The Pixels and Sounds of Emotion: General-Purpose Representations of Arousal in Games

Konstantinos Makantasis, Antonios Liapis *Member, IEEE*, and
Georgios N. Yannakakis, *Senior Member, IEEE*

Abstract—What if emotion could be captured in a general and subject-agnostic fashion? Is it possible, for instance, to design general-purpose representations that detect affect solely from the pixels and audio of a human-computer interaction video? In this paper we address the above questions by evaluating the capacity of deep learned representations to predict affect by relying only on audiovisual information of videos. We assume that the pixels and audio of an interactive session embed the necessary information required to detect affect. We test our hypothesis in the domain of digital games and evaluate the degree to which deep classifiers and deep preference learning algorithms can learn to predict the arousal of players based only on the video footage of their gameplay. Our results from four dissimilar games suggest that general-purpose representations can be built across games as the arousal models obtain average accuracies as high as 85% using the challenging leave-one-video-out cross-validation scheme. The dissimilar audiovisual characteristics of the tested games showcase the strengths and limitations of the proposed method.

Index Terms—General-purpose representation, subject-agnostic, arousal modelling, pixels, audio, games, CNN, classification, preference learning

---

1 INTRODUCTION

DESIGNING autonomous agents capable of performing equally well across different tasks is a long term vision of artificial intelligence (AI) [1]. Towards realizing such a vision, the recent groundbreaking study of Minh *et al.* [2] introduces the idea of *general-purpose* deep-learned representations for controlling agents capable of performing well across different tasks. These agents, in particular, managed to achieve superhuman performance in playing 2D Atari games by merely observing the pixels of the screen. As impressive as such a result might be, the derived agents are restricted to act in a particular set of deterministic environments and achieve a clearly- and objectively-defined goal: to maximize their score. Arguably, however, several of the most interesting problems that AI is requested to solve— such as emotion recognition and artificial psychology—have ill-posed and subjectively-defined target functions under non-deterministic contexts.

Inspired by the core principles of Mnih *et al.* [2], in this paper we transfer and introduce the idea of general-purpose representations to the field of affective computing. We thus reframe the user-specific way in which affective detection normally operates and, instead, we investigate the degree to which general-purpose representations can learn to predict emotion. As videos of interaction capture a user's behavior, we base our investigations on the assumption that the audiovisual information contained in such a video can
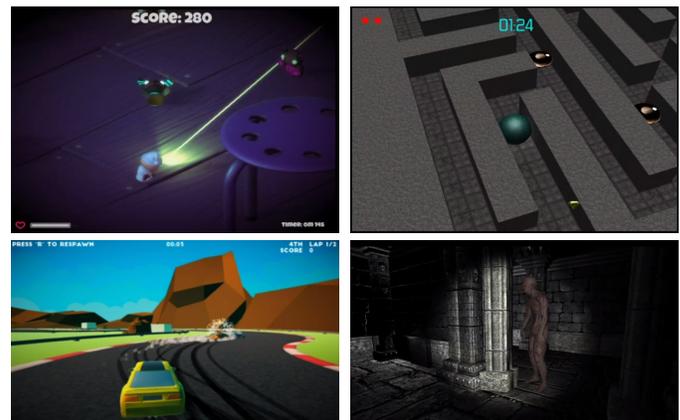

Fig. 1: Screenshots from Survival Shooter (top left), Maze-Ball (top right), Solid Rally (bottom left) and Sonancia (bottom right) games used in this study.

hold information about both the interaction context and the elicited affective patterns, and thus it can be a predictor of the user's experience. Our key hypothesis is that we can construct accurate models of affect based only on the audiovisual information of videos of interaction; as in [2], we test this hypothesis in the domain of games. In particular we attempt to predict a game's arousal level relying solely on the audiovisual information of game footage.

Games provide complex yet well-defined environments, which are designed to elicit increased levels of player engagement and motivation [3], [4]. Players, during their interaction with games, produce gameplay footage that has the unique property of overlaying the game context onto

---

- K. Makantasis, A. Liapis and G. N. Yannakakis are with the Institute of Digital Games, University of Malta, Msida 2080, Malta (e-mail: konstantinos.makantasis@um.edu.mt, antonios.liapis@um.edu.mt, georgios.yannakakis@um.edu.mt).

aspects of playing behavior and affect; this suggests that players' affect is embedded in the gameplay context. That embedding, in turn, renders the explicit fusion of context with affect manifestation unnecessary—a dominant practice within affective computing [5], [6], [7], [8]. Although we focus on the games domain, our approach is general and potentially applicable to different human-computer interaction domains, since it relies on raw audiovisual information. Such information fuses the interaction context with the affect of the user as manifested during the interaction.

Given the bimodal (audio and visuals) nature of the affect modeling task, we use a two-stream deep neural network—both as a classifier and as a preference learner—that considers audiovisual gameplay footage information and predicts the player's annotated arousal. The first stream is a Convolutional Neural Network (CNN) that processes visual information as pixels of video frame sequences. The second stream is a fully connected network that processes the game audio information of the considered sequence of video frames. Via late fusion, we propagate the audiovisual information to a fully connected network that predicts arousal. We test the methodology across four dissimilar 3D games and their corresponding gameplay footage (see Fig. 1). All gameplay videos have been annotated by the players themselves (first-person) in terms of arousal using the *RankTrace* [9] continuous annotation tool. Our experimental evaluation validates our hypothesis in most games and suggests that we can derive highly accurate models of affect using general-purpose representations that rely solely on audiovisual information of the interaction. In particular, the two representations (deep classifier and preference learner) predict arousal for three of the games examined with an average classification accuracy that reaches between 82% and 83% on average using the demanding leave-one-video-out cross-validation scheme. The under-performing models in one of the games lead to an insightful discussion regarding the limitations of the proposed approach and the environments it is best suited for.

Our work is novel in several ways. First, we derive accurate models of affect in different games without relying on any direct manifestation of emotion or modality of user input. Instead, our work is one of the first approaches towards modeling players' affect through general-purpose representations of information embedded solely in the context of interaction. Our methodology, thereby, yields affect models that are general and user-agnostic. Second, to the best of our knowledge, this study is the first attempt to derive a function that maps directly audiovisual gameplay information—such as pixels and audio features—to game experience across different games. Finally, via the employed two-stream deep network, we investigate the degree to which each modality, as well as their fusion, can be used as a predictor for such a mapping in affective computing. The paper builds upon and significantly extends the preliminary results of Makantasis *et al.* [10], which map the visual information of gameplay footage to players' arousal in one game. Specifically, in this paper we explore two different modalities of the game footage: besides the visual information we also exploit audio information in an attempt to yield more accurate representations of arousal. Moreover, we approach the arousal modeling problem using two different learning paradigms—a binary classification and a preference learning task—and we compare their performance quantitatively and their top-performing arousal models qualitatively. Finally, we test the generality of the proposed methodology across four heterogeneous games with regards to both the audiovisual information they offer to the deep learner and the arousal patterns they elicit.

The remainder of the paper is organized as follows. Section 2 reviews related work regarding affect modeling in games and videos. Section 3 describes the games, the employed datasets and the data pre-processing approach we followed. Section 4 lists the key elements of our methodology including the architectures of the learning models for both binary classification and preference learning. In Section 5 we experimentally validate and analyse our models across the four games. Finally, Section 6 discusses our main findings and Section 7 concludes this study.

2 RELATED WORK

This section surveys key literature on affect modelling relevant to the proposed approach of mapping audiovisual data from gameplay videos for predicting affect.

2.1 Models of Affect Based on Audiovisual Information

Audiovisual information has been at the core of interest for both eliciting and modeling emotions in affective computing [11], [12], [13]. Typically, videos feature the face or the body of one or more humans and their emotions are modeled via non-verbal (visual and vocal) cues [14], [15] due to theoretical frameworks and evidence supporting that such modalities can convey emotion [14], [15], [16], [17], [18]. Visual information is related to the dynamic patterns of human face(s) modeled via facial cues [19], [20], body postures [21], [22], gestures [23] or gait [24], [25]. Vocal information relies on audio signals which are used to construct acoustic and voice quality cues based on the pitch, the energy, the frequency and the spectrum of the signal [26], [27].

A number of earlier studies base the construction of affect models on ad-hoc features of an image. Indicatively, Liu *et al.* [28] combined traditional hand-crafted image features such as SIFT [29] and Histogram of Oriented Gradients [30] as inputs to machine learning models for emotion recognition in the wild. Yao *et al.* [31] hand-crafted image features based on Local Binary Patterns [32] for facial image emotion recognition. Recent advances in deep learning, however, enable the automatic construction of features via convolutional neural networks (CNNs); CNNs were first applied in [33] to predict dimensional affective scores from videos, but the small scale datasets challenged the training of deep models of affect. The need for effectively training CNNs triggered the development of medium- and large-scale affect datasets such as the Celebrity Face in the Wild [34], the Facial Expression Recognition 2013 Dataset [35] and the Aff-Wild database [36]. Based on these datasets, Ng *et al.* [37] used transfer learning and CNNs for emotion recognition through visual cues and Kollias *et al.* [38] combined CNNs with recurrent neural networks to model arousal and valence. Finally, in [5] facial expressions were fused with videos of advertisements for predicting whether viewers liked the videos or not.

Regarding emotion recognition via audio data, Eyben *et al.* [39] conducted a detailed study on audio emotion features. The authors constructed GeMAPS, a concise feature set with 62 audio features. Recent studies show that fusing audio and visual information results in more accurate models than those of a single modality [40]. In [41] energy and spectral audio features are fused with visual information for emotion prediction in short video clips, while in [42] audiovisual data is used to train a deep neural network for recognizing affect in real-world environments.

The approach presented in this paper can be seen as unconventional for modeling affect. Following our preliminary study [10] on general-purpose pixel-based models of affect, in the current study we use audiovisual information of human-computer interaction as the sole input for modeling the affect of the human across different tasks (i.e. games). The role of the audiovisual interaction footage is thus twofold: the audiovisual information contained in the footage is used to model affect as the context that both elicits and manifests emotion without the need of other external manifestations of affect. The proposed approach is a general method for modeling affect solely via videos containing sound and do not contain either facial/bodily expressions or vocal cues of humans. The experimental validation of the proposed approach—at least within the games domain—suggests that this *subject-agnostic* perspective is not only possible, but it also yields highly accurate models of affect in games with particular audiovisual characteristics.

2.2 Affect Modeling in Games

Affect modeling within the domain of games refers to modeling the behavior and the affective responses of players [4], [43]. A player model receives as input a modality (or a set of modalities) regarding the player, such as gameplay data and/or physiological measurements, and attempts to predict aspects of the in-game behavior or the player experience. Indicatively Pedersen *et al.* [44] combined gameplay data (e.g., number of deaths) with game level features to predict players' reported affect using *Super Mario Bros* (Nintendo 1985) as a testbed. Shaker *et al.* [45] used the same testbed to predict players' affect based on players' posture during gameplay. Recently, Melhart *et al.* [46] managed to successfully model the moment-to-moment engagement level of *PUBG* (PUBG Corporation, 2017) streamed videos by considering the chatting behavior of its viewers. Martinez *et al.* used various deep learning methodologies to capture player experience via gameplay metrics and physiology [47], [48]. Finally, Camilleri *et al.* [49] attempted to create arousal models that are general across different games relying solely on gameplay metrics.

This study advances the state of the art in player modeling as the proposed model of affect is based solely on the audiovisual information contained in gameplay footage. The majority of studies that analyze and extract information from gameplay footage rely on contextual information about the game such as structural and game level elements, physics and mechanics of the game (e.g. [45], [50]). Moreover, the most common approaches for analyzing player experience heavily rely on direct measurements from players under well-defined experimental settings; the modalities that are usually considered include facial expression and head pose [45], speech [51] and physiology [47], [52].

Building upon and significantly extending the preliminary study of Makantasis *et al.* [10], our methodology models players' experience without any *a priori* contextual knowledge about the game. Instead, it uses general-purpose deep learned representations of gameplay footage (i.e. pixels and audio files) as it ignores functional aspects of the game per se. As a result, our approach does not require any direct in-game feature or manifestation of affect (e.g. via physiology, speech, or facial expression), it is not intrusive, and it allows the rapid collection and processing of vast amounts of data. As gameplay videos are largely available online in massive quantities—e.g. via service such as Twitch[1] and Mixer[2]—the proposed approach is potentially generalizable to any game with available audiovisual content.

2.3 Video Affective Content Analysis

Video affective content analysis has been an active research area focusing on classifying and retrieving videos based on their affective content. While conventional content-based video analysis relies on generic semantic content, video affective content analysis tries to identify videos that elicit certain emotions in their viewers [53]. Recent research adopts either direct or implicit approaches. Direct approaches infer the affective content of videos directly from the related audiovisual features, while implicit approaches detect affective content from videos based on an automatic analysis of a user's spontaneous response while consuming the video [54]. Below we discuss direct approaches since they are closely related to the present study.

Hanjalic and Xu [55] proposed one of the first direct approaches for video affective content analysis, using handcrafted features of audio and visual information of video segments to model arousal and valence. Since then, extracting audiovisual features and exploiting machine learning methods to model emotion is the most common practice in video affective content analysis [56], [57], [58], [59]. More recent work takes advantage of deep learning to automatically generate deep features to describe audiovisual information, such as features of motion and scene cues [60]. Wang *et al.* [61] use a generative adversarial network to classify emotion of videos, while Mitenkova *et al.* [62] use the output of a pretrained network on face images [63] as input to a tensor regression layer for prediction arousal and valence levels. Zhu *et al.* [64] propose a multimodal deep quality embedding network and a deep learning affective classifier to efficiently process noisy affect data.

Although our study relates to video affective content analysis studies, it is conceptually different. Video affective content analysis tries to model and predict the emotion elicited by a video to a viewer. In contrast, this paper focuses not on the content creator's side, as we aim to model the emotional state of a player while they are playing the game.

3 DATASET AND DATA PROCESSING

To test the performance and the generality of the proposed approach we used frames and sound from four dissimilar

1. https://www.twitch.tv/
2. https://mixer.com/microsoft

games: *Survival Shooter*, *Maze Ball*, *Solid Rally* and *Sonancia*. Figure 1 shows a screenshot of each game. In this section, we describe the games, the datasets obtained from these games and the data cleaning process we followed. Participants were recruited via snowball sampling and were primarily university students who are casual gamers and/or follow courses in game design and ICT, with no prior experience in affect annotation. A different set of participants was used for *Solid Rally* and *Sonancia* whereas the same set of participants was used for *Survival Shooter* and *Maze Ball* [49]. Prior to annotation, all participants were presented with an introductory screen that describes arousal as "the intensity of gameplay no matter whether you like the game or not. High arousal can be a feeling of readiness, tension, excitement or exhilaration. Low arousal can be a feeling of fatigue, boredom, calmness or relaxation".

3.1 Testbed Games

To test how general-purpose input representations can be used for modelling affect, we selected the four games due to their dissimilarities. The games belong to different genres, with different mechanics, camera perspective, pace, visual and audio design. Specifically, Survival Shooter is a fast-paced shooter game that requires accurate aiming and constant movement. Maze Ball is a slow-paced physics game that needs accurate timing of movement. Sonancia is a horror game that elicits negative emotions and disorientation. Finally, Solid Rally is a fast-paced racing game that simulates a multi-player experience with AI drivers. Moreover, the camera perspective is top-down in Survival Shooter, third-person in Maze Ball and first-person in Sonancia and Solid Rally. The dissimilarities between the four games make them ideal for testing the degree to which accurate models of affect can be based on general-purpose input representations. We should also highlight that different sets of players played and annotated three of the four games.

3.1.1 Survival Shooter

Survival Shooter (SS) [49] tasks the player to shoot down as many hostile toys as possible while avoiding collisions with them. Hostile toys spawn at predetermined areas of the level and move towards the player's avatar. The avatar is equipped with a laser gun, which can kill a toy with a few shots. Every toy killed adds to the player's score. Background music plays throughout the gameplay of SS; while the player is firing the laser gun, the volume of music lowers, and the dominant sound is the weapon sound. Sound effects play when the avatar collides with hostile toys, when a toy is killed, and when the player runs out of life. The duration of the gameplay is 60 seconds.

The SS data used in this study was collected from 25 different players (10 females) aged from 19 to 54 (median age 24). Most players considered themselves good or expert players (70%) while the rest considered themselves novice or non-gamers. Each player played the game and then annotated her recorded gameplay footage in terms of arousal; this play-annotation cycle occurred twice. The first-person annotation process was carried out using the *RankTrace* tool [9], [65] which allows the continuous and unbounded annotation of arousal using the Griffin PowerMate wheel

interface. Gameplay footage was recorded at 30 frames per second (30Hz), while RankTrace provided 4 annotation samples per second (4Hz).

3.1.2 Maze Ball

Maze Ball (MB) [66] (or Space Maze [67]) is a 3D game that served as testbed in multiple studies investigating affect detection in games [4], [47], [48], [49], [66]. The player controls a cyan ball in a maze which contains dark ball-shaped enemies and three diamond-shaped tokens of different colors. The player has to avoid colliding with the enemies patrolling the maze, collect all the tokens and move the ball to a predefined goal point (only shown after all three tokens are collected) within 90 seconds. Each collected token adds to the player's score. The game ends either when the player runs out of time or collides with enemies twice. Background music plays during the entirety of gameplay, and sound effects play when the player obtains a token and when the player collides with an enemy.

A total of 25 players provided data for the MB dataset (the same set of players as in SS) [49]. Similarly to SS, each player conducted a play-annotation cycle twice, using RankTrace and the Griffin PowerMate wheel for annotation. MB game footage was also recorded at 30Hz.

3.1.3 Solid Rally

Solid Rally (SR) tasks the player to drive their car through a closed circuit for two laps. In each race, the player competes against three opponents, and the goal is to finish the race on the highest possible position. Within the track, there are several checkpoints at predetermined locations: passing through a checkpoint adds to the player's score. The car engine makes sounds throughout the gameplay of Solid Rally, and a sound effect plays during car crashes. The game ends either after two laps or after 90 seconds of playing.

SR data was collected from 17 players (7 females) aged from 23 to 55 (median age 32); almost half of them (47%) were novice or non-gamers, 35% considered themselves expert players and the rest played games only occasionally. Each player conducted a play-annotation cycle twice using the RankTrace annotation tool provided by the PAGAN framework [65]. Game footage was recorded at 60Hz but downsampled to 30Hz to match the sampling rate of the other three games.

3.1.4 Sonancia

Sonancia (SON) [68] is a horror game taking place in a haunted dungeon divided into rooms. The players' task is to find the old statue while avoiding and outrunning monsters. The level is procedurally generated, including the number of rooms, the positioning of lights and monsters. Background audio plays throughout the game, and changes based on the room the player is in. The only sound effect is a low-volume growl when a monster sees the player.

SON data was collected from 14 players (5 females) aged from 25 to 34; 36% of them played games everyday, 45% played frequently or casually while the rest rarely on never played. Each participant performed a play-annotation cycle twice, using RankTrace and the Griffin PowerMate wheel for annotation. Gameplay footage duration varies from 31 to 173 seconds, and is recorded at 30Hz.

3.2 Data cleaning

For all datasets we omit short gameplay videos with a duration under 15 seconds, in order to maintain an appropriate balance between sufficient gameplay and a player's cognitive load. This rule yields 43 videos for SS (7 short videos were omitted), 50 videos for MB, 34 videos for SR and 28 videos for SON (no video was omitted).

Since our approach is based on statistical machine learning, we explicitly assume that gameplay frames and sounds can characterize a player's arousal through an unknown mapping that machine learning aims to discover. To preserve the soundness of this assumption, we identified and omitted outlier videos whose annotations are not *consistent* with the gameplay. For all games, all players play the same level, which has a specific structure, e.g. for the SS game the toys keep spawning at predetermined areas and time instances. Coupled with the fact that the duration of each session is relatively short (60, 90, 90 and maximum 173 seconds for SS, MB, SR and SON respectively), the possible states of the gameplay are restricted. Based on this observation, we assume that arousal annotation traces should, on average, exhibit a specific pattern and we can thus omit outlier videos that deviate substantially from this pattern.

In particular, we denote a playthrough as an outlier if its annotation trace is dissimilar to an annotation trace that can be considered representative for the whole dataset. Since RankTrace provides continuous and unbounded arousal annotations, initially we normalise all annotation traces to $[0, 1]$. Then we consider the median of all annotation traces as the representative annotation trace for the whole dataset and compute the distance between the annotation trace of every gameplay footage and the representative trace using the Dynamic Time Warping (DTW) [69] algorithm. DTW is widely used for measuring the similarity between two timeseries that may vary in length. The distribution of distances for each game indicates that the density is mainly concentrated on one cluster. Based on that, we omit outliers above a DTW distance threshold. For the SS game we omit videos where the distance to the median (representative) annotation trace is larger than 0.135; for MB, SR and SON the corresponding thresholds are 0.195, 0.4 and 0.2. After removing outliers, the SS, MB, SR and SON datasets contain, respectively, 37, 45, 33 and 25 videos. Figure 2 depicts the average arousal trace of the cleaned dataset for each game.

4 LEARNING AUDIOVISUAL MODELS OF AFFECT

This study investigates the degree to which information coming from footage of the player's interaction with a game—i.e. pixels of frames and sound of a gameplay video—can act as sole predictors of a player's affective state. The RankTrace annotation tool provides continuous values of arousal, and thus it may seem natural to view the arousal estimation problem as a regression task. In this study, however, we wish to develop a user-agnostic and general approach for predicting affect without making any assumptions regarding the value of the output which may, in turn, result in biased and user-specific models [70]. Instead, we view the challenge of arousal prediction as both a classification and a ranking task.
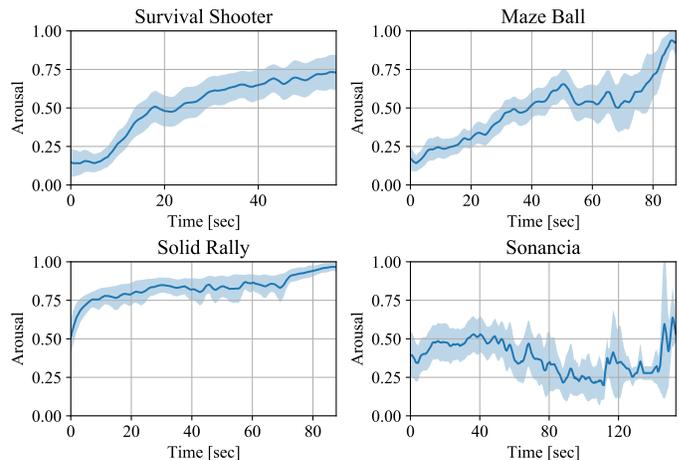

Fig. 2: Arousal annotation per game, averaged from signals processed at 4 Hz after min-max normalization per session. Shaded areas indicate 95% confidence interval. As the duration of Sonancia sessions varies, the average arousal is derived from ever-fewer sessions as time progresses resulting in higher deviations of the average arousal signal.
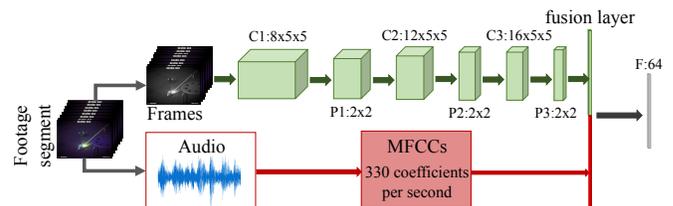

Fig. 3: The architecture of the proposed deep learning model. The convolutional, max pooling and fully connected layers are denoted by "C", "P" and "F" respectively. The green stream corresponds to the processing of visual information, while the red stream to the audio information.

This section first outlines our approach for processing the input and the output of both binary classification and preference learning models of affect (see Section 4.1) and follows with the details of the machine learning models we employ for mapping gameplay frames and sound to players' arousal (see Section 4.2).

4.1 Training Data Preparation

An obvious question that arises when a learning model is faced with *video* data is how many frames it should consider. The authors in [71] and [72] argue that a relatively small number of subsequent frames are adequate for representing the core elements and the content of a scene. Following this advice, we train our arousal models and evaluate their performance using small segments of footage with durations ranging from 0.25 to 3.0 seconds. We view the duration of a segment as a hyperparameter of both modeling approaches and we report results regarding the performance of the learning models for varying segment length. We construct those segments by splitting the videos using non-overlapping windows. The frames of those segments represent the visual information of the gameplay. To reduce the computational cost of training and evaluating the

learning models, we convert the frames of gameplay videos from RGB colour to grayscale and resize them to 72×96 pixels for SS and MB datasets and 72×115 pixels for SR and SON datasets; doing so results in a more compact yet general-purpose representation.

As far as *audio* data is concerned, we compute the Mel Frequency Cepstral Coefficients (MFCCs) [73] corresponding to the sound of each footage segment. MFCCs have been successfully used for audio classification and retrieval schemes [74], [75] as they can represent the spectral properties of audio data in a compact fashion.

To construct models of *arousal*, independently of the method used, we fix the range of the annotation values of each footage to $[0, 1]$ using min-max normalization. Then, we synchronize the recording frequency of videos (30Hz) with annotations (4Hz) by extrapolating annotation values to each non-annotated frame. Finally, the arousal value associated with each segment is the average of the annotation values of the frames belonging to it.

4.2 Deep Learning Models of Affect

To learn the unknown mapping between gameplay pixels, sounds and arousal we employ and train deep learning models to infer such a function. The deep learning models receive as input both the frames of the footage segments and the computed MFCCs and fuse those two streams of information. The learning architecture that processes and fuses the audiovisual information—for both binary classification and ranking—is depicted in Fig. 3.

The video stream feeds a convolutional neural network that contains three convolutional layers with 8, 12, and 16 filters, respectively. The size of the filters for all convolutional layers is 5×5 pixels. A max-pooling operation of size 2×2 pixels follows each convolutional layer. The output of the convolutional stream (a feature vector of 640 elements for the SS and MB datasets and 1056 elements for the SR and SON datasets) represents the visual information of the input in a compact fashion. We should emphasize that the convolutional stream exploits both spatial and temporal information of the video frames. It exploits the spatial information by learning spatial filters (filters applied along the spatial dimension of the input). Moreover, since the learning model processes sequences of frames that exhibit temporal relations, the learned spatial filters implicitly capture and encode the temporal information of the inputs.

The audio stream receives the MFCCs as its input and propagates it directly to the fusion layer. The network does not process the MFCCs before fusing the visual and the audio streams since MFCCs are already a compact representation of the sound included in a video segment. The fusion layer, initially, concatenates the MFCCs (330 elements for each second of footage) and the features constructed by the convolutional (video) stream; it then propagates the information to a fully connected layer with 64 nodes. All aforementioned nodes use the ReLU activation function.

All the hyperparameters of the employed model, i.e., the number and the size of hidden layers, the activation functions and the approach for fusing the two information streams, are empirically selected to balance two different criteria: (a) the computational cost of training and evaluating the model and (b) the sample complexity for avoiding
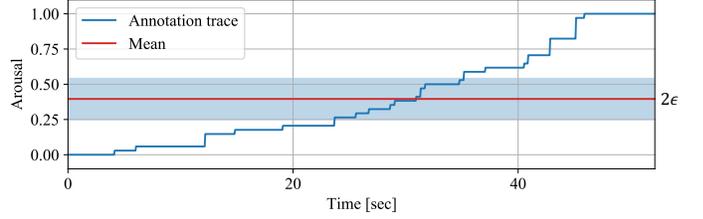

Fig. 4: Class splitting procedure. Samples with annotations above the shaded area belong to the high arousal class, while samples with annotations below the shaded area belong to the low arousal class. The value of the uncertainty threshold ϵ defines the width of the shaded region.

under-/over-fitting. The model described above has approximately $6.5 \cdot 10^4$ trainable parameters.

4.2.1 Deep Classifier

The task of arousal classification is formulated as follows; we denote $x \in \mathbb{R}^{h \times w \times c}$ and as $z \in \mathbb{R}^p$ the raw visual and audio information of a gameplay footage respectively, where h, w and c stand for height, width and length of the video segment, and p for the length of the video's audio stream. Let $\xi(x)$ and $\zeta(z)$ represent the transformations of raw information to informative features. In our case, $\xi(\cdot) \in \Xi$, where Ξ denotes all possible functions that can be modeled by the CNN described in Section 4.2, and $\zeta(\cdot)$ is the function that transforms audio information to MFCCs. Having in our disposal a training set $\mathcal{D} = \{(x_i, z_i, y_i)\}_{i=1}^N$ of N samples, where $y_i \in \{0, 1\}$ for $i = 1, \cdots, N$, and a class of functions $\mathcal{F}$ that map $\xi(x)$ and $\zeta(z)$ to $(0, 1)^2$, our derived model of affect corresponds to

$$f^*, \xi^* = \arg \min_{f \in \mathcal{F}, \xi \in \Xi} \frac{1}{N} L(f(\xi(x_i), \zeta(z_i)), y_i), \quad (1)$$

for $i = 1, \cdots, N$. $L(\cdot)$ is the negative log-likelihood loss. In our case $\mathcal{F}$ is the class of functions computed by feed-forward fully connected networks with one hidden layer of 64 neurons and 2 output neurons activated by the softmax function (see Fig. 3). The fact that we minimize the loss in (1) with respect to both f and ξ indicates that our model is end-to-end trainable, i.e. the weights of the CNN (feature construction of visual input) and the classifier are optimized simultaneously.

For training the binary classifier we transform the continuous annotation values of the segments into binary values (low and high arousal) by using the mean of the annotation trace of each video as the class splitting criterion (Fig. 4). We opt for the mean value of the annotation trace as it is the most intuitive and unbiased way to split a continuous, unbounded annotation trace. Finally, we employ a threshold parameter (ϵ) to determine a region around the mean within which annotation values are labeled as uncertain and ignored during classification (see the shaded area in Fig. 4).

4.2.2 Deep Preference Learner

The preference learner indicates, via its output, which one of two input segments is associated with a higher arousal value. By denoting a function $g_i(f(\xi(x_i^A), \zeta(z_i^A)) - f(\xi(x_i^B), \zeta(z_i^B)) \to (0, 1)^2$ for the i-th (A, B) pair of inputs

and a dataset $\mathcal{D} = \{x_i^A, z_i^A, x_i^B, z_i^B, y_i\}_{i=1}^N$, of N pairs, the derived preference learner corresponds to

$$f^*, \xi^* = \arg \min_{f \in F, \xi \in \Xi} \frac{1}{N} L(g_i, y_i). \qquad (2)$$

In our case, $g(\cdot)$ is the softmax function. Based on this formulation, the preference learner employed here—similarly to RankNet [47], [76]—can be seen as a binary classifier which takes as input a pair of samples and outputs 1 if the first sample in the pair is ranked higher, and 0 otherwise. Again, the output nodes employ the softmax activation.

Similarly to the ϵ parameter of binary classification, in preference learning we employ a threshold δ which determines if the absolute difference between the mean arousal values of two segments is high enough for the segments to be considered as a preference pair (i.e. a datapoint for training). Based on δ we create input pairs by comparing them in both ways; i.e. we use both (a, b) and (b, a) pairs, where a and b represent the audiovisual information of two different segments. This approach gives us a perfectly balanced dataset for deep preference learning.

5 EXPERIMENTAL RESULTS

This study aims to test the hypothesis that there is a general-purpose learnable mapping of gameplay footage representation to players' affect. Towards this direction, we use the two-stream neural network (see Section 4.2) for classifying game video and audio segments as high or low arousal, and for ranking them. For all the experiments in this paper we report the average cross-validation accuracy and the 95% confidence following the demanding leave-one-video-out cross-validation scheme [10], [77] which offers a highly conservative estimate for the generalization capacity of the models. To avoid model overfitting we employ early stopping by randomly selecting 4 videos of the training set to form the validation set. Early stopping is activated if the classification accuracy on the validation set does not improve for 30 training epochs. We compare the performance of the model against a baseline model which always outputs the most common class in the training set. The baseline accuracy for preference learning is always 50%, since we have a perfectly balanced dataset (see Section 4.2.2).

5.1 Classifying Arousal

To investigate the impact of the two input modalities (video frames and audio) on the performance of the model, we report the classification accuracy of three model types: two single-stream (unimodal) neural networks which are trained on either the visual or the audio information, and the two-stream bimodal neural network (see Section 4.2) which is trained on both visual and audio information. Figure 5 reports the average classification accuracy values obtained for different input modalities for all games, across different thresholds ϵ for omitting uncertain values near the annotation's mean value. For $\epsilon = 0$, all segments of a trace are labelled high or low if their mean arousal value is above or below the mean value of the entire annotation trace (μ), respectively; for $\epsilon > 0$, segments with mean arousal values within $[\mu - \epsilon, \mu + \epsilon]$ are omitted from the data (see Fig. 4). Note that preliminary experiments established that splitting
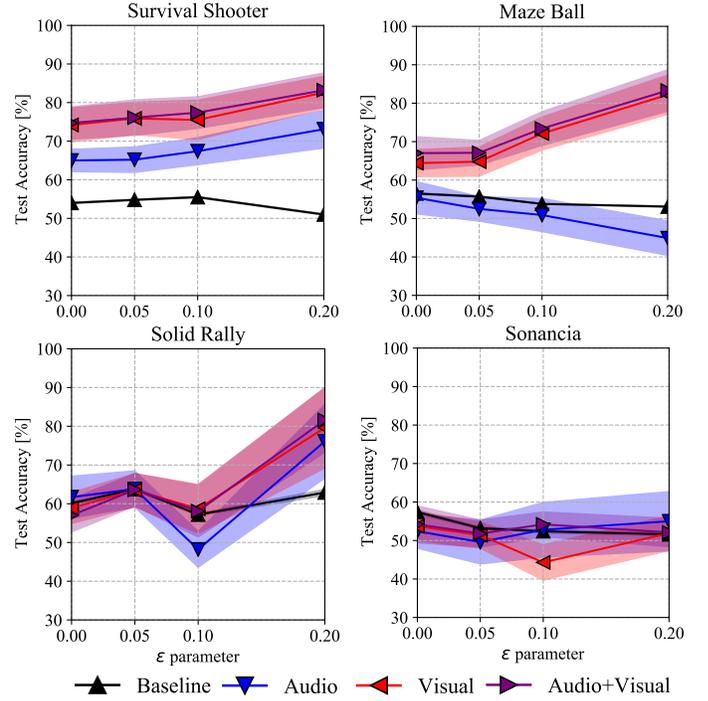

Fig. 5: Average classification accuracy (%) on the test set across the two modalities and different uncertainty threshold values (ϵ). The time window is 0.5 seconds and shaded areas indicate the 95% confidence interval.

the traces into segments of 0.5 seconds (see Section 5.3) led to the highest accuracies across the different modalities and parameters, and as such the exploration of the best ϵ in Fig. 5 focuses on a time window of 0.5 seconds, while Table 1 presents the size of the employed datasets.

Unsurprisingly, the deep learning models perform better when both audio and visual inputs are considered. For SS and MB the bimodal classifier reaches accuracies as high as 30% above the baseline classifier, but only 1% to 3% above the visual-only classifier. Similarly, for SR the bimodal classifier reaches accuracies 20% above the baseline classifier and 2% above the unimodal visual-only model. For these three datasets, most of the information regarding arousal is stored within the pixels of the video, while audio seems to play a minor role. For SS, however, audio can also be a good predictor of arousal, since even audio-only classifiers reach accuracies between 10% and 20% above the baseline. The same holds for SR when $\epsilon = 0.2$, in which case the audio-only model reaches accuracies nearly 15% above the baseline. On the other hand, audio-only models cannot predict arousal for the MB dataset and the SR dataset (when $\epsilon < 0.2$), as accuracies are on par or worse than the baseline. For MB, audio information per se is not a reliable predictor of arousal, most likely because sound effects are rather sparse (see Section 5.5). However, it can contribute to the model's predictive capacity when combined with visual information. For the SR dataset, audio information does not seem to affect the performance of the model when it is combined with visual information. For the SON dataset, both bimodal and unimodal models perform on par or worse than the baseline. In this case, neither visual nor audio

TABLE 1: Sizes of the employed datasets for different uncertainty thresholds and time windows. For splitting the dataset into training and testing sets we follow a leave-one-video-out scheme. For the validation set (early stopping) we randomly select and use four videos in the training set.

Classification

| | Time Window (t): 0.5 second | | | | ϵ=0.2 | | | | |
	ϵ=0.00	ϵ=0.05	ϵ=0.1	ϵ=0.2	t=0.25	t=0.5	t=1.0	t=2.0	t=3.0
SS	3,381	3,102	2,621	1,972	3,698	1,972	1,002	4,83	345
MB	5,989	5,574	4,379	2852	5,393	2,852	1,419	700	448
SR	4,925	3,608	2,225	711	1,446	711	358	180	119
SON	4,719	3,707	2,977	1,846	3,474	1,846	905	426	282

Preference Learning

| | Time Window (t): 2 seconds | | | | | δ=0.6(SS/MB), 0.4(SR), 0.75(SON) | | | |
	δ=0.0	δ=0.2	δ=0.4	δ=0.6	δ=0.75	t=0.5	t=1.0	t=2.0	t=3.0
SS	20,916	13,804	7,532	3,860	2,138	67,576	16,380	3,860	1,584
MB	43,090	25,740	13,854	6,072	2,766	104,258	25,330	6,072	2,488
SR	39,813	14,146	4,898	2,324	1,358	41,588	9,940	4,898	888
SON	43,844	25,108	10,364	3,502	1,282	26,527	6,509	1,282	532

information is a reliable predictor of arousal. We believe that this occurs due to the specific nature and design of the game; SON is a horror game with a delayed effect in arousal which may not be captured by the class splitting criterion (one of the limitations of our study discussed in Section 6). As far as the design of the game is concerned, visual information in SON comes in highly vignetted frames with no HUD elements, which makes the vision-based pattern recognition task difficult and ambiguous. The background audio of the game also changes suddenly when the player moves from one room to another. Since these changes do not follow a specific pattern, audio information encoded in MFCCs cannot be easily associated with arousal.

In summary, for 3 out of 4 datasets, the high performance obtained by varying the uncertainty bound indicates that the mapping between general-purpose representations of audiovisual gameplay information and arousal can be learned statistically with very high accuracy. Results for the SON dataset indicate that the performance of our models depends on the specific nature of the game, as well as on the underlying assumptions of our approach (see Section 6).

5.2 Ranking Arousal

Similar to Section 5.1, we investigate how different modalities affect the performance of the preference learner by training three preference learning models with different inputs. The preference learner compares two input segments and outputs which segment has a higher arousal value. Based on preliminary experiments we focus on the best time window for the preference learner, which is 2 seconds.

The δ parameter (as defined in Section 4) sets the minimum absolute difference between the mean annotation value of two segments that can be considered as a preference. In this section, we investigate the performance of the preference learner in terms of average classification accuracy for 5 different values of δ, i.e., $\delta = \{0.0, 0.2, 0.4, 0.6, 0.75\}$, and for the different input modalities. For all the experiments presented, we follow the leave-one-video-out validation procedure using segments of 2 seconds. Figure 6 summarises the results of this investigation, and the sizes of the datasets are presented in Table 1.

The preference learner achieves up to 32%, 28%, 22% and 11% higher accuracy than the random baseline for the SS,
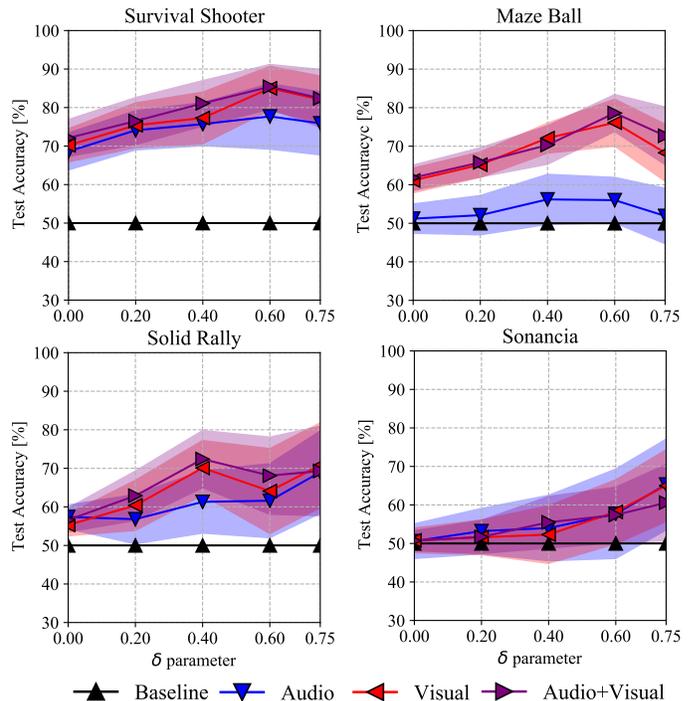

Fig. 6: Average accuracy (%) of the preference learner on test set across the two modalities and different uncertainty threshold values (δ). The time window is 2.0 seconds and shaded areas indicate the 95% confidence intervals.

MB, SR and SON dataset, respectively. As with the classifier, the models that exploit both audio and visual input perform better than unimodal models. While high values for δ yield pairs of inputs that have significantly different annotation values, this also results in smaller datasets. According to Fig. 6, for SS and MB we obtain the highest performance values when $\delta = 0.6$, for SR when $\delta = 0.4$ and for SON when $\delta = 0.75$. These threshold values seem to balance between highly informative and comparable inputs, and adequately large dataset size for training (see Table 1).

5.3 Impact of the Time Window

In all experiments presented so far we investigated the performance of the arousal model by keeping the time window of the input signal constant (0.5 seconds for classification and 2 seconds for preference learning). In this section, we vary the time window while retaining the best δ and ϵ values found in Sections 5.1 and 5.2 respectively. We assume that the duration of the gameplay videos affects the model performance for three reasons: first, the length of the window determines directly the size of the dataset; second, the duration of footage segments determines the amount and the quality of the audiovisual information contained in a segment (i.e. the longer the segment, the richer the information contained in it); third, the duration of the window affects the ground truth arousal values as those are averaged from the window's annotation trace.

Figure 7 (left) summarizes the impact of window duration on the accuracy of our proposed two-stream (audio and visual) neural network for the classification task. For all results, the uncertainty threshold value is fixed to its best
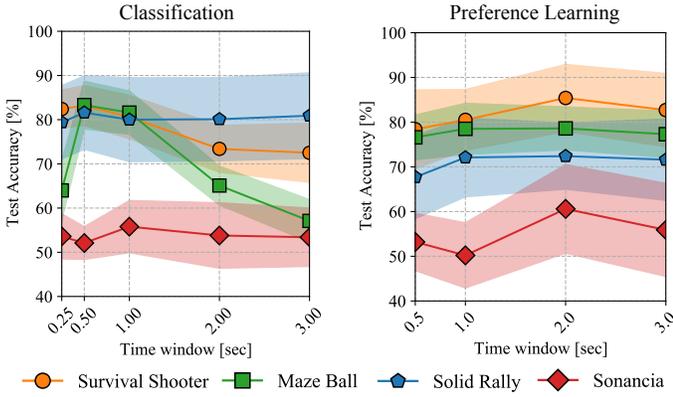
Fig. 7: Average accuracy (%) on test set for the best audiovisual model across different time windows, for classification and preference learning. Shaded areas indicate the 95% confidence intervals.

Fig. 8: Kendall's τ across modelling approaches, games, and best time windows for each approach. Values are averaged from the leave-one-video out cross-validation and error bars denote the 95% confidence interval.

value: $\epsilon = 0.2$. For the SS game, accuracy is consistently over 70% for all durations. However, the model achieves the best performance for 0.5 second windows, and accuracies drop for longer windows. It appears that the fast pace of the game does not favor inputs of long duration since their ground truth annotation values are over-smoothed. For the MB game, the accuracies deviate wildly in different time windows. In particular, the performance of the model is over 80% for segments of 0.5 or 1 seconds, and less than 65% for shorter segments ($\sim$0.25 second). Contrary to SS, MB is a slow paced game. Therefore, in this game it seems that short segments do not contain sufficiently rich information for the classification task, and thus do not contribute towards the efficient training of the model. Both games (especially MB) perform worse in segments over 1 second, also due to the fact that the size of the dataset becomes too small for training (for 3 seconds and $\epsilon = 0.2$ the datasets for SM and MB are only 345 and 448 segments). For the SR dataset the classification accuracy is $\sim$80% for all time windows. However, it shows wide confidence intervals due to the small number of training samples (as shown in Table 1, for 2 or 3 second windows less than 200 samples are retained).

Figure 7 (right) similarly visualizes the impact of time window length on the accuracy of the preference learner using audiovisual input, and with the best δ value ($\delta = 0.6$ for SS and MB, $\delta = 0.4$ for SR and $\delta = 0.75$ for SON). As indicated in Section 5.2, the best performance for both datasets is obtained for segments of 2 seconds. For the SS dataset the learner that uses 2 second segments as input performs 5% and 2.7% better that the models that use 1 and 3 second segments, respectively. For the MB and the SR games, all models perform almost the same irrespective of the time window considered, although for short time windows (0.5 seconds) the performance drops in both games. For SON the preference learner performs best for 2 seconds time windows (11% above the baseline), and 3 seconds (6% above the baseline). This suggests that Sonancia, as a horror game, elicits affect in a delayed fashion and thus requires longer segments of gameplay to be considered. Due to the way that preference learning processes the dataset, the number of preferences increases exponentially compared to
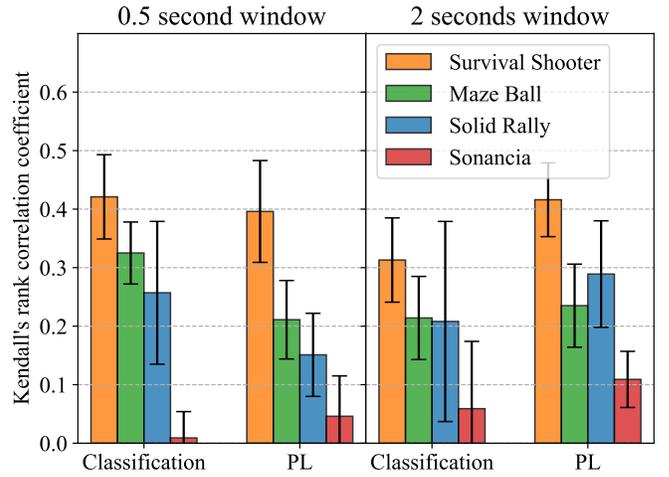
the number of segments themselves: based on Table 1, the number of preferences at 0.5 seconds are in the power of 10^4 (up to 50 times the number of preferences at 2 seconds). For segments of 0.25 seconds, the dataset explodes and training becomes problematic due to computational effort.

5.4 Classification vs. Ranking

In this section we compare the preference learner against the binary classifier. While both methods yield high accuracies for three of the four games, such a metric is not appropriate for conducting a fair comparison between the methods as the set of input-output pairs for the two approaches is not the same [48]. Following the method introduced by Martinez *et al.* [48], we compare the two approaches based on the global orders of arousal they produce when they are fed with inputs that belong to the same gameplay footage. The orders produced by the models are evaluated against the ground truth global order which is derived by the arousal annotation values. Inspired by [48], [70] we compare the methods using the Kendall's rank correlation coefficient (τ), which measures the ordinal association between two rankings [78]. We calculate τ on the test video in a leave-one-video-out cross-validation process, and report the 95% confidence intervals across all videos in each dataset.

For both approaches we use the trained models presented in Sections 5.1 and 5.2 which achieve the best classification accuracy: for classification, the best models are with $\epsilon = 0.2$ and a time window of 0.5 seconds for all games, and for preference learning the best models are with $\delta = 0.6$ for SS and MB, $\delta = 0.4$ for SR and $\delta = 0.75$ for SON and a time window of 2 seconds. Fig. 8 also shows the τ values for models trained on both time windows for both approaches, for a more holistic comparison. The average Kendall's τ for both datasets indicates—unsurprisingly— that the produced orderings are positively correlated to the ground truth independently of the method used. For the SS and SR games both approaches seem to perform almost the same for their best models (with no significant

differences). For MB the classifier yields higher τ values than the preference learner and for SON the best preference learner yields higher τ values than the best classifier (which is also at 2 seconds), but these differences are not statistically significant. As evident from Section 5.3, the classifier performs worse at 2 seconds windows (except for SON) while preference learning performs worse at 0.5 second windows compared to each method's optimal time window. It should be noted that the way classification and preference learning process the data results in a very different treatment (classes versus ranks) which makes a completely fair comparison very difficult. Indicatively, classification with 2 second windows and $\epsilon = 0.2$ operates on a dataset of size $483, 700, 180$ and 426 for SS, MB, SR and SON respectively, versus $3, 860$, $6, 072$, $4, 898$ and $1, 282$ for preference learning (with the best δ values per game). Therefore, using the best models for each approach even if the input is different (specifically, the number of frames used as input to the CNN, and the number of MFCCs for audio) is the most straightforward comparison as the number of samples (with the chosen ϵ, δ parameters) are in the same order of magnitude.

Based on the comparison above, we conclude that a binary classifier can reach comparable accuracies to the preference learner, or higher in the case of MB. The accuracy of the binary classifier comes at the cost of the resolution of the output (as there are only two classes). If the problem requires larger output resolution (e.g., high, medium and low arousal), it is not clear how a 3-class classifier could produce such orderings. On the other hand, preference learning models can always produce orderings via pairwise comparisons of inputs and they appear to be more robust across time windows and across all games tested.

5.5 Analysis of Findings

The experiments presented in this paper showed that it is possible to construct accurate models of players' arousal based on general-purpose representations of gameplay footage. The results obtained across different input modalities also indicate that the visual information is key for the efficiency of the models. Moving towards higher degrees of model expressivity and explainability, in this section we attempt to identify the features of the gameplay video that contribute more to the output of the arousal models. One way to achieve this is by visualizing the areas of the frame that have the highest influence on the model's prediction. For that purpose we use the Gradient-weighted Class Activation Mapping (GCAM) method [79]. For any given input, GCAM computes the gradient of the output neuron with respect to the neurons of a convolutional layer. By multiplying the given input with the computed gradient, we obtain a heatmap that indicates how much each area of the input contributes to the output.

Figure 9 depicts the activation maps for a sample footage segment for different games and learning paradigms; for visualisation purposes, we display the last frame of the segment. We observe that aspects of the heads-up display (HUD) affect the arousal prediction. For SS, the score located at the top centre of the screen—which keeps increasing during the progression of the game as the player kills more and more hostile toys—contributes significantly to arousal

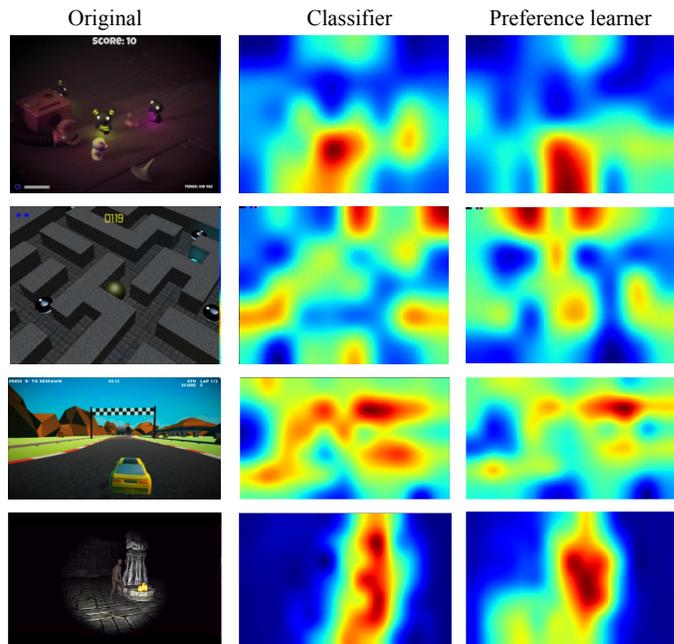Original   Classifier   Preference learner

Fig. 9: Activation maps of a sample footage segment for all games. We display the last frame of the segment.

prediction. The pixels of the avatar and the hostile toys, however, seem to have the highest impact on the outcome independently of the method used. Similarly, the time indicator on the HUD of the MB game contributes highly to the arousal prediction regardless of the learning paradigm used. Besides the time indicator, the location of the ball, the enemies and the tokens appear to have a substantial impact on arousal prediction. Interestingly, the HUD element of the player's health was not considered for either game. For SR, the HUD elements and the player's avatar (car) do not seem to be important features for either approach; the focus is instead on level elements immediately in front of the car such as the finish line or the loop in the horizon. For SON, it is obvious that the lack of HUD and dark visuals (only a small part of the screen contains information) confuse the classifier, although both approaches identify the statue (the goal of the game) and the monster as important features.

As a general comment from our qualitative analysis, there are two key differences in how sound influences the arousal models in different games. On the one hand, sound effects in SS follow shooting and enemy deaths which are common events and information-rich (e.g. killing an enemy means that the player survives longer), while for MB sound effects are rare since they trigger when a token is picked (with three tokens in the game) or when the player dies (which will not be a common event). Sound effects are more common and can thus be exploited better in SS, which explains the low performance of audio channels in arousal detection for MB. SR on the other hand has a persistent sound from the engine, making some frequencies in the MFCC near-constant. Finally, in Sonancia there are almost no sound effects (only if a monster sees the player) and the background audio changes based on the room in a non-diegetic way [80]. The sound design in SON is thus expected to confuse the models of affect as the soundscape

variation is sparse and only captures the gameplay context very indirectly.

6   DISCUSSION

The most common approaches for modelling affect rely on direct observations and measurements of human behaviour. The proposed study, to the best of our knowledge, presents one of the first attempts to model affect via general-purpose representations of information that comes solely from gameplay footage that does not display human behaviour directly. Human behaviour is embedded into the gameplay footage, e.g. as avatar movement and actions, since emotion is manifested through and annotated on the video per se.

We exploit bimodal audiovisual information to build and test a deep neural network for predicting players' arousal states via two different approaches; binary classification and preference learning. Results show, on the one hand, that building such models is possible, and on the other, that these models can be highly accurate in most cases. Moreover, more robust and accurate models can be constructed when the dataset is pruned from ambiguous data.

In this study, we make several assumptions which implicitly indicate possible limitations of our approach. First, we use the mean value of each annotation trace to split data into high and low arousal classes. Although this criterion is intuitive and straightforward, it makes sense only for stationary processes. In our study, this criterion results in robust annotations due to the short duration of gameplay. For long playthroughs, however, this assumption will not hold and using the mean as the class splitting criterion may produce misleading classification and preference learning results. Second, we use a representative annotation trace (median trace) to detect and remove outliers. In other words, our data cleaning methodology considers only the distribution of the annotation values. In our study, such a methodology can efficiently detect and remove outliers since the games considered can be played in specific ways, and gameplay duration is short. For sandbox games or long play sessions, a data cleaning methodology that takes into consideration simultaneously the input and the output distributions should be used (i.e., the joint distribution of audiovisual information and annotation values). Finally, the data points used for training the affect models are generated sequentially. Thus the annotation of a data point at a specific time instance might depend on the annotation value of the data point generated before. Our models, however, are not able to exploit this information. To take advantage of this kind of information, models that explicitly take into consideration the temporal ordering of data, such as LSTMs, should be used.

The differences in performance among the four games also illuminate some concerns regarding the impact of the game environment on the feasibility of general-purpose models. As discussed in Section 3 and Section 5.5, each game is different in terms of what the player sees (camera perspective, color scheme, illumination), hears (background audio, variety and volume of sound effects) and performs (control schemes, actions per minute, degree of immediate feedback, available actions, clarity of game goal). Based on these differences, it is expected that the player also feels (and annotates) differently in each game (see Figure 2). While in games such as Survival Shooter highly accurate models of affect could be trained via either approach, in Sonancia specifically the performance of the classifier was not better than the baseline and the preference learner could reach accuracies of $\sim 60\%$ with the best parameter setup. It can be gleaned that arcade games with fast-paced interactions (such as SS and SR), a top-down camera perspective that shows more of the level (SS and MB), distinct forms and colors to distinguish game objects (MB and SR), loud sound effects tied to game events (SS) could help the model predict affect from the audiovisual signals alone. In contrast, Sonancia has none of the above design patterns; moreover, the actions that a player takes (e.g. choosing a room to go into) do not have immediate gameplay (and, one would assume, affective) impact as a monster could be hiding in a remote part of the room she chooses to go. Future work should explore where the limits are in terms of game environments and visual, audio or interaction design for which this method can be applied. While it is expected that high-contrast and fast-paced arcade games such as the ATARI games studied by Mnih *et al.* [2] will work very well for this method, it is unclear whether audiovisual signals in time windows of a few seconds would work well for e.g. role-playing games (which require long interactions), visual novels (where the story consequences are not displayed visually or immediately) or turn-based games (where real-time windows are irrelevant). Exploring how these different design patterns affect the quality of predictive affect models based on audiovisual data alone can be useful not only for affective computing but also for game design, as it can inform designers how to maximize the emotional impact of their content.

While this study is one of the first attempts at the challenging task of predicting affect states from general-purpose gameplay footage information, the results are promising and point to a number of extensions in future work. In this paper, our models require training on each particular game; while the method is robust and general-propose, the models themselves have not been tested for their generality. To test for the model's generality, a future direction would be to devise leave-one-game-out validation schemes once our game corpus becomes even larger. Such a cross-validation scheme would allow us to test the degree to which certain characteristics of audiovisual information are general predictors of arousal and transferable to other games. In terms of the model's input, we use grayscale frames to represent the visual information and MFCCs for the audio information. Grayscale frames and MFCCs can compactly represent the audiovisual information of gameplay footage and reduce the computational cost of training the models. These representations, however, can be enhanced without sacrificing the generality of our approach. In terms of sounds, MFCCs can be fused with the concise GeMAPS feature set [39] which has been successfully used for voice recognition and affective computing applications. As far as the representation of visual information is concerned, it can be enhanced by using RGB channels or hand-crafted channels that include low-level image information [81]. For example, exploiting hue and saturation information could better detect the red monsters present in Sonancia. In terms

of output (affect labels), we use the mean arousal value within a time window. That is an intuitive approach, which, however, can be further investigated and refined. For example, amplitude and average gradient [9], [49] could be used for processing annotations within a time window.

Beyond arousal, the method's robustness needs to be tested for other affective dimensions—including valence and dominance—or continuous affective states such as engagement [46]. Beyond games, the method appears to be generalizable to any rich human computer interaction domain that interweaves the context of interaction with user behavior and user affect, such as mobile app interaction and web navigation. Additional experiments in datasets of that type, however, need to be performed to validate this hypothesis.

7 CONCLUSIONS

In this paper we introduced a general methodology for predicting affect solely from audiovisual aspects of human computer interaction. Our hypothesis is that arousal embedded in affective interaction can be modeled accurately without considering any user manifestation of affect besides the pixels and the sound of the interaction. The hypothesis was tested in digital games, a domain that interweaves affect with audiovisual content through gameplay interaction. The audiovisual content in games has a dual role: it is both the elicitor of affect and the context of the interaction. We developed two deep learning paradigms for mapping directly from pixels and audio of videos to the annotated arousal of gameplay: a deep classifier and a deep preference learner, both using a combination of CNN and feedforward architectures. Our experimental results across four dissimilar games suggest that arousal can be predicted with very high accuracies via such general-purpose representations (as high as 85%) as long as the audiovisual feed captures the gameplay context accurately (which depends on the game's design). The fusion of the two modalities (gameplay pixels and sounds) unsurprisingly appears to be beneficial for the predictive capacity of the models. Our key findings also show that activation maps can visualize the areas on the screen that lead to high arousal—in our case primarily the score, the avatar and the enemies. The GCAM visualization increases the explainability of the models [82] and can be very useful for game designers when adjusting the appearance or in-game function of the game elements to increase or decrease the elicited emotion of certain events.

This paper defines one of the first steps towards the creation of general representations of affect by studying arousal detection in games. The results showcase that it is possible to detect arousal accurately by only considering low-level contextual information of the interaction. The key findings are relevant to any application area within affective computing and directly applicable to domains of rich human computer interaction that consider user affect.

ACKNOWLEDGMENTS

REFERENCES

[1] B. Goertzel and C. Pennachin, *Artificial general intelligence*. Springer, 2007, vol. 2.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[3] D. Melhart, A. Azadvar, A. Canossa, A. Liapis, and G. N. Yannakakis, "Your gameplay says it all: Modelling motivation in tom clancy's the division," in *Proceedings of the IEEE Intl. Conference on Games*, 2019.

[4] G. N. Yannakakis and J. Togelius, *Artificial Intelligence and Games*. Springer Nature, 2018, vol. 2.

[5] D. McDuff, R. el Kaliouby, D. Demirdjian, and R. Picard, "Predicting online media effectiveness based on smile responsesgathered over the internet," in *Image and Vision Computing*, 2014.

[6] C. Ringer and M. A. Nicolaou, "Deep unsupervised multi-view detection of video game stream highlights," in *Proceedings of the Intl. Conference on the Foundations of Digital Games*, 2018.

[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[8] L. Zafeiriou, S. Zafeiriou, and M. Pantic, "Deep analysis of facial behavioral dynamics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.

[9] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation," in *Proceedings of the IEEE Intl. Conference on Affective Computing and Intelligent Interaction*, 2017, pp. 158–163.

[10] K. Makantasis, A. Liapis, and G. N. Yannakakis, "From pixels to affect: A study on games and player experience," in *Proceedings of the IEEE Intl. Conference on Affective Computing and Intelligent Interaction*, 2019.

[11] R. Picard, "Affective computing," MIT, Tech. Rep., 1995.

[12] G. Bryant and H. C. Barrett, "Vocal emotion recognition across disparate cultures," *Journal of Cognition and Culture*, vol. 8, no. 1-2, pp. 135–148, 2008.

[13] R. A. Calvo, S. D'Mello, J. M. Gratch, and A. Kappas, *The Oxford handbook of affective computing*. Oxford University Press, USA, 2015.

[14] S. Pini, O. B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, and B. Huet, "Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild," in *Proceedings of the ACM Intl. Conference on Multimodal Interaction*, 2017, pp. 536–543.

[15] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on affective computing*, vol. 1, no. 1, pp. 18–37, 2010.

[16] Z. Ambadar, J. Schooler, and J. Cohn, "Deciphering the enigmatic face. the importance of facial dynamics in interpreting subtle facial expressions," *Psychological Science*, vol. 16, 2005.

[17] J. Bassili, "Emotion recognition: The role of facial movement and the relative importance of upper and lower areas of the face," *Journal of personality and social psychology*, vol. 37, no. 11, 1979.

[18] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[19] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (CERT)," in *Proceedings of the IEEE Intl. Conference on Automatic Face & Gesture Recognition*, 2011.

[20] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan, "Automatic recognition of facial actions in spontaneous expressions," *Journal of Multimedia*, vol. 1, no. 6, 2006.

[21] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," in *Proceedings of the IEEE Intl. Conference on Affective Computing and Intelligent Interaction*, 2007.

[22] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed, "Automatic recognition of non-acted affective postures," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 4, pp. 1027–1038, 2011.

[23] D. Glowinski, A. Camurri, G. Volpe, N. Dael, and K. Scherer, "Technique for automatic emotion recognition by body gesture analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2008.

[24] J. M. Montepare, S. B. Goldstein, and A. Clausen, "The identification of emotions from gait information," *Journal of Nonverbal Behavior*, vol. 11, no. 1, pp. 33–42, 1987.

[25] S. Li, L. Cui, C. Zhu, B. Li, N. Zhao, and T. Zhu, "Emotion recognition using Kinect motion capture data of human gaits," *PeerJ*, 2016.

[26] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1-2, pp. 227–256, 2003.

[27] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient mfcc extraction method in speech recognition," in *Proceedings of the IEEE Intl. symposium on circuits and systems*, 2006.

[28] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the Intl. Conference on multimodal interaction*, 2014, pp. 494–501.

[29] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[31] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the ACM on Intl. Conference on Multimodal Interaction*, 2015, pp. 451–458.

[32] T. Ahonen, A. Hadid, and M. Pietikäinen, "Face recognition with local binary patterns," in *Proceedings of the European conference on computer vision*, 2004, pp. 469–481.

[33] Y. Baveye, E. Dellandrea, C. Chamaret, and L. Chen, "Deep learning vs. kernel methods: Performance for emotion prediction in videos," in *Proceedings of the IEEE Intl. Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 77–83.

[34] X. Zhang, L. Zhang, X.-J. Wang, and H.-Y. Shum, "Finding celebrities in billions of web images," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 995–1007, 2012.

[35] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *Proceedings of the Intl. Conference on Neural Information Processing*, 2013, pp. 117–124.

[36] S. Zafeiriou, D. Kollias, M. A. Nicolaou, A. Papaioannou, G. Zhao, and I. Kotsia, "Aff-wild: Valence and arousal 'in-the-wild' challenge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 34–41.

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

[38] D. Kollias, M. A. Nicolaou, I. Kotsia, G. Zhao, and S. Zafeiriou, "Recognition of affect in the wild using deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 26–33.

[39] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[40] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proceedings of the ACM Intl. Conference on Multimodal Interaction*, 2016, pp. 506–513.

[41] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the ACM Intl. Conference on Multimodal Interaction*, 2015, pp. 467–474.

[42] P. Tzirakis, S. Zafeiriou, and B. Schuller, "Real-world automatic continuous affect recognition from audiovisual signals," in *Multimodal Behavior Analysis in the Wild*. Elsevier, 2019.

[43] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. André, "Player modeling," in *Artificial and Computational Intelligence in Games (Dagstuhl Seminar 12191).*, 2012, pp. 45–59.

[44] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience in super mario bros," in *Proc. of the Intl. Conf. on Computational Intelligence and Games*, 2009.

[45] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE Transactions on System, Man and Cybernetics*, vol. 43, no. 6, 2013.

[46] D. Melhart, D. Gravina, and G. N. Yannakakis, "Moment-to-moment Engagement Prediction through the Eyes of the Observer: PUBG Streaming on Twitch," in *Proceedings of the Intl. Conference on the Foundations of Digital Games*, 2020.

[47] 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.

[48] H. P. Martinez, G. N. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!" *IEEE transactions on affective computing*, vol. 5, no. 3, pp. 314–326, 2014.

[49] E. Camilleri, G. N. Yannakakis, and A. Liapis, "Towards general models of player affect," in *Proceedings of the IEEE Intl. Conference on Affective Computing and Intelligent Interaction*, 2017, pp. 333–339.

[50] M. Guzdial, N. Sturtevant, and B. Li, "Deep static and dynamic level analysis: A study on infinite mario," in *Proceedings of the AIIDE workshop on Experimental AI in Games*, 2016.

[51] T. Kannetis, A. Potamianos, and G. N. Yannakakis, "Fantasy, curiosity and challenge as adaptation indicators in multimodal dialogue systems for preschoolers," in *Proceedings of the 2nd Workshop on Child, Computer and Interaction*, 2009, pp. 1–6.

[52] G. N. Yannakakis, H. P. Martinez, and M. Garbarino, "Psychophysiology in games," in *Emotion in games*. Springer, 2016, pp. 119–137.

[53] S. Wang and Q. Ji, "Video affective content analysis: a survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 410–430, 2015.

[54] Y. Baveye, C. Chamaret, E. Dellandréa, and L. Chen, "Affective video content analysis: A multidisciplinary insight," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 396–409, 2017.

[55] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE transactions on multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[56] L. Canini, S. Benini, and R. Leonardi, "Affective recommendation of movies based on selected connotative features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 4, pp. 636–647, 2012.

[57] Y. Cui, S. Luo, Q. Tian, S. Zhang, Y. Peng, L. Jiang, and J. S. Jin, "Mutual information-based emotion recognition," in *The Era of Interactive Media*. Springer, 2013, pp. 471–479.

[58] A. Yazdani, E. Skodras, N. Fakotakis, and T. Ebrahimi, "Multimedia content analysis for emotional characterization of music video clips," *EURASIP Journal on Image and Video Processing*, vol. 2013, no. 1, p. 26, 2013.

[59] M. Xu, C. Xu, X. He, J. S. Jin, S. Luo, and Y. Rui, "Hierarchical affective content analysis in arousal and valence dimensions," *Signal Processing*, vol. 93, no. 8, pp. 2140–2150, 2013.

[60] Y. Yi and H. Wang, "Multi-modal learning for affective content analysis in movies," *Multimedia Tools and Applications*, vol. 78, no. 10, pp. 13 331–13 350, 2019.

[61] S. Wang, G. Peng, Z. Zheng, and Z. Xu, "Capturing emotion distribution for multimedia emotion tagging," *IEEE Transactions on Affective Computing*, 2019.

[62] A. Mitenkova, J. Kossaifi, Y. Panagakis, and M. Pantic, "Valence and arousal estimation in-the-wild with tensor methods," in *Proceedings of the IEEE Intl. Conference on Automatic Face & Gesture Recognition (FG 2019)*. IEEE, 2019, pp. 1–7.

[63] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," British Machine Vision Association, 2015.

[64] Y. Zhu, Z. Chen, and F. Wu, "Affective video content analysis via multimodal deep quality embedding network," *IEEE Transactions on Affective Computing*, 2020.

[65] D. Melhart, A. Liapis, and G. N. Yannakakis, "Pagan: Video affect annotation made easy," in *Proceedings of the IEEE Intl. Conference on Affective Computing and Intelligent Interaction*, 2019.

[66] G. N. Yannakakis, H. P. Martínez, and A. Jhala, "Towards affective camera control in games," *User Modeling and User-Adapted Interaction*, vol. 20, no. 4, pp. 313–340, 2010.

[67] Y. Knight, H. P. Martínez, and G. N. Yannakakis, "Space maze: Experience-driven game camera control," in *Proceedings of the Intl. Conference on the Foundations of Digital Games*, 2013, pp. 427–428.

[68] P. Lopes, A. Liapis, and G. N. Yannakakis, "Modelling affect for horror soundscapes," *IEEE Transactions of Affective Computing*, vol. 10, no. 2, pp. 209–222, 2017.

[69] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the Intl. Conference on Knowledge Discovery and Data Mining*, 1994, pp. 359–370.

[70] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, 2018.

[71] K. Schindler and L. J. Van Gool, "Action snippets: How many frames does human action recognition require?" in *Proceedings of the IEEE Intl. Conference on Computer Vision and Pattern Recognition*, 2008, pp. 3025–3032.

[72] K. Makantasis, A. Doulamis, N. Doulamis, and K. Psychas, "Deep learning based human behavior recognition in industrial workflows," in *Proceedings of the IEEE Intl. Conference on Image Processing (ICIP)*, 2016, pp. 1609–1613.

[73] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstral coefficients on the power spectrum," in *Proceedings of the IEEE Intl. Conference on Acoustics, Speech, and Signal Processing*, 2001, pp. 73–76.

[74] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez, "Automatic detection of pathological voices using complexity measures, noise parameters, and mel-cepstral coefficients," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 2, pp. 370–379, 2010.

[75] T. C. Nagavi, S. Anusha, P. Monisha, and S. Poornima, "Content based audio retrieval with mfcc feature extraction, clustering and sort-merge techniques," in *Proceedings of the IEEE Intl. Conference on Computing, Communications and Networking Technologies*, 2013.

[76] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proceedings of the Intl. Conference on Machine learning*, 2005, pp. 89–96.

[77] M. Kearns and D. Ron, "Algorithmic stability and sanity-check bounds for leave-one-out cross-validation," *Neural computation*, vol. 11, no. 6, pp. 1427–1453, 1999.

[78] H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of Measurement and Statistics*, pp. 508–510, 2007.

[79] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE Intl. Conference on Computer Vision*, 2017, pp. 618–626.

[80] I. Ekman, "Meaningful noise: Understanding sound effects in computer games," in *Proceedings of Digital Arts and Cultures*, 2005.

[81] K. Makantasis, E. Protopapadakis, A. Doulamis, N. Doulamis, and C. Loupos, "Deep convolutional neural networks for efficient vision based tunnel inspection," in *Proceedings of the IEEE Intl. Conference on Intelligent Computer Communication and Processing (ICCP)*, 2015, pp. 335–342.

[82] J. Zhu, A. Liapis, S. Risi, R. Bidarra, and G. M. Youngblood, "Explainable ai for designers: A human-centered perspective on mixed-initiative co-creation," in *Proceedings of the IEEE Conference on Computational Intelligence and Games*, 2018.
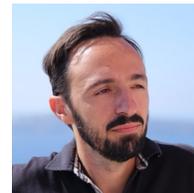
Antonios Liapis is a Lecturer at the Institute of Digital Games, University of Malta, where he bridges the gap between game technology and game design in courses focusing on human-computer creativity, digital prototyping and game development. He received the Ph.D. degree in Information Technology from the IT University of Copenhagen in 2014. His research focuses on Artificial Intelligence as an autonomous creator or as a facilitator of human creativity. His work includes computationally intelligent tools for game design, user models for the design process, gameplay, or visual preference, as well as evolutionary computation and deep learning. He has published over 100 journal, conference and workshop papers on these topics, and has received several awards for his research contributions and reviewing effort. He has served as general chair in four conferences (EvoMusArt 2018-2019, GALA 2019, FDG 2020), and as a Guest Editor in three special issues in IEEE TRANSACTIONS OF GAMES, ACM Journal on Computing and Cultural Heritage, and the International Computer Games Association Journal.

Georgios N. Yannakakis (S'04-M'05-SM'14) is a Professor and Director of the Institute of Digital Games, University of Malta, and a co-founder of modl.ai. He received the Ph.D. degree in Informatics from the University of Edinburgh in 2006. He does research at the crossroads of artificial intelligence, computational creativity, affective computing, advanced game technology, and human-computer interaction. He has published more than 260 papers in the aforementioned fields and his work has been cited broadly. He is currently an Associate Editor of the IEEE TRANSACTIONS ON GAMES and the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION and used to be Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING journal. He has been the General Chair of key conferences in the area of game artificial intelligence (IEEE CIG 2010) and games research (FDG'13, '20). Among the several rewards he has received for journal and conference publications he is the recipient of the *IEEE Transactions on Affective Computing Most Influential Paper Award* and the *ACII 2017 Best Paper Award*. He is a senior member of the IEEE.

Konstantinos Makantasis received his computer engineering diploma and his Master degree from the Technical university of Crete (TUC, Greece). In 2016 Dr. Makantasis received his PhD from the same school working on detection and semantic analysis of objects and events through visual cues. Currently, he is a MSCA IF Widening Fellow at the Institute of Digital Games, University of Malta working on tensor-based machine learning methods for affect modeling. He is mostly involved and interested in computer vision, machine learning / pattern recognition and probabilistic programming. He has more than 45 publications in international journals and conferences on computer vision, signal and image processing and machine learning.