigger less evident emotion manifestations, or those manifestations could be due to cognitive processing of the game state versus the visceral reactions of *Outlast*. We consider the corpus of "let's play" videos, which combine gameplay footage and multimodal affect manifestations, a fertile ground for research on how visuals, audio, and—in this paper—level design [10] impact players' emotions.

VII. Conclusion

This paper proposed a method for capturing affect in gameplay videos that include both the voice and the face of the player as they narrate their experience. Deriving arousal, fear, and surprise via pre-trained models on the player's voice, utterances, and face camera, we collected an extensive dataset (almost 9 hours) of different YouTube streamers traversing the same map in the *Outlast* horror game. Experiments on the relationship between level design features (including architecture, illumination, pre-scripted events and gameplay affordances) and players' affect manifestations showcase the importance of jump scares in this particular genre. Moreover, this paper establishes a methodology for processing affect and level design based on the room that the player is in, using expert labels for both the room arrival times and its contents. Future work should explore automating the annotation process, or coupling expert labels with audiovisual information of the gameplay footage itself. The proposed method offers a new avenue for affective computing research based on in-the-wild but emotionally rich data in streamed gameplay with live commentary.

Ethical Statement

This paper performs experiments on a corpus of affect labels collected from public data available on YouTube. Since YouTube data is not public domain, the published *Outlast*

Asylum Affect Corpus contains derived data (affect signals) and expert annotations, with links to the original YouTube videos; this ensures that the content creators retain control over their content. To the best of our knowledge, this work has no negative or deceptive applications, and will not exacerbate existing privacy or discriminatory issues. Application of this methodology is not expected to disrupt the processes or livelihoods of YouTube streamers or the industry at large.

REFERENCES

[1] E. Xylakis, A. Najm, D. Michael-Grigoriou, A. Liapis, and G. N. Yannakakis, "Eliciting and annotating emotion in virtual spaces," in *Proceedings of the Education and Research in Computer Aided Architectural Design in Europe (eCAADe) Conference*, 2023.

[2] L. Gregorians, P. F. Velasco, F. Zisch, and H. J. Spiers, "Architectural experience: Clarifying its central components and their relation to core affect with a set of first-person-view videos," *Journal of Environmental Psychology*, vol. 82, 2022.

[3] E. Xylakis, A. Liapis, and G. N. Yannakakis, "Architectural form and affect: A spatiotemporal study of arousal," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2021.

[4] P. Lopes, A. Liapis, and G. N. Yannakakis, "Targeting horror via level and soundscape generation," in *Proceedings of the AAAI Artificial Intelligence for Interactive Digital Entertainment Conference*, 2015.

[5] A. Shemesh, R. Talmon, O. Karp, I. Amir, M. Bar, and Y. J. Grobman, "Affective response to architecture–investigating human reaction to spaces with different geometry," *Architectural Science Review*, vol. 60, no. 2, pp. 116–125, 2017.

[6] G. N. Yannakakis and D. Melhart, "Affective game computing: A survey," *Proceedings of the IEEE*, 2023.

[7] D. Kollias and S. Zafeiriou, "Aff-Wild2: Extending the Aff-Wild database for affect recognition," *arXiv preprint arXiv:1811.07770*, 2018.

[8] A. Liapis, G. N. Yannakakis, M. J. Nelson, M. Preuss, and R. Bidarra, "Orchestrating game generation," *IEEE Transactions on Games*, vol. 11, no. 1, pp. 48–68, 2018.

[9] C. W. Totten, *Architectural Approach to Level Design*. CRC Press, 2019.

[10] A. Liapis, G. N. Yannakakis, and J. Togelius, "Computational game creativity," in *Proceedings of the International Conference on Computational Creativity*, 2014.

[11] E. Joosten, G. Van Lankveld, and P. Spronck, "Influencing player emotions using colors," *Journal of Intelligent Computing*, vol. 3, no. 2, 2012.

[12] C. McCall, G. Schofield, D. Halgarth, G. Blyth, A. Laycock, and D. J. Palombo, "The underwood project: A virtual environment for eliciting ambiguous threat," *Behavior research methods*, vol. 55, 2022.

[13] S. Niedenthal, "Patterns of obscurity: Gothic setting and light in Resident Evil 4 and Silent Hill 2," *Horror video games: Essays on the fusion of fear and play*, pp. 168–180, 2009.

[14] M. Ferreira, A. Pinha, M. Fonseca, and P. Lopes, "Behind the door: Exploring horror VR game interaction and its influence on anxiety," in *Proceedings of the International Conference on the Foundations of Digital Games*, 2023.

[15] S. Graja, P. Lopes, and G. Chanel, "Impact of visual and sound orchestration on physiological arousal and tension in a horror game," *IEEE Transactions on Games*, vol. 13, no. 3, pp. 287–299, 2020.

[16] P. Ekman, "Argument for basic emotions," *Cognition and Emotion*, vol. 6, pp. 169–200, 1992.

[17] Z. Zhao, Q. Liu, and F. Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[18] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band DenseNets for audio source separation," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017, pp. 21–25.

[19] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: Closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, 2023.

[20] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the Neural Information Processing Systems Conference*, 2020.

[21] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the International Conference on Machine Learning*, 2023.

[22] P. Lopes, G. N. Yannakakis, and A. Liapis, "RankTrace: Relative and unbounded affect annotation," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2017.

[23] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, 2018.

[24] K. Makantasis, A. Liapis, and G. N. Yannakakis, "The pixels and sounds of emotion: General-purpose representations of arousal in games," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, 2023.

[25] K. Pinitas, D. Renaudie, M. Thomsen, M. Barthet, K. Makantasis, A. Liapis, and G. N. Yannakakis, "Predicting player engagement in Tom Clancy's The Division 2: A multimodal approach via pixels and gamepad actions," in *Proceedings of the International Conference on Multimodal Interaction*, 2023.

[26] D. Cramer, *Fundamental statistics for social research: step-by-step calculations and computer techniques using SPSS for Windows*. Routledge, 2003.

[27] D. Melhart, A. Liapis, and G. N. Yannakakis, "Towards general models of player experience: A study within genres," in *Proceedings of the IEEE Conference on Games*, 2021.

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[29] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, W. Che, X. Yu, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," in *Proceedings of the International Conference on Machine Learning*, 2023.

.