Varying the Context to Advance Affect Modelling: A Study on Game Engagement Prediction

Kosmas Pinitas, Nemanja Rasajski, Matthew Barthet, Maria Kaselimi
Konstantinos Makantasis, Antonios Liapis, Georgios N. Yannakakis

Institute of Digital Games, University of Malta, Msida, Malta.
Email: {kosmas.pinitas,matthew.barthet, maria.kaselimi,
konstantinos.makantasis, antonios.liapis, georgios.yannakakis}@um.edu.mt

***Abstract*—Affective computing faces a pressing challenge: the limited ability of affect models to generalise amidst varying contextual factors within the same task. While well recognised, this challenge persists due to the absence of suitable large-scale corpora with rich and diverse contextual information within a domain. To address this challenge, this paper introduces a *GameVibe*, a novel corpus explicitly tailored to confront the lack of contextual diversity. The affect corpus is sourced from 30 First Person Shooter (FPS) games, showcasing diverse game modes and designs within the same domain. The corpus comprises 2 hours of annotated gameplay videos with engagement levels annotated by a total of 20 participants in a time-continuous manner. Our preliminary analysis on this corpus sheds light on the complexity of generalising affect predictions across contextual variations in similar affective computing tasks. These initial findings serve as a catalyst for further research, inspiring deeper inquiries into this critical, yet understudied, aspect of affect modelling.**

***Index Terms*—affect modelling, domain generalisation, engagement, video games, FPS games**

I. Introduction

Domain generalisation refers to the process of constructing models capable of generalising amongst varying contextual factors within the same task. This process remains an unresolved challenge in affective computing (AC) and artificial intelligence (AI) at large [1]. Although several datasets have been introduced to combat this challenge in domains such as object recognition [2], [3], constructing generalisable models of affect still requires access to large corpora with diverse contexts and affect annotators. Digital games offer a compelling case for understanding engagement dynamics, due to their inherent complexity and interactive nature. Unlike traditional media, digital games offer a dynamic experience introducing a multitude of contextual factors that can influence engagement levels, such as game mechanics or visuals [4]. However, player affect modelling corpora feature at best few contexts (i.e. games), hindering research on domain generalisation [5]–[7].

Motivated by the lack of multimodal corpora for the study of affect dynamics across different contexts, this paper introduces *GameVibe*, a game affect corpus consisting of 30 First Person Shooter (FPS) games annotated by 20 participants in terms of engagement using the PAGAN annotation tool [8]; each video is annotated by 5 participants. The 30 games all belong in the same subdomain (FPS genre) but vary substantially in terms of game mode, audiovisual style, and mechanics. The *GameVibe* corpus is introduced for the study and analysis of *affective game computing* [4] across different games of the same genre. In this first paper featuring this corpus we attempt engagement modelling *within* the same game: training a model on different videos of one game and testing it on an unseen video of the same game. Hence, we test the *generalisability* of engagement models on unseen videos as annotated by unseen participants, not included in the training data. This paper gauges how easy it is to model engagement using audiovisual information from the gameplay footage alone. Inspired by previous experiments in this vein [9], [10], we assume that sufficient affect information is existent and interwoven within the *audiovisual gameplay footage*. We employ two pre-trained state-of-the-art Transformer encoders to create high-level representations of gameplay pixels and audio, which are then used to train models for predicting viewer engagement independently for each one of the 30 games. We then analyse the impact of audiovisual gameplay context on the performance of the obtained models.

This paper is novel in several ways. First, to the best of our knowledge, this is the first time a diverse corpus of popular and commercial-standard games has been used for the study of viewer engagement. Second, the paper presents a general-purpose methodology for modelling affect by fusing pixel and sound information processed via pre-trained Transformer architectures. Third, the initial results presented in this paper serve as the baseline for this new multimodal corpus, inspiring deeper inquiries into generalisable game affect models [4].

II. Background

This section reviews relevant work on multimodal affect models, and engagement as an affect dimension in particular.

A. Multimodal Affect Modelling

AC studies affective phenomena by developing computational models that can capture affect manifestations [11]. Since visual elements (such as colour and composition) can act as emotion elicitors, many affect models rely on such visual

This project has received funding from the European Union's Horizon 2020 programme (grant agreement No 101021714). Makantasis was supported by Project ERICA (GA: REP-2023-36) financed by the Malta Council for Science & Technology, for and on behalf of the Foundation for Science and Technology, through the FUSION: R&I Research Excellence Programme.

cues. Breuer and Kimmer [12] employed Convolutional Neural Networks (CNNs) for diverse facial expression recognition tasks, while Ng et al. [13] fine-tuned CNNs to detect emotions on smaller datasets. Recognising the potential of games to effectively trigger emotional responses, Makantasis et al. [10] trained CNNs to correlate gameplay footage with arousal, while Pinitas et al. [14] employed pre-trained Vision Transformers and neuroevolution to train a preference learner to predict arousal in gameplay footage from arcade games.

Beyond visual cues, the capacity of other input modalities as affect predictors, such as audio and physiology, has also been studied extensively [15]–[18]. Martinez *et al.* [19] pioneered the use of CNNs for detecting affect through physiological signals. In a similar vein, Makantasis *et al.* [9] used CNNs to model arousal based on raw video footage and sound frequencies while Zhang *et al.* [20] employed a Convolutional Long Short-Term Memory (LSTM) network and a 1D-CNN to extract spatio-temporal facial and bio-sensing features. Recently. Yang *et al.* [21] employed a novel low-dimensional cluster-based contrastive learning algorithm for emotion recognition in conversations, achieving state-of-the-art performance across different benchmarks.

This paper is motivated by the above studies, but focuses on *generalisable* multimodal affect modelling. The purpose of the *GameVibe* corpus is to investigate the degree to which multimodal information can be used to derive models of affect able to generalise well within the same task under varying contextual factors. While this paper evaluates generalisability within the same game, the format of the corpus allows for future research on generalisability across games.

B. Engagement Modelling

Engagement stands out as a crucial component in human-computer interaction (HCI), serving as a multifaceted construct including affective responses of the user [22]. A growing body of research focuses on modelling various aspects of user engagement. Indicatively, Dermouche and Pelachaud [23] leveraged facial expressions, head movements, and gaze cues to predict user engagement in real-time dyadic interactions using an LSTM. Ting *et al.* [24] modelled student engagement within virtual learning environments using Bayesian Networks. Recently, Pan *et al.* [25] proposed an interpretable CNN for estimating streamer engagement from videos.

Over the last few years, games have emerged as a compelling domain for HCI and AC research [4], [26], prompting a new research in engagement modelling within games. Notably, Melhart *et al.* [27] leveraged viewers' chat logs as a proxy for engagement, using a small neural network to predict moment-to-moment gameplay engagement solely based on game telemetry. Similarly, Xue *et al.* [28] introduced a Dynamic Difficulty Adjustment framework aimed at maximising player engagement, reflected in play time. Finally, Pinitas *et al.* [6] employed pre-trained models and time-conditioning to predict long-term engagement in the commercial shooter game *Tom Clancy's The Division 2* by Ubisoft.

TABLE I: Features of the *GameVibe* corpus

Number of Participants	20 (5 per session)
Number of Gameplay Videos	120 (30 per session)
Video database size	2 hours
Number of Elicitors	30 games
Gameplay video duration	1 minute
Annotation Perspective	Third-person
Annotation Type	Continuous unbounded
Affect Labels	Engagement

Similar to the above research, this paper studies engagement modelling within the genre of FPS games but focuses instead on the relationship between different game contexts and the validity of the engagement models obtained.

III. The *GameVibe* Corpus

The *GameVibe* corpus [29], which is available for download[1], consists of audiovisual footage of gameplay from 30 FPS games (detailed in Section III-A) annotated for viewer engagement by 20 participants in total (detailed in Section III-B). To derive models of engagement (see Section IV), we process both the annotation traces and the audiovisual data as described in Section III-C.

A. The Games

The *GameVibe* corpus consists of annotated gameplay videos from 30 dissimilar, popular commercial FPS games; see Fig. 1 and Table I. The selection of games and their corresponding gameplay videos was based on several criteria. Primarily, we aimed to encompass a broad spectrum of audiovisual stimuli for engagement annotation, incorporating diverse graphical styles (such as photo-realistic, retro, cartoon-like, etc.) and gameplay modes (including Battle Royale, single player, and Deathmatch, elaborated in Section V-C). Additionally, we ensured that the selected videos did not include any comments from players or users. Instead, they solely featured the sounds from the game itself. Finally, all videos were limited to a maximum of 15 seconds of non-gameplay content, such as cut scenes or transition animations.

B. The Corpus

Below we describe the processed multimodal data from the gameplay videos, and the affect annotation process.

1) Corpus Modalities: We consider the two available modalities of game context information: video frames and in-game audio. The former modality consists of a series of high-resolution and low-resolution videos of in-game footage sampled at 30Hz: normally 1280×720 pixels for more recent games and 541×650 for older games. The duration of each gameplay segment is 60 seconds. The auditory information is extracted from the video and consists of stereo sounds sampled at 44 kHz. In older games this usually corresponds to midi-type of background music; in more recent games the audio usually consist of dynamic sound environments that respond to player actions.

979-8-3315-1643-7/24/$31.00 ©2024 IEEE.

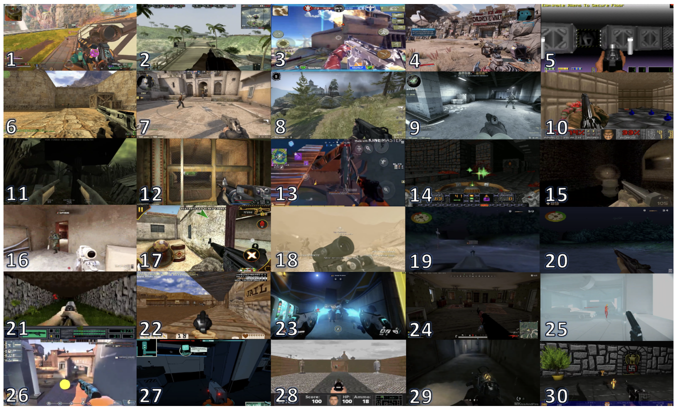


Fig. 1: Screenshots from the 30 different FPS games annotated for engagement. List of game titles: (1) Apex Legends; (2) Battlefield 1942; (3) Blitz Brigade; (4) Borderlands 3; (5) Corridor 7; (6) Counter Strike 2016; (7) Counter Strike 2018; (8) Counter Strike 2019; (9) Counter Strike: Global Offensive; (10) Doom; (11) Dusk; (12) Far Cry 1; (13) Fortnite; (14) Heretic; (15) Hrot; (16) Insurgency; (17) Modern Combat: Sandstorm; (18) Medal of Honor 2010; (19) Medal of Honor 1999; (20) Medal of Honor: Pacific Assault; (21) Operation Bodycount; (22) Outlaws; (23) Overwatch 2; (24) PUBG; (25) Superhot; (26) Team Fortress 2; (27) Void Bastards; (28) Wolfenstein 3D; (29) Wolfenstein New Order; (30) Wolfram Wolfenstein.

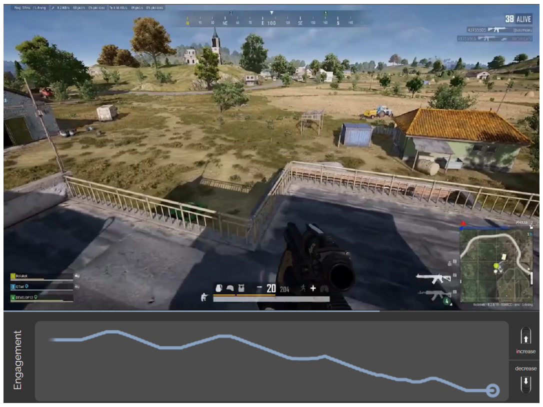


Fig. 2: Engagement annotation via the RankTrace [30] tool of the PAGAN platform [8]

2) Engagement Annotation: Gameplay videos were annotated in terms of the viewer's own engagement while watching, in a first-person manner. This annotation task was carried out across four different sessions with each session being annotated by 5 different (randomly assigned) participants. Participants were provided a concise definition of engagement as follows: *"A high level of engagement is associated with a feeling of tension, excitement, and readiness. A low level of engagement is associated with boredom, low interest, and disassociation with the game"*. Participants were then asked to annotate 30 short (1 minute) FPS gameplay videos, one video per game. The order of the 30 videos was randomised to minimise participants' habituation effects. We use different gameplay videos per session as the corpus is solicited to study the generalisability of affect models across varying contexts. Participants could pause the annotation process at any time by pausing the video itself. Each session lasted approximately 30 minutes per participant and thus, collectively, the 4 sessions (30 min each) offer 2 hours of annotated gameplay videos in total (see Table I).

Collecting reliable first-person engagement labels simultaneously for multiple games is impossible due to the high cognitive load of the task [5], [31]. Hence, we argue that offering short videos as stimuli for affect annotation provides a good tradeoff between annotation reliability and richness of engagement stimuli. Note that, given the 30 games in the corpus, each participant is required to be physically in the lab for 30 minutes which is the maximum time for engagement annotation in games as reported in the literature [6]. The 20 annotators involved in this study (5 per session) were affiliated with the *University of Malta* (research staff and graduate students). All annotators completed all annotation tasks in the same room, ensuring consistent room and lighting conditions. Additionally, the same machine and input/output devices, including a screen for visual stimuli, headphones for auditory stimuli, and a mouse scroll wheel were offered for the annotation task.

Data collection was carried out in two phases. First, each participant was asked to perform two simple yet controlled Quality Assurance (QA) tests (one visual and one auditory QA test) to ensure the annotators' reliability [32]. After the QA tests and a small break, the participant was asked to watch 30 randomly ordered 1 minute videos (one video per game) and annotate engagement in a continuous manner using the RankTrace [30] annotation tool of the PAGAN platform [8] (see Fig. 2). It is worth noting that care was taken to ensure data was collected and analysed respecting GDPR and ethical principles of AI and games research [33]. The core properties of the corpus are summarised in Table I.

C. Data Pre-Processing

The data preprocessing approach follows best practices for multimodal affect modelling as described in relevant studies [6], [10]. In particular, each video of a session is split into non-overlapping time windows of 3 seconds. The time windows of the input modalities (frames and sound) are shifted by 1 second to the annotation time window, accounting for the reaction time [34] between stimulus (gameplay) and annotation. Each time window consists of a sequence of frames and the corresponding sound.

For the visual modality we convert all videos into frames. As each video is 60 seconds long and sampled at 30Hz we end up with $1,800$ frames per video and 90 frames per time window. Several studies have shown that not all of these frames carry the same amount of information [35], and that many of them can be considered redundant; particularly, consecutive frames

[1] https://osf.io/p4ngx/

[36]. We follow a similar practice to reduce the computational load and we thus sample 16 RGB frames (downscaled to 224×224×3 pixels) in constant intervals within each time window. We preserve the number of colour channels under the assumption that transforming the frames to grayscale would omit vital visual information. For the sound modality, we extract audio clips within 3 second time windows and convert them from stereo to mono by averaging across the two channels (see Section IV-A).

When it comes to the engagement traces (i.e. 1 trace per annotator, 5 traces per gameplay video), we perform a min-max normalisation, transforming the unbounded engagement values to $[0, 1]$ on a per-trace basis. Each engagement annotation trace is similarly processed into time windows of 3 seconds, deriving an average engagement value for each time window per annotator.

IV. Modelling Engagement

As a preliminary experiment, this paper presents a study on engagement modelling within the same game, using an unseen video to test accuracy of a support vector machine model (see Section IV-C) trained on three videos from the same game. Importantly, both the participants and the video context are different between training and test sets (see Section V-A). Given the time-continuous traces of engagement, we treat this as a preference learning task (see Section IV-B). We leverage pre-trained models to derive the latent embeddings from audio and pixels of the video (see Section IV-A).

A. Representing Visual and Auditory Cues

We exploit pre-trained Transformers [37] to create high-level representations of audiovisual information. Transformers' multi-head self-attention modules can effectively capture the global spatio-temporal dependencies of the content. This paper uses two different pre-trained Transformer architectures with frozen parameters as feature extractors for the frame and audio modalities respectively.

The encoder which extracts features from *frames* is fed with 16 scaled-down RGB images as input and processes them via a Vision Transformer (ViT) [38] architecture that outputs 768 features. The ViT is pre-trained on Kinetics-400 [39] via Masked Video Distillation [40]. The encoder which extracts *audio* features, instead, processes monophonic sound via a Transformer architecture that outputs 527 multi-label class probabilities. This Transformer-based architecture, also called BEATs, is pre-trained on the AS-2M [41] dataset via Masked Audio Modelling. [42].

B. Preference Transformation

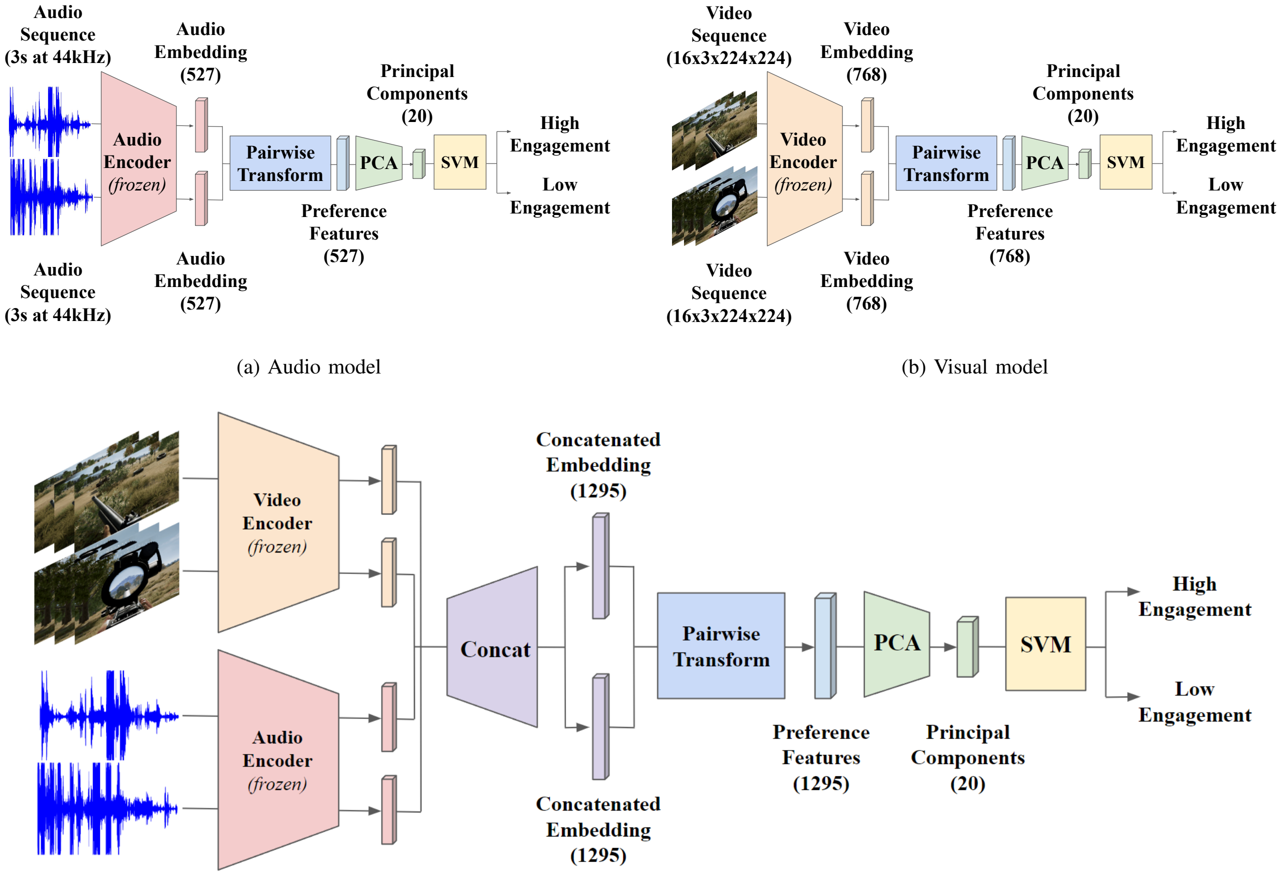
Preference Learning (PL) involves learning to rank data points, and is a suitable learning paradigm for any supervised task as long as the labels represent ordinal relationships [43]. PL is an appealing framework for affective computing since ordinal representations of emotion seem to offer more reliable and valid models of affect [44]. Drawing inspiration from previous work [45], [46] we formulate the engagement prediction task as a PL problem through a *pairwise transformation* process (see Fig. 3). In this paper, due to the short duration of the stimuli, we assume a global ranking of all time windows within the 1-minute video rather than only considering preferences between consecutive time windows [47]. Initially, time windows (20 per video) are selected (with their gameplay representations and corresponding 5 annotation values, one per participant) and pairs are formed (380 pairs in total per video). For each pair, the difference between the same annotator's engagement values is calculated. Each difference is assigned a preference: positive differences correspond to engagement increase and are labelled as 1, negative differences correspond to engagement decrease (-1), and no change as 0. Unlike some previous work which used an ambiguity threshold [9], [14], here we treat "no change" labels (0) as direct equality between engagement values of the two time windows. Subsequently, the agreement among all five annotators is tallied. If at least 4 out of 5 annotators agree on the preference label (i.e. if the annotations exhibit the same preference), the pair is retained; otherwise, it is discarded. Based on extensive PL work [6], [10], [44], we also discard pairs where the dominant label is "no change" as it demonstrates no clear preference. This process is repeated across all pairs within a video, filtering out labels where annotators disagree substantially in their assessment; this ensures the reliability of the annotated corpus. When it comes to the transformation of the audio and visual data, we compute the difference of latent representations for each pair of time windows. Finally, we form the same pairs of time windows in reverse order, yielding balanced preference datasets of approximately 514 samples per game, on average.

Since the latent embeddings of the audio or visual data are fairly large for the size of the training dataset (from 527 up to $1,295$ parameters, see Fig. 3), we use dimensionality reduction techniques before processing the data for preference learning (see Section IV-C). Specifically, we apply Principal Component Analysis (PCA) [48] (see Fig. 3) on the latent embeddings to produce a 20-dimensional feature vector used as input for training. The same training set employed to train the preference learner (see Section V-A) is used to train the PCA module, and the test data is projected onto the same 20-dimensional space.

C. Engagement Predictor

The 20 principal components extracted from PCA module serve as the input of the engagement predictor. We explored several methods and architectures for the downstream task of modelling engagement. Due to the limited volume of data available per game, we report results employing RankSVM [49], a widely used ranking method based on Support Vector Machines (SVM). We used a simple SVM with an RBF kernel, regularisation parameter $C = 1$ and coefficient $\gamma = 5\%$ of the data variance, trained on top of the extracted features to classify the preference labels (see Fig. 3).

979-8-3315-1643-7/24/$31.00 ©2024 IEEE.



(a) Audio model

(b) Visual model

(c) Fusion model

Fig. 3: The three model architectures employed for engagement preference learning, using pixel information, sound information, or both modalities.

V. Results

This section presents the results from the engagement modelling task (on a per-game basis) described in Section V-A.

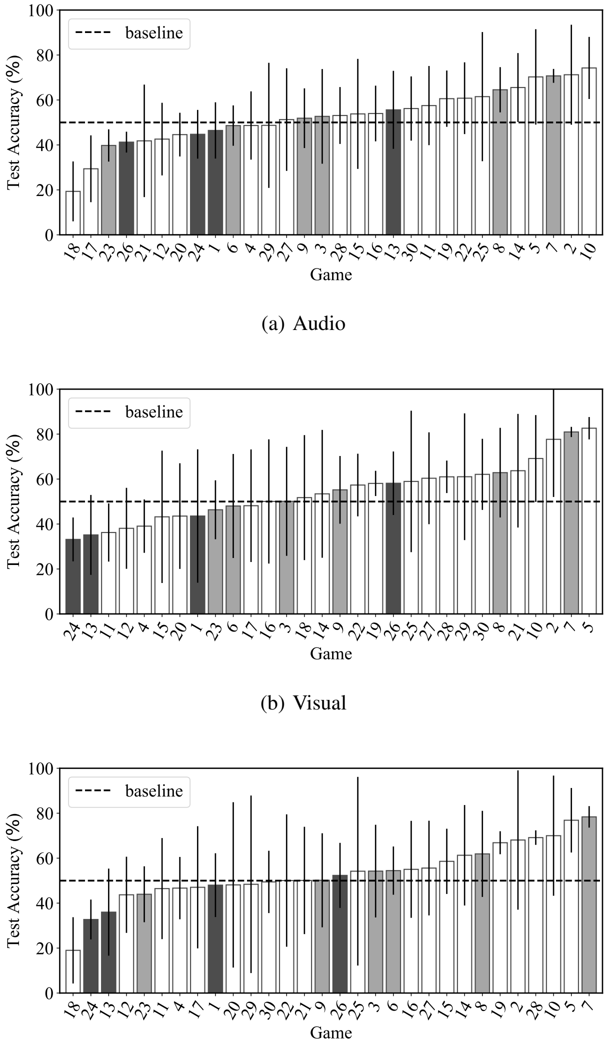
A. Experimental Protocol

Performance of engagement across each game independently is assessed via a demanding leave-one-session-out cross-validation strategy. We split the data into training and test sets, ensuring that data in each set belong to different sessions and thereby vary in terms of context and annotators (see Section III-B2). The hyperparameters of PCA and SVM were optimised using a grid-search protocol. For PCA, we determined that 20 principal components explained more than 95% of the variance in the training set, after testing between 10 and 100 components in steps of 10. For SVM, we considered C values ranging from 0.1 to 1 in steps of 0.1, ultimately selecting $C = 1$ as the best fit for the training data. The parameter $\gamma = 5\%$ corresponds to the inverse value of the number of PCA components. The performance of models is evaluated in terms of accuracy; baseline performance is 50% due to the pairwise transformation followed (see Section IV-B). Statistical significance is established via the 95% confidence interval on the Student's t-distribution due to the limited number of samples (4 folds) and the deterministic nature of SVM [50].

B. Engagement Models Per Game

Figure 4 shows the average test accuracy of the models across the 30 games of the *GameVibe* corpus. Evidently, models trained solely on the audio modality perform poorly, with 18 out of 30 games above the baseline (4 games are significantly better) and 9 games yielding an accuracy above 60%. The best accuracy obtained is 74% for the *Doom* game. Surprisingly, the accuracy of audio-only models is very low for *Modern Combat: Sandstorm* (29%) and *Medal of Honor 2010* (19%). Models trained on frames alone perform slightly better with 19 out 30 games above the baseline (5 games are significantly better). Out of those, 10 games yield accuracy values higher than 60%. The best accuracy (83%) is obtained for the *Corridor 7* game. Training on the fused audiovisual modalities results in test accuracy values between the two unimodal experiments: 18 out of 30 games mark higher average accuracy values than the baseline (4 games significantly better) with 9 games above 60%. The best accuracy for the bimodal input is *Counter Strike 2018* game, with 78.4% test accuracy.

979-8-3315-1643-7/24/$31.00 ©2024 IEEE.



(a) Audio

(b) Visual

(c) Bimodal Fusion

Fig. 4: Average test accuracy for engagement preference. Error bars depict 95% confidence intervals. The *Single Player*, *Deathmatch* and *Battle Royale* games are depicted, respectively, as white, grey and black bars. The dotted line corresponds to the majority class baseline. The x-axis labels correspond to the game IDs displayed in Fig. 1.

To examine how the participants' annotation patterns impacted the models, we calculate the dominant label (engagement increasing or decreasing) across all time windows and videos per game (i.e. 20 annotation traces per game, 5 per video). We expect that game datasets where one engagement relationship is more dominant leads to more reliable models, despite the efforts taken to balance the dataset by reversing the order of the pairs (see Section IV-B). Indeed the Pearson's correlation coefficient between the dominance of one label over the other and the test accuracy ($\rho = 0.23$) is statistically significant ($p < 0.05$). Figure 5 shows the tradeoff between the dominance of one label and model accuracy, indicating the majority label across all videos per game.

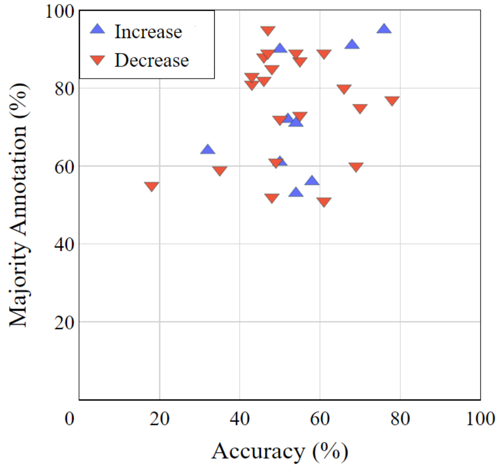


Fig. 5: Scatter plot of the majority annotation trend and the average test accuracy of the fusion model across all 30 games.

C. Engagement Modelling Across Game Modes

We examine the predictive power of our engagement models across different contexts, specifically three game modes. In *Single Player* games the gameplay often revolves around completing objectives or progressing through a narrative storyline. *Deathmatch* games are multi-player games where the primary objective is to eliminate opponents and score the most kills within a set time limit. Finally, *Battle Royale* games feature many players competing in a shrinking play area; players must survive by eliminating their opponents while scavenging for weapons, items, and resources. The *GameVibe* corpus comprises of 20 *Single Player* games, 6 *Deathmatch* games and 4 *Battle Royale* games.

Figure 6 illustrates the average performance of the games in a specific game mode. While *Deathmatch* and *Single Player* games can be predicted from gameplay frames with an average accuracy of 57% and 56%, respectively, *Battle Royale* games mark a 42% accuracy on average, way below the baseline. We contend that *Battle Royale* games feature many unpredictable variables, but also slower and more calculated gameplay; this makes the task of engagement annotation itself more challenging based on 1-minute stimuli.

VI. Discussion

This paper introduced a large corpus of diverse stimuli (as FPS games) and a total of 600 annotation traces of engagement on 120 gameplay videos. Based on a preliminary analysis using outputs of pre-trained models of visuals and audio, we observe that achieving an accurate prediction of engagement using visuals and audio alone is challenging. Visual inputs are slightly better predictors than audio input (with a relative increase of 8% in test accuracy values averaged across games), while their fusion performs somewhere in-between. An analysis of the original stimuli indicates that

979-8-3315-1643-7/24/$31.00 ©2024 IEEE.

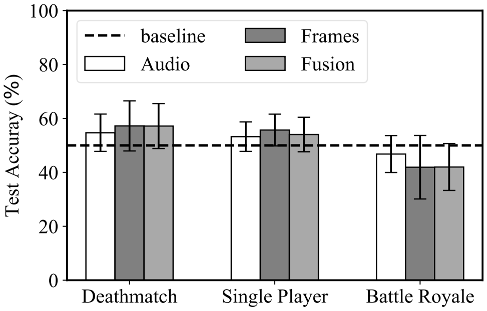


Fig. 6: Average test accuracy per game mode. Error bars depict 95% confidence intervals.

models' accuracy is impacted by both gameplay mode and annotation trends.

As expected, the variety of the corpus collected (with 30 games published between 1992 and 2023) makes the engagement modelling task challenging. While the use of pre-trained Transformer-based models offers a quick way of deriving latent vectors, these models are not trained on similar corpora and are expected to give better embeddings on more realistic audiovisual depictions. Exploring different pre-trained models, or fine-tuning these models [51] to more contextual corpora (e.g. older computer graphics and sounds), may lead to more nuanced results. On the other hand, modern games with more realistic audiovisual styles tend to have more nuanced, complex gameplay (e.g. the *Battle Royale* subgenre) which makes the actual engagement annotation task more difficult and thus leads to larger deviations of the affective ground truth. Future work should explore leveraging more annotated videos per game or providing longer stimuli, especially for modern games, in order to improve the within-game context variation of this corpus. This is especially true given the insights about videos with mostly descending engagement, as future work could explore the temporal and causal sequence [52] of events and their impact on engagement.

The extensive *GameVibe* corpus collected and reported in this paper is intended to address the generalisation challenge in AC by providing high-quality,engaging, yet diverse stimuli and their annotations. In the preliminary analysis of this paper, we focus on assessing how easy it is to predict engagement of an *unseen* video annotated by *unseen* participants but within the same game. Given that each video shows different gameplay contexts and, often, new game levels or pre-scripted events, this already tests generalisation of engagement modelling to some degree. Future work, however, should explore the limits of generalisation by testing how well-performing engagement models perform on entirely unseen games in a zero-shot manner. More ambitiously, generalisation can be tested by training models on multiple games and testing them on unseen games, similarly to [53]; this could shed light on the impact of gameplay mode as discussed in Section V-C. Beyond this corpus, we hope that this paper highlights the issue of generalisation within affect modelling and opens up a broader avenue of research using (diverse) games as affect elicitors. Such research would be fundamental for games user research [54], human-computer interaction, and AC at large.

VII. Conclusions

This paper presented the *GameVibe* corpus, designed to facilitate research on domain generalisation in affect modelling. We focus on engagement modelling in the context of a game genre, namely 30 dissimilar FPS games. For this paper, we did not explore generalisability across games but we tested both unimodal (audio and visual) and bimodal approaches using pre-trained Transformer-based models for predicting engagement in unseen footage of the same game the model is trained on. Our findings demonstrate varying degrees of success across different games, highlighting the complexity of modelling engagement across diverse gaming experiences. The study also emphasised the impact of game modes, suggesting that the nature of player interactions impacts the predictability of viewer engagement. Finally, our analysis reveals the influence of gameplay context on both model predictions and annotator behaviours. We hope that the *GameVibe* corpus and the preliminary engagement modelling experiments on a per-game basis reported in this paper will expedite further research at the crossroads of AC and domain generalisation.

Ethical Impact Statement

This paper presents a dataset of affect annotations collected from participants in a laboratory setting. Participants provided informed consent for the data collection process, and all personally identifiable information was removed from the dataset. The protocol was approved by the University Research Ethics Committee of the University of Malta. The dataset is made publicly available to support further studies and scientific reproducibility. To the best of our knowledge, this work does not pose a significant risk of being used for negative or deceptive applications and does not exacerbate existing privacy or discriminatory issues.

References

[1] K. Zhou, Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy, "Domain generalization: A survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4396–4415, 2022.

[2] D. Li, Y. Yang, Y.-Z. Song, and T. M. Hospedales, "Deeper, broader and artier domain generalization," in *Proc. of the IEEE Intl. Conf. on computer vision*, 2017, pp. 5542–5550.

[3] R. A. Bafghi and D. Gurari, "A new dataset based on images taken by blind people for testing the robustness of image classification models trained for imagenet categories," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 16 261–16 270.

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

[5] D. Melhart, A. Liapis, and G. N. Yannakakis, "The Arousal video Game AnnotatIoN (AGAIN) dataset," *IEEE Trans. on Affective Computing*, vol. 13, no. 4, pp. 2171–2184, 2022.

[6] 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 *Proc. of the Intl. Conf. on Multimodal Interaction*, 2023, pp. 488–497.

979-8-3315-1643-7/24/$31.00 ©2024 IEEE.

[7] K. Karpouzis, G. N. Yannakakis, N. Shaker, and S. Asteriadis, "The platformer experience dataset," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 712–718.

[8] D. Melhart, A. Liapis, and G. N. Yannakakis, "PAGAN: Video affect annotation made easy," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2019, pp. 130–136.

[9] K. Makantasis, A. Liapis, and G. N. Yannakakis, "The pixels and sounds of emotion: General-purpose representations of arousal in games," *IEEE Trans. on Affective Computing*, 2021.

[10] ——, "From pixels to affect: A study on games and player experience," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*, 2019.

[11] R. W. Picard, *Affective computing*. MIT press, 2000.

[12] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[13] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proc. of the Intl. Conf. on multimodal interaction*, 2015, pp. 443–449.

[14] K. Pinitas, K. Makantasis, A. Liapis, and G. N. Yannakakis, "RankNEAT: outperforming stochastic gradient search in preference learning tasks," in *Proc. of the Genetic and Evolutionary Computation Conf.*, 2022, pp. 1084–1092.

[15] S. M. S. A. Abdullah, S. Y. A. Ameen, M. A. Sadeeq, and S. Zeebaree, "Multimodal emotion recognition using deep learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 02, pp. 52–58, 2021.

[16] N. Sebe, I. Cohen, and T. S. Huang, "Multimodal emotion recognition," in *Handbook of pattern recognition and computer vision*. World Scientific, 2005, pp. 387–409.

[17] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[18] E. Lieskovská, M. Jakubec, R. Jarina, and M. Chmulík, "A review on speech emotion recognition using deep learning and attention mechanism," *Electronics*, vol. 10, no. 10, p. 1163, 2021.

[19] H. P. Martinez, Y. Bengio, and G. N. Yannakakis, "Learning deep physiological models of affect," *IEEE Computational intelligence magazine*, vol. 8, no. 2, pp. 20–33, 2013.

[20] Y. Zhang, M. Z. Hossain, and S. Rahman, "Deepvanet: A deep end-to-end network for multi-modal emotion recognition," in *Procceedings of the 18th Intl. Conf. on Human-Computer Interaction (INTERACT)*, 2021, p. 227–237.

[21] K. Yang, T. Zhang, H. Alhuzali, and S. Ananiadou, "Cluster-level contrastive learning for emotion recognition in conversations," *IEEE Trans. on Affective Computing*, 2023.

[22] J. J. Appleton, S. L. Christenson, D. Kim, and A. L. Reschly, "Measuring cognitive and psychological engagement: Validation of the student engagement instrument," *Journal of school psychology*, vol. 44, no. 5, pp. 427–445, 2006.

[23] S. Dermouche and C. Pelachaud, "Engagement modeling in dyadic interaction," in *Proc. of the Intl. Conf. on Multimodal Interaction*, 2019, pp. 440–445.

[24] C.-Y. Ting, W.-N. Cheah, and C. C. Ho, "Student engagement modeling using bayesian networks," in *Proc. of the IEEE Intl. Conf. on Systems, Man, and Cybernetics*, 2013, pp. 2939–2944.

[25] 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 Trans. on Games*, 2023.

[26] G. N. Yannakakis and J. Togelius, *Artificial intelligence and games*. Springer, 2018, vol. 2.

[27] D. Melhart, D. Gravina, and G. N. Yannakakis, "Moment-to-moment engagement prediction through the eyes of the observer: Pubg streaming on twitch," in *Proc. of the Intl. Conf. on the Foundations of Digital Games*, 2020, pp. 1–10.

[28] S. Xue, M. Wu, J. Kolen, N. Aghdaie, and K. A. Zaman, "Dynamic difficulty adjustment for maximized engagement in digital games," in *Proc. of the Intl. Conf. on World Wide Web Companion*, 2017, pp. 465–471.

[29] M. Barthet, M. Kaselimi, K. Pinitas, K. Makantasis, A. Liapis, and G. N. Yannakakis, "Gamevibe: A multimodal affective game corpus," *arXiv preprint arXiv:2407.12787*, 2024.

[30] P. Lopes, G. N. Yannakakis, and A. Liapis, "RankTrace: Relative and unbounded affect annotation," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*, 2017, pp. 158–163.

[31] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *Proc. of the IEEE Intl. Conf. and Workshops on Automatic Face and Gesture Recognition*, 2013.

[32] M. Barthet, C. Trivedi, K. Pinitas, E. Xylakis, K. Makantasis, A. Liapis, and G. N. Yannakakis, "Knowing your annotator: Rapidly testing the reliability of affect annotation," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction Workshops and Demos*, 2023.

[33] D. Melhart, J. Togelius, B. Mikkelsen, C. Holmgård, and G. N. Yannakakis, "The ethics of ai in games," *IEEE Trans. on Affective Computing*, 2023.

[34] S. Mariooryad and C. Busso, "Analysis and compensation of the reaction lag of evaluators in continuous emotional annotations," in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction (ACII)*, 2013, pp. 85–90.

[35] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 436–450, 2012.

[36] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[38] 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.

[39] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The Kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[40] R. Wang, D. Chen, Z. Wu, Y. Chen, X. Dai, M. Liu, L. Yuan, and Y.-G. Jiang, "Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2023, pp. 6312–6322.

[41] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. of the IEEE Intl. Conf. on acoustics, speech and signal processing*, 2017, pp. 776–780.

[42] S. Chen, Y. Wu, C. Wang, S. Liu, D. Tompkins, Z. Chen, and F. Wei, "BEATs: Audio pre-training with acoustic tokenizers," *arXiv preprint arXiv:2212.09058*, 2022.

[43] J. Fürnkranz and E. Hüllermeier, "Pairwise preference learning and ranking," in *Proc. of the European Conf. on Machine Learning*. Springer, 2003, pp. 145–156.

[44] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Trans. on Affective Computing*, vol. 12, no. 1, pp. 16–35, 2018.

[45] A. R. Naini and C. Busso, "Preference learning labels by anchoring on consecutive annotations," in *Proc. of INTERSPEECH Conf.*, 2023.

[46] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Proc. of INTERSPEECH Conf.*, 2018.

[47] E. Xylakis, A. Liapis, and G. N. Yannakakis, "Architectural form and affect: A spatiotemporal study of arousal," in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2021.

[48] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (PCA)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[49] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. of the eighth ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, 2002, pp. 133–142.

[50] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer New York Inc., 2001.

[51] X. He, C. Li, P. Zhang, J. Yang, and X. E. Wang, "Parameter-efficient fine-tuning for vision transformers," *arXiv preprint arXiv:2203.16329*, vol. 3, 2022.

[52] K. Makantasis, K. Pinitas, A. Liapis, and G. N. Yannakakis, "The invariant ground truth of affect," in *Proc. of the ACII Workshop on What's Next in Affect Modeling?*, 2022.

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

979-8-3315-1643-7/24/$31.00 ©2024 IEEE.

[54] A. Drachen, L. E. Nacke, and P. Mirza-Babaei, *Games User Research*. Oxford University Press, 2018.