Journal on Multimodal User Interfaces (2021) 15:373–391
https://doi.org/10.1007/s12193-021-00364-0

ORIGINAL PAPER

MUMBAI: multi-person, multimodal board game affect and interaction analysis dataset

Metehan Doyran[1] · Arjan Schimmel[1] · Pınar Baki[2] · Kübra Ergin[3] · Batıkan Türkmen[2] · Almıla Akdağ Salah[1,2,4] · Sander C. J. Bakkes[1] · Heysem Kaya[1] · Ronald Poppe[1] · Albert Ali Salah[1]

Received: 24 May 2020 / Accepted: 22 January 2021 / Published online: 22 February 2021

Abstract

Board games are fertile grounds for the display of social signals, and they provide insights into psychological indicators in multi-person interactions. In this work, we introduce a new dataset collected from four-player board game sessions, recorded via multiple cameras, and containing over 46 hours of visual material. The new MUMBAI dataset is extensively annotated with emotional moments for all game sessions. Additional data comes from personality and game experience questionnaires. Our four-person setup allows the investigation of non-verbal interactions beyond dyadic settings. We present three benchmarks for expression detection and emotion classification and discuss potential research questions for the analysis of social interactions and group dynamics during board games.

Keywords Affective computing · Facial expression analysis · Board games · Social interactions · Group dynamics · Multimodal interaction · Game experience

1 Introduction

Multiplayer board games are excellent tools to stimulate specific interactions for both children and adults. Players of board games exhibit a wealth of social signals, related to decisions made in the game, game outcome, and game progress. Consequently, multi-player board games can be used to study affective responses to game events and other players and emotion contagion, possibly in interaction with personal and interpersonal factors. Board games have been adopted for therapeutic purposes by psychologists [49,72], for example, to assess behavioural patterns, cognitive abilities, and attitudes [26,53], but the setting allows other analyses in the areas of affective computing and social signal processing, as well as for applications in the area of education and serious games.

Using board games for affect analysis presents several methodological challenges. First, games elicit valuable behavioural and affective responses, but these responses

✉ Metehan Doyran
m.doyran@uu.nl

Arjan Schimmel
a.schimmel@uu.nl

Pınar Baki
pinar.baki@boun.edu.tr

Kübra Ergin
kubraergin3@gmail.com

Batıkan Türkmen
batikan.turkmen@boun.edu.tr

Almıla Akdağ Salah
a.a.akdag@uu.nl

Sander C. J. Bakkes
s.c.j.bakkes@uu.nl

Heysem Kaya
h.kaya@uu.nl

Ronald Poppe
r.w.poppe@uu.nl

Albert Ali Salah
a.a.salah@uu.nl; salah@boun.edu.tr

[1] Department of Information and Computing Sciences, Utrecht University, Princetonplein 5, 3584CC Utrecht, The Netherlands

[2] Department of Computer Engineering, Boğaziçi University, 34342 Bebek, Istanbul, Turkey

[3] Sahibinden.com, Istanbul, Turkey

[4] Department of Industrial Design Engineering, Delft Technical University, Postbus 5, 2600 AA Delft, The Netherlands

are relatively scarce and brief [27]. Second, exhibited play behaviour typically cannot be annotated objectively. There is a fair amount of subjectivity as a result of the annotator's insight into human affect and decision-making processes, but also player personality and motivation, the state of the game, and the dynamics of the social context. Finally, manually coding a player's behaviour is inherently time-intensive. As such, while the potential for employing board games as an analysis tool is clear, at present it generally is time-consuming for a therapist to fully exploit this potential.

An effective computational approach to behavioural analysis would mitigate these challenges. With rapid developments in automated behaviour analysis, it is becoming feasible to automatically process large amounts of play observations and, if needed, prepare indices for therapists. This has important advantages. First, depending on the observed behaviour, a limited number of observations may suffice for accurate analysis. Second, multiple modalities such as the face, body, and voice can be analysed simultaneously. Third, automated analysis can be expected to be significantly more efficient than manual analysis. There are, however, also drawbacks to fully automatic analysis, including limited generalization capabilities of such algorithms, their dependence on rich annotations, and the lack of a semantic grounding that complicates the interpretation of rare events and idiosyncratic displays.

In this paper, we investigate the feasibility of automated analysis of multimodal behaviour in multiplayer board games. We focus on adults interacting with each other while playing different types of board games. We introduce the Multi-Person, Multimodal Board Game Affect and Interaction Analysis Dataset (MUMBAI), a dataset with recordings of 62 game sessions, each involving four players. Our setup includes the recording of videos of interacting players and the game board, the collection of personality traits for each player, and an assessment of the game experience after each played game. Using MUMBAI, we investigate to what extent we can derive information on the emotional states and social interactions of adults from recordings of their behaviour. We discuss the possibility of analysing the relationship between self-reported personality traits, apparent gameplay behaviour, and self-reported game experience.

This paper makes the following contributions:

1. We introduce a multi-person dataset of recorded game sessions with two sets of segment-level annotations of affect, self-reported game experience questionnaires, game outcomes, and additional personality tests. This is the first publicly available database for board game settings that goes beyond dyadic interactions[1].
2. We present benchmark evaluation results using state-of-the-art feature extraction and emotional expression classification methods.
3. We investigate the feasibility of automated analysis of multimodal behavior in multiplayer board games, as well as the link between personality traits and game experience.

This paper significantly extends our eNTERFACE'19 paper [69] by completing the annotations of the database, providing evaluations with new feature sets, presenting three benchmark tests, and in-depth discussion of several aspects of the database.

We proceed with a discussion of related work on multimodal behaviour analysis for games and publicly available datasets on multi-person interaction. Section 3 introduces our dataset, MUMBAI. In Sect. 4, the annotation and feature extraction processes are detailed. We present benchmark results for several automated analysis tasks in Sect. 5 and conclude with a discussion in Sect. 6.

2 Related work

Research in the domain of *affective computing* focuses on equipping computer programs with the ability to sense affective cues exhibited by humans [56]. The application of affective analysis during gameplay can directly benefit game design and evaluation activities, as well as investigations that use games as rich platforms for observing human behaviour, such as by psychologists who observe, describe, and quantify behaviours during long-term therapy. Since the type of features that can be automatically derived from human behaviour analysis is vast [64,65], a comprehensive review is not included here. Rather, we focus on the automatic analysis of player behaviour during co-located gameplay.

We focus on a scenario where multiple persons are sitting around a table to play a game with materials on it. The most interesting behaviours during such a scenario involve responses to the game events or other players, such as displays of frustration, anger, elation, boredom, excitement, disappointment, concentration, puzzlement, expectation, pride, and shame. There are affective constructs (such as emotion models) that can be used to describe the state of a person during a potentially emotional interaction. Automatic analysis focuses on the behavioural indicators that may point out the presence of affective states, and more prominently, on the apparent exhibits of affective and communication-related cues. The social context is relevant for the interpretation of such indicators, and -at least for the moment- the expertise required to properly relate indicators to constructs is mostly beyond automatic analysis tools, except for very simple and straightforward affective states. However, these tools can

[1] The videos, annotations, as well as questionnaires are made publicly available at https://github.com/dmetehan/MUMBAI.

provide the experts with valuable insights and save time in analysis.

In the remainder of this section, we will give a brief survey on multimodal behaviour analysis for games and play, focusing on facial affect as the most prominent source of signals.

2.1 Multimodal behaviour analysis

The face is regarded as the most expressive part of the body [54], which provides the most discriminative features for a range of affect recognition tasks [23,29,36], and there are works specialised in processing faces during gameplay or other activities such as problem-solving (e.g. [37,44]). Gaze is in particular shown to be a good indicator of a person's engagement with an activity [71,74]. On the other hand, the use of the bodily motions alone in affect recognition is less common than using facial expressions [62]. Some expressions may be better inferred from the body than the face [2,10]. Kleinsmith and Bianchi–Berthouze have surveyed the different techniques that have been used in the field of affective body expression recognition [40].

The challenges of affect analysis with a broad range of affective states include the relatively rare observation and subtle expression of some affective behaviours, as well as the imbalance of sample distributions [36,45]. We stress the importance of studying affective behaviours in natural conditions, in contrast to the analysis of acted material. Facial displays by themselves may be insufficient for fully capturing these states automatically, as the face is also deformed via non-emotional speech. With the use of multiple modalities, a system can increase performance by taking into account the redundant and complementary information in the various channels. A combination of facial and bodily modalities is most widely used for automatic analysis of interactions [70]. Filntisis [24] et al. addressed affect recognition during child–robot interaction and illustrated how the combination of face and bodily cues in a machine learning algorithm could yield better results than the use of a single modality. A similar finding in the application domain of serious games was reported by Psaltis et al. [60] where decision level fusion was employed and the individual modalities were fused with the help of confidence levels .

For facial expressions, Ekman and Friesen introduced the Facial Action Coding System (FACS) [21,22], which provides an objective way to describe facial movements of the face in terms of action units. However, there is no clear and unambiguous mapping from action units to expressions; there are only indicators for a number of expressions, some strongly correlated, and some not. For example, the upwards movement of lip corners, coded as AU12, is a good indicator of a smile. Yet it does not immediately tell us whether it is due to genuine enjoyment, or used as a social back-channel signal [17]. Manual FACS coding is a lengthy and expensive process.

Currently, there is no widely-adopted coding scheme for bodily behaviour. Body language associated with certain emotions is usually described by how specific body parts move, but it is much more idiosyncratic [13,76]. Movement and posture are the most common meaningful low-level features that are widely used [40]. In addition to these features, machine learning researchers use representations that rely on feature extraction (e.g. via deep learning), which are more difficult to interpret than movement and posture [58].

How to represent affect is still up for debate [62]. In 1981, Kleinginna created an overview of the definitions of emotion that existed until then [39], and listed 92 different definitions. Since then, there have been many works on affect and what it precisely is (e.g. [52,73]). A working definition is given by Desmet [15]: *"emotions are best treated as a multifaceted phenomenon consisting of the following components: behavioural reactions (e.g. approaching), expressive reactions (e.g. smiling), physiological reactions (e.g. heart pounding), and subjective feelings (e.g. feeling amused)"*. This definition agrees with our aims, as in this work, we create a dataset where participants' subjective feelings during gameplay and their expressive reactions can be analysed more closely.

2.2 Multimodal behaviour analysis for games

In the domain of (video) game playing, numerous investigators have focused on assessing the affective state of the involved players, including for clinical purposes [18,19]. For human players this task is important, as typically the overall experience is moderated by group dynamics, and as individual players may often achieve a competitive advantage when they succeed in correctly assessing another player's affective state (cf. numerous *eSports* (electronic sports) games and *Poker*). Indeed, as with card games, players often seek to extract hidden information from their opponents by analysing social signals such as speech, body motion, and facial expressions [34].

In this context, one may observe that a certain body of work has focused on automatically modelling player stress levels. For instance, Mavromoustakos-Blom et al. [7] proposed a multi-modal approach towards stress response modeling in the competitive *League of Legends* video game; a leading eSports title, where they collected wearable physiological sensor data, mouse & keyboard logs, and in-game data in order to study the relationship between player stress responses and in-game behaviour.

Analogously, studies with competitive video-game playing have shown that there is a strong correlation between players' cognitive and in-game performance [4,8,75]. More specifically, Bonny et al. [8] examined the correlation between cognitive and in-game player skills, revealing that players with higher levels of gaming expertise respond faster to decisions that rely on spatial memory. In addition, it was shown that in video games the subjectively perceived difficulty can be assessed automatically through facial expression analysis [50]. In a related work, Mavromoustakos-Blom et al. employed recordings of players' facial expressions during competitive *Hearthstone* games to analyse the correlation between in-game player affective responses and subjective post-game self-reports [51]. Correlation analyses between in-game and post-game variables reveal that players' facial expressions and eye gaze measurements are correlated to both players' attention to the opponent and their mood influenced by the opponent.

Finally, a notable application area of multimodal behaviour analysis for games is *training* of professional players. For instance, Korotin et al. [41] have collected affective data from professional *Counter-Strike* players for this purpose. Hung et al. [33] are presently developing a player training tool which is to recognise player affective states and provide personalised training sessions. Indeed, this research direction is similar to simulation training, in which the affective states of groups of participants are analysed [46,78] in order to provide more effective learning experiences.

2.3 Game behaviour datasets

Some existing datasets provide researchers with audio, visual, or audiovisual data to aid research on affective computing and social interaction analysis. Mimicry database [77], Static Multimodal Dyadic Behavior (MMDB) database [61], Tower Game Dataset [67], PInSoRo database [43], Haggling database recorded within the CMU Panoptic Studio [35], and GAME-ON dataset [47] are among the most important resources to study social interactions during gameplay. The interactions in these datasets happen between adults, between a child and an adult, or a child and a robot. Table 1 summarises available game behaviour datasets and their characteristics.

The Multimodal Mimicry database [77] is recorded during two experiments: a discussion on a political topic and a role-playing game, respectively. The annotation consists of a number of social signaling cues (mimicry) and "conscious/non-conscious" labels illustrating the status of these cues. Mimicry is an automatic process that regulates social interactions [48].

The Tower Game Dataset [67] consists of audio-visual recordings of two players, and the work focuses on the joint attention and entertainment during a cooperative tower building game. Although the game allows a rich interaction setting, the wearable cameras on the participants may create an unnatural feeling and therefore affect the interaction.

The Multimodal Dyadic Behavior (MMDB) Dataset [61] focuses on dyadic interactions between adults and 15- to 30-month old children. The examiner performs a set of scripted actions, such as showing a ball or a book, and the child's responses are labelled into 17 binary behaviour classes. The dataset provides a valuable multimodal resource for social engagement analysis for infants.

The PInSoRo dataset [43] has recordings of both child–child and child–robot interactions during free play without any adult intervention. This dataset is annotated using three different social interaction codes: task engagement, social engagement, and social attitude, respectively.

Haggling DB [35] has multiview audiovisual data of triadic interactions between two sellers and a buyer in a haggling setting. The prediction tasks are for speaking status, social formations, and body gestures, using both interpersonal and intrapersonal features.

The GAME-ON dataset is specifically collected for studying task cohesion and social cohesion amongst friends [47]. Three friends play an escape game together and try to cooperate with each other in order to win the game. Additionally, the authors have collected self-reported questionnaires, consisting of a participant's emotional state and their perception of leadership, warmth, and competence of their group members.

The datasets we reviewed here, as well as other social interaction datasets in the literature, are mostly based on dyadic and triadic interactions. Some of them are scripted and some use wearables to capture data. The Idiap Wolf Corpus [32] is an interesting game corpus that has audio-visual recordings of 8–12 people during a role-playing game, where roles vary between deceptive and non-deceptive. Conversation is the main interaction channel and the deceptive players try to hide their expressions in order to succeed.

In the Multi-Person, Multimodal Board Game Affect and Interaction Analysis Dataset introduced by this paper, four board game players are recorded simultaneously during each session, which affords for more complex interactions between participants. We have focused on *cooperative board games*, in which players win and lose the game together. These games create a rich setting for social interaction, and there are plenty of opportunities to observe emotional contagion. We chose the Magic Maze game as our main board-game for our data collection, because of its rules that forbid speaking for most of the game and further enforce people to use non-verbal communication signals like face and body expressions. We also include other games (both cooperative and competitive) to allow comparative analyses.

Table 1 Recent game behaviour datasets

Name	Year	Modality	Subj.	Subj. per sess.	Sessions	Annotations	Labels
The Idiap Wolf Corpus [32]	2010	V + A	36	8-12	15	Semi-automatic, discrete	Speaking segments, deceptive/non-deceptive roles, decisions in the game
Mimicry Database [77]	2011	V + A	40	2	54	Semi-automatic, discrete	Behavioural expressions (smile, head nod, head shake, body leaning away, body leaning forward) mimicry/non mimicry conscious/unconscious
The Tower Game Dataset [67]	2015	V + A	39	2	112	Manual, continuous	Eye gaze, body language, simultaneous movement, tempo similarity coordination and imitation are rated using a six-point Likert scale
MMDB Dataset [61]	2013	V + A + P	121	2	160	Manual, discrete	Engagement, responsiveness, attention shifts, facial expressions, gestures and vocalizations
PInSoRo Dataset [43]	2018	V + A	120	1 or 2	75	Manual, discrete	Task engagement, social engagement, social attitude
Haggling Database [35]	2019	V + A	120	3	180	Manual, automatic, continuous	3D full body motion, 3D point clouds, speaking status, and timing
GAME-ON Dataset [47]	2020	V + A + M	51	3	17	Self-reported	Group environment, warmth & competence, competitivity, emotions, leadership, motivation
MUMBAI	*2020*	*V*	*58*	*4*	*62*	*Manual, discrete, self-reported*	*Emotional moments, game related emotions, game experience questionnaire, selfreported personality test*

V video, *A* audio, *P* physiological, *M* motion capture

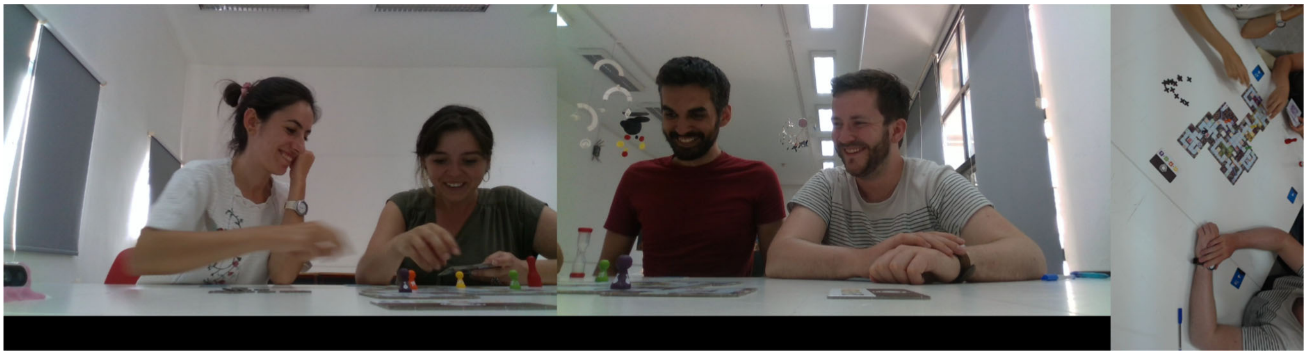


Fig. 1 A screenshot from the recording stream, where four players respond to a player's mistake

Table 2 The games played in MUMBAI

Type	Games	Sessions	Mins.	Players
Coop.	Magic Maze	39	405	57
	Pandemic	2	78	4
	The Mind	1	6	4
Comp.	Qwixx	10	203	17
	Kingdomino	8	140	17
	King of Tokyo	2	73	5

(*Coop* cooperative, *Comp* competitive, *Mins* minutes)

3 MUMBAI dataset

In this section, we introduce the Multi-Person, Multimodal Board Game Affect and Interaction Analysis Dataset (MUMBAI). The dataset features participants playing cooperative (co-op) and competitive board games. Every game session consisted of four participants, recorded by two separate cameras, and an additional recording of the board game itself to allow for the detection of in-game events. Every participant filled in a HEXACO personality test [3] and after every game, they completed the in-game and social modules of the Game Experience Questionnaire (GEQ) [57]. In total, there are 62 recorded sessions. MUMBAI contains manual annotations of affect, self-reported personality, and game experience data, as well as automatically extracted facial features and bodily landmarks.

The study received ethical approval from the Internal Review Board for Ethical Questions by the Scientific Ethical Committee of the university. In this section, we describe the games, participants, recordings, annotations, and questionnaires.

3.1 Games

According to Schaefer and Reid [68], there are four categories of games for therapeutic use: communication games, problem-solving games, ego-enhancing games, and socialization games. MUMBAI features two types: communication games and ego-enhancing games, respectively. In communication games, competition plays a smaller role, and inter-player communication is the key [79]. Ego-enhancing games trigger stress, feelings of competition, and challenge. This potentially leads to conflicts between game players, creating emotional states such as frustration, disappointment, anger, but also relief, triumph, and elation.

Each game session consisted of four participants that played one of six multiplayer games (see Table 2). Some people participated in more than one game session (See Fig. 5). The game that was played was chosen by the participants. Before playing, the rules of the game were explained by the experimenters. We summarize the six games.

Magic Maze is the most played game in MUMBAI. It is a cooperative game, players win by collectively managing four game characters to explore a maze. These characters need to steal items from specific locations in the maze and use specific escape locations to complete the task against a running hourglass. Players do not take turns and are allowed to move whenever they can. Each player has a complementary set of moves. The game is played in real-time and if the hourglass (green circle in Fig. 2) runs out, the players lose the game. Players are not allowed to speak with each other but can place a big red cone (red circle in Fig. 2) in front of another player to nudge that player to do something. In Magic Maze, players naturally show emotions due to the tension generated by the game. The most stress-related emotions arise when the hourglass is about to run out. More positive responses are typically observed when using the red cone, as it is usually placed in front of a player with a lot of enthusiasm. The player who gets it often displays frustration or confusion (e.g. left player in Fig. 2). A game of Magic Maze takes around 10–15 min.

Pandemic is a cooperative game where players try to save the world from an epidemic. Players work together to keep the outbreaks of diseases under control, while at the same time finding the cures for these diseases. The game decides where

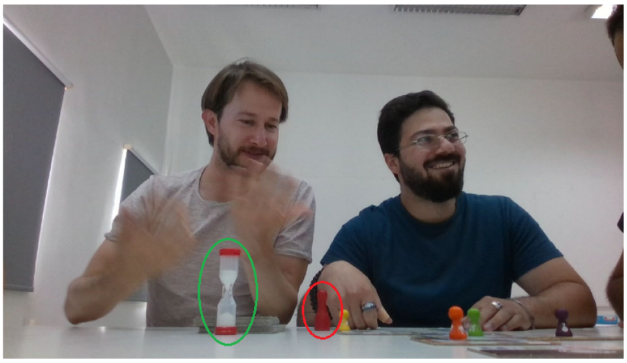


Fig. 2 Moment in a Magic Maze session, where the red cone was just placed in front of the player on the left, who is confused about what is expected

the next outbreak is, based on a deck of cards which players need to draw from every turn. This creates a lot of tension in these moments because depending on which card is drawn, the game can swing in favor of the players or it will become harder to win. As each player has a different unique role with specific advantages, and players need to discuss their moves with others to get the best results. A game of Pandemic takes approximately 40 min.

The Mind requires players to work together to play cards in the right order. Every player receives the same number of cards, ranging from 1 to 100 and without duplicates. The players need to play the lowest card, compared to all the cards in every hand, on the table. After that, the second-lowest card is played, and so on until all cards are played in order. The problem is that the players do not know each others' cards and they are not allowed to speak with each other. Players are not allowed to talk to each other, which often results in hilarious moments when players look intensely at each other and wait for the others to do something.

Qwixx is a competitive game, primarily based on luck. Players throw dice every turn and take some of them to cross off numbers, based on certain restrictions, on their own sheet. Each action disables a number of future actions, thus the game requires the players to calculate and take risks. At the end of the game, the player with the most crosses wins. The emotions that are shown during a Qwixx game are mostly moments of surprise, both happy and sad when players see the results of the dice throw. Another commonly occurring type of emotion is 'schadenfreude,' i.e., enjoyment of another player's demise. When a player cannot cross something off, other players typically enjoy these moments.

Kingdomino is also a competitive game, where players compete to create the most valuable kingdom. Every turn, players take a piece of land to place it in their kingdoms. The pieces work just like domino stones and have similar placement restrictions. New pieces are revealed at the start of every turn. This typically evokes emotions such as positive and negative surprise. A player's choices directly influence the other players, as the piece of land can only be chosen by one player. This creates moments of friction between the players. In Kingdomino, players sometimes display boredom when a player takes a long time to think. Players also take the opportunity to talk to other players to convince them to take a certain piece.

King of Tokyo is a competitive game mostly based on luck. In each turn, a player throws some custom dice to determine what action they can perform. Based on the dice outcome, a player receives points, currency, lives, or attack points. All results are positive for the player and can be used to become stronger, except for the attacks, which can be used to damage other players. When this happens, friction often arises between players. Typically, all players pay attention to each throw of the dice. If there is a possibility to attack, players will plead and argue about who will be attacked.

3.2 Participants

In MUMBAI 58 participants were recorded at a workshop venue and a board game cafe. The participants who were recruited at the workshop venue were asked to play a board game for this project. All of them were other workshop participants. Some of those participants wanted to play more board games and came back multiple times. At the board game cafe, the participants were notified beforehand via social media groups. These participants were generally of a higher board game experience level. The participants usually stayed for a couple of games. The demographic information provided by the participants themselves consists of sex (40 male, 18 female), age (Fig. 3), and how often they play board games (Fig. 4). Participants are instructed about the rules of the games and when they should fill the questionnaires (personality test when they sign in to participate and GEQ at the end of each game session).

The participants varied from beginners to experienced players. The most experienced player assumed the role of 'game master'. If there was no experienced player, one of the authors played as the game master. The game master's role was to answer questions and to make sure the game rules were adhered to. Figure 5 shows the game count histogram where the 11+ bar consists of three game masters who participated in 28 games on average.

3.3 Recording

The setup for the recordings can be seen in Fig. 6. Players sit side-by-side in pairs at two sides of a table. Two cameras are placed opposite to record both pairs (see left and middle frames in Fig. 1). A third camera recorded the

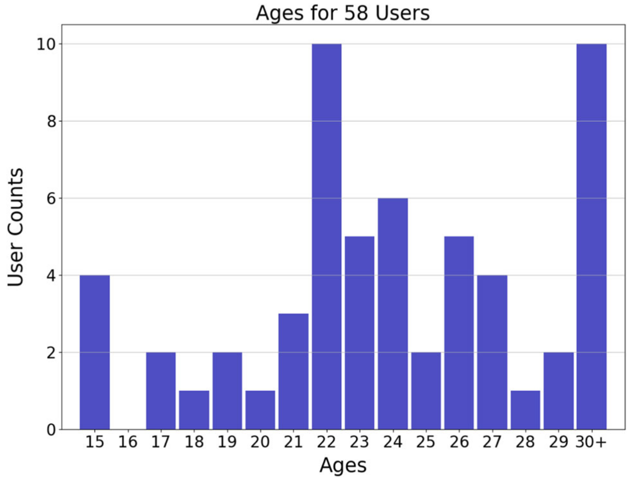


Fig. 3 Age histogram for all participants

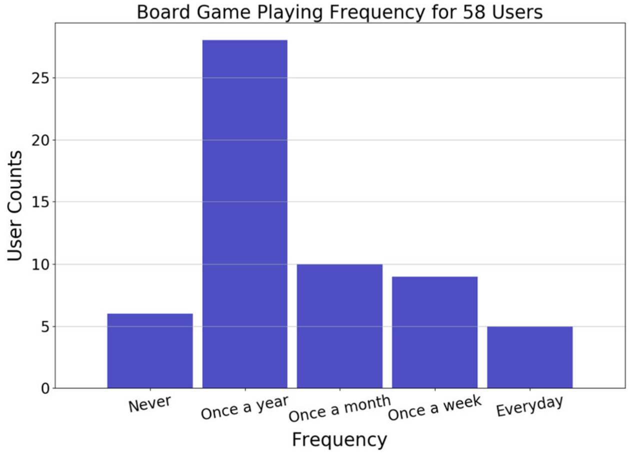


Fig. 4 Board Game playing frequency histogram for all participants

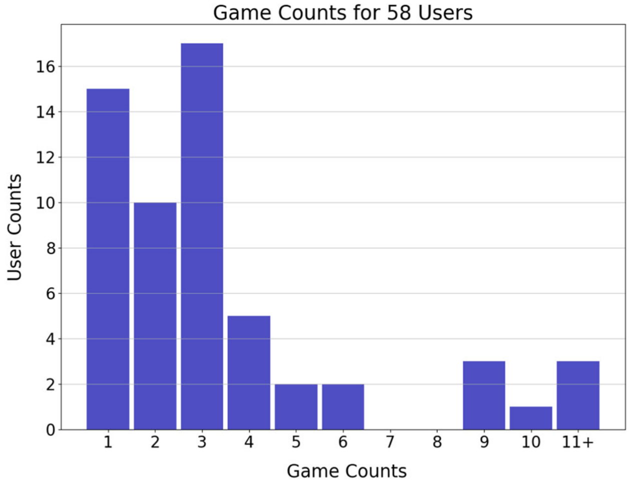


Fig. 5 Game count histogram for all participants

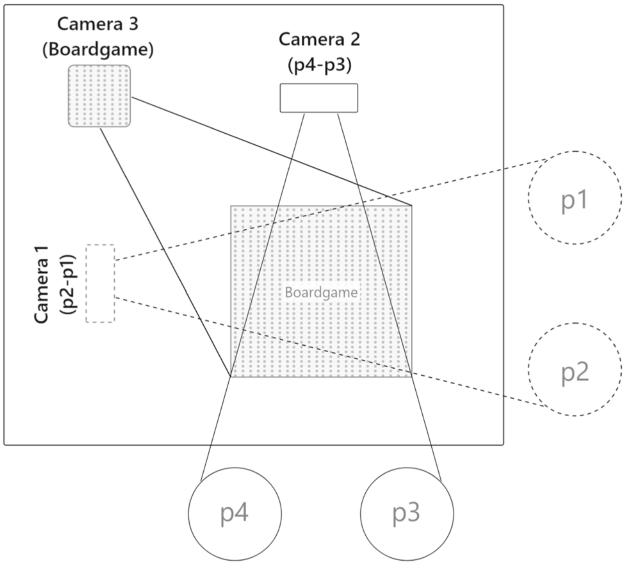


Fig. 6 Recording setup

board and was placed slightly higher to have a better view (right frame in Fig. 1). The three videos were first synchronized and then merged into a single one (Fig. 1) using Open Broadcaster Software[2] (OBS) for annotation purposes only. Automatic processing is performed on the individual streams. The videos of the participants have a resolution of 1280 × 720 and the recording of the board game state (right frame in Fig. 1) has a resolution of 800 × 448. All recordings were captured at 30 fps. We decided not to focus on the audio in the recordings, because our recordings took place in a noisy environment. Furthermore, our participants were from different nationalities and they were not using their native language.

4 Data annotation

MUMBAI contains manual, self-reported and automated annotations. Manual annotations include game outcome and player affect, and are discussed in Sect. 4.1. The self-reported personality and game experience tests are introduced in Sect. 4.2. The automated feature annotation is detailed in Sect. 4.3.

4.1 Manual annotations of affect

We provide two sets of manual, affective annotations. Set A contains segment-level annotations of expressive moments. Using ELAN[3], we annotated for each player the deviations from a neutral facial expression using seven different

[2] https://obsproject.com/.

[3] https://tla.mpi.nl/tools/tla-tools/elan/.

Table 3 Expressive moments annotation

Label	Name	Meaning
+	Positive	The most positive expressions: laughter, open mouth smiles, excitement, cheering
+?	Small positive	Somewhat positive expressions: gentle smiles, peaceful happiness
'No label'	Neutral	Represents the state of the player that is generally shown throughout the game. Each player has a different neutral state, so annotations are done considering the most occurring state of that player
–?	Small negative	Somewhat negative expressions: short frowns, lowered mouth corners, slight scowl
–	Negative	The most negative expressions: looking angry, extensive scowl, closed eyes with tightened lips
f	Focus	Not ranked in valence space. Player gives full attention to the board game: narrowed eyes and lower blink rate
f?	Small focus	Less intense version of the focus label
x	Non-game event	For example taking a call or talking with another person outside of the game

labels, see Table 3. The set of expressions important for gaming experience include pleasure, but also boredom and anxiety [66]. In particular, the latter two are important for the flow experience, as the game establishes a balance between skill and challenge [14]. Other important emotions are triumph (i.e. moments of success in the game) and frustration/confusion, particularly for learning experience [66].

Each recording was annotated by one of two annotators. To assess the reliability of the manual coding, four videos were annotated by both annotators. The annotators were free to choose starting and ending of each annotated segment. To allow for structured comparison, we used a sliding window approach with various length and stride selections to map annotations to different sets of segments. The inter-rater reliability, calculated using Cohen's Kappa [12], was 0.735 for binary neutral class vs the rest and 0.669 for all categories using 50 frames long (1.67 s) segments and 16 frames (0.53 s) stride. The label distribution for Set A annotations appears in Fig. 7. It is clear that positive displays are more common than negative ones.

We collected a second set of annotations, Set B, from 54 annotators. These annotations consider more specific game-related emotions, see Table 4. Figure 8, and Fig. 9 show instances of these emotions. Annotators used expressive moment annotations and classified them into one of four categories. Annotators used the ELAN software to view the expressive moment annotations and to classify the existing segments with their new annotations. We preferred this option over a tool where we show only the segments because, with the ELAN software, the annotators have the freedom to watch non-labelled segments to get more context about the session. Each annotator annotated three randomly selected videos. Each video is annotated by three to six annotators. The average inter-rater reliability, calculated using Fleiss' Kappa [25], was 0.381 ranging from 0.013 to 0.659 per video. To increase the annotation quality, we calculated inter-rater reliability between each pair of annotators using Cohen's Kappa and for each video, we chose the annotator pair with the highest agreement. After this process, the average inter-rater reliability increased to 0.573 ranging from 0.289 to 0.859. We combined all of the annotations into a single annotation by using the agreed labels of the annotator pair with the highest inter-rater agreement for each recording. When there was a disagreement, we performed majority voting over all annotators. For the tie situations, we chose the highest voted

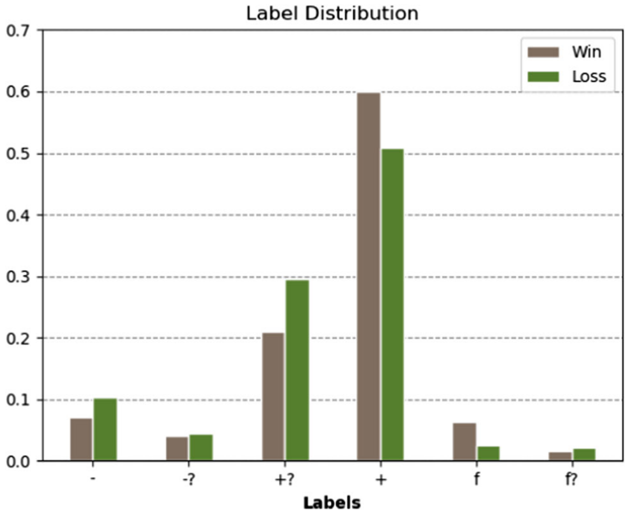


Fig. 7 Annotation Set A label distribution for won and lost games. Table 3 shows class mappings

Table 4 Game related emotion labels

Label	Emotions
A	Anxious/frustrated/disappointed/angry
B	Bored
C	Confused
D	Delighted/happy

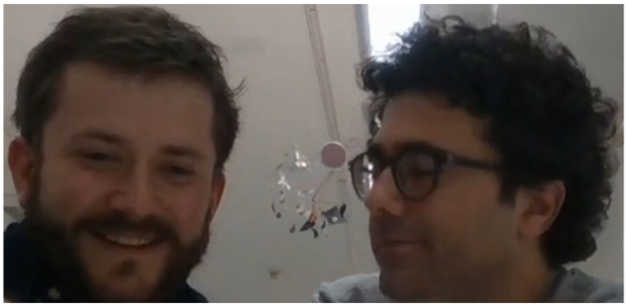


Fig. 8 Examples for delighted/happy (D) on the left and disappointed (A) on the right (see Table 4)

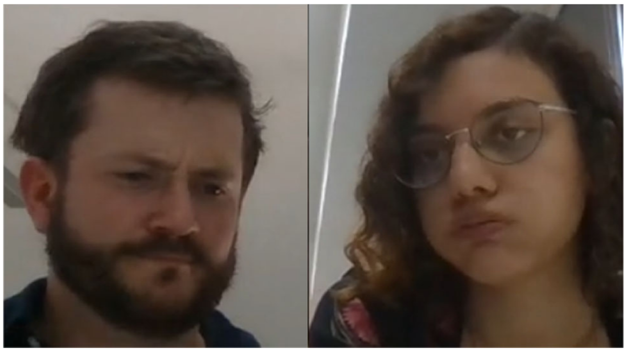


Fig. 9 Examples for confused (C) on the left and bored (B) on the right

label between the best annotator pair. Only for one segment the tie still remained and we then chose the top annotator's label.

We visualize the label distribution of Set B annotations separately for win and loss outcomes, in Fig. 10. Delighted/Happy (class D) is the dominant label. Both game outcomes have very similar label distribution. We believe that our game choice, Magic Maze, is responsible for this balance. In Magic Maze, the players are most of the time intrigued by the fast-paced gameplay, and they become only aware that they are losing the game when the time runs out. The label distribution would be different in games where the win/loss probability can be estimated by looking at the board state before the end of the game.

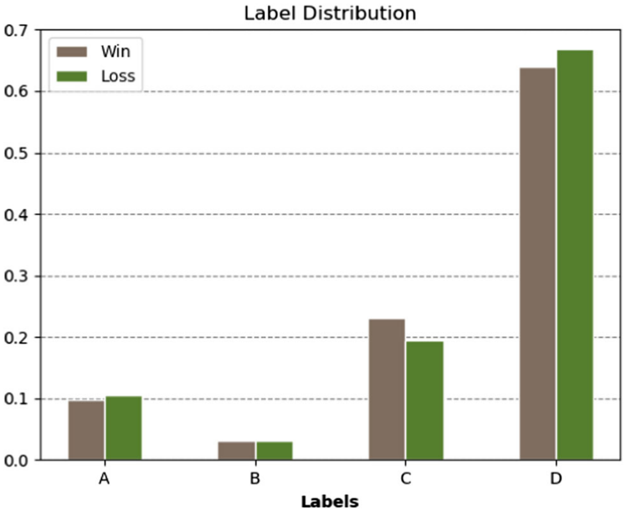


Fig. 10 Set B label distribution for won and lost games. Table 4 shows class mappings

4.2 Self-reported personality and game experience

The participants filled in two different questionnaires, which provided an opportunity to investigate the role of personality traits and game experience on game outcome and displayed affect.

Each participant filled in a 60-item HEXACO-PI-R test (HEXACO-60) [3] to assess personality along six dimensions: Honesty–Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. Participants rated 60 statements from 1 (strongly disagree) to 5 (strongly agree). We also translated the original HEXACO-60 test to Turkish, because 48 out of 58 participants' native language was Turkish and some of them were not fluent in English to answer all of the questions. Our translation consisted of three levels. One person translated the test to Turkish, then another person without seeing the original English test translated the Turkish test back to English. A third person compared the back-translated test with the original English test and marked the questions with discrepancies. This process was repeated until all the differences were acceptable, meaning that all questions in the Turkish test had the same meaning as the original English test.

After playing a game session, each player filled in a Game Experience Questionnaire (GEQ [57]). This form was also translated to Turkish with the same process explained above. The GEQ consists of four separate modules, which can be used individually. We used the in-game and social presence modules to evaluate the participant's experience during the game, and to evaluate empathy, negative feelings, and behavioural involvement with the other players, respectively. Players filled in the GEQ every time they participated in a game session. This gave MUMBAI 248 GEQ tests, which can be combined with the HEXACO-60 tests and in-game moments.

There are several tools for measuring player experience in games, and each tool looks at a slightly different set of psychological constructs [1]. Some of these tools look at constructs that cannot be observed in a 'laboratory' study, such as the socio-cultural context of gameplay. GEQ is one of the most frequently employed tools, and more importantly, has items for positive and negative affect. However, its factor structure is recently criticised to be unstable [42]. Subsequently, we do not analyse GEQ results in detail here, but provide the responses with the dataset to serve as a source of further insights.

We checked the Pearson Correlation coefficient between GEQ in-game module and HEXACO dimensions. For doing that we first took the mean of all the GEQ forms filled by each person after the Magic Maze games and calculated the seven dimensions of the in-game module. Our findings in Fig. 11 indicate a moderate association between the 'Emotionality' dimension of HEXACO with four dimensions of GEQ. While 'Positive Affect' and 'Competence' are negatively correlated, 'Negative Affect' and 'Tension' have a positive correlation. Lee and Hashton[4] describe people with high 'Emotionality' as "experience anxiety in response to life's stresses, feel a need for emotional support from others", and people with low 'Emotionality' as "feel little worry even in stressful situations, have little need to share their concerns with others, and feel emotionally detached from others". These definitions are in line with our findings because the Magic Maze game creates a stressful experience by making people worry about making mistakes, running out of time, and losing the game. Also, being a cooperative game, it creates social pressure on the players by making them responsible to each other.

4.3 Automated feature annotation

We automatically annotated face and body features from the recordings of the Magic Maze games. All players have played this game. The extracted features are provided as part of the dataset, to facilitate comparisons and to allow researchers to focus on the social aspects of the game such as group structure and personality.

4.3.1 Face and head features

We used an open-source face analysis tool, OpenFace 2.2[5] [5, 6], to locate faces in the video frames, and to extract facial landmark locations, head pose, eye gaze, and facial expressions. The processing we applied below is used for our experiments and can be optimized further.

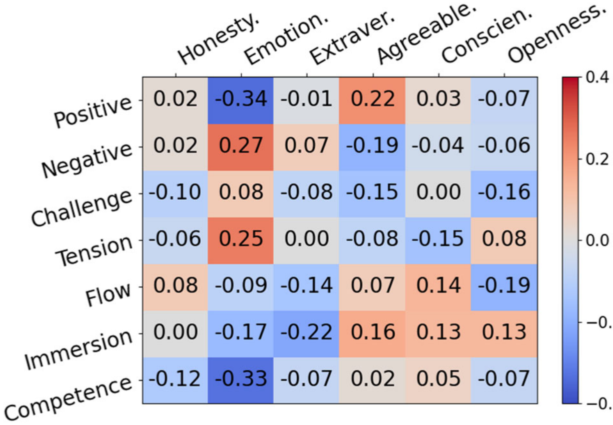


Fig. 11 Correlation between GEQ and HEXACO dimensions (honesty–humility, emotionality, extraversion, agreeableness, conscientiousness, openness to experience)

OpenFace face detections are not consistently mapped to the same person throughout the videos. Therefore, our first step is to assign all detections to the correct identity. In each video, we have two players sitting side-by-side. As their relative positions do not change, tracking the nose landmark locations is sufficient to determine whether the output of OpenFace belongs to the left or right person in view. During a play session, it sometimes happens that a player reaches for something far away or the session administrator stands up. A person then might appear in the recording of the other two players. To eliminate these unwanted faces, we check for clusters of outliers that correspond to incidental face detections. To determine whether this process correctly labels the players, we selected from each video five random frames with more than two OpenFace detections and manually checked the outputs. From these frames, 93.7% of the faces was correctly labeled.

OpenFace provides a confidence score for each detection, which we used to exclude false or problematic detections from the feature set. In preliminary experiments on the validation set, we observed that a confidence threshold of 0.5 performed the best. Confidence thresholding gives us an improved feature set, but with a cost of more missing frames. To counteract this problem we interpolated the missing detections. After interpolation, we smoothed out the noisy landmark locations with a Savitzky-Golay smoothing filter [59]. We selected this filter's window length (15) and polynomial order (3) empirically on the validation set.

The processed OpenFace outputs are then used to extract features over segments of a video. We used the same segment settings as in the annotations of Set A (Sect. 4.1). The features are divided into two categories: low-level and high-level features. We calculated low-level features together with first-

[4] https://hexaco.org/scaledescriptions.

[5] https://github.com/TadasBaltrusaitis/OpenFace.

Table 5 The facial action units used in the analysis

Action unit	Corresponding action
AU-04	Lowering of the brow
AU-05	Raising of the upper eye lid
AU-06	Raising of the cheeks
AU-07	Tightening of the eye lid
AU-09	Wrinkle in the nose
AU-15	Lowering of the lip corner
AU-20	Stretching of the lip
AU-23	Tightening of the lips
AU-26	Dropping of the jaw

and second-order derivatives and using summarizing functionals over segments for head movement, gaze movement, and facial action units.

- **Head movement:** OpenFace provides us with the location of the head in millimeters with respect to the camera. The location is given in 3D coordinates. We calculate the absolute movement of the head. The velocity and acceleration are calculated as the first- and second-order derivatives of the position with respect to time. OpenFace also provides the rotation of the head. These values can be seen as pitch, yaw, and roll. We calculate the absolute rotation to determine velocity and acceleration. For every segment, the mean and variance are calculated for the 3D coordinates of movement and pitch, yaw and roll for rotation. This provides us with 24 features for head movement: [3 mean + 3 variance velocity] + [3 mean + 3 variance acceleration] of 3D Cartesian coordinates and the same 12 summarizing functionals for the 3D rotational coordinates.
- **Gaze behaviour:** The angle of the gaze is produced by taking the average of the gaze vectors of both eyes. This creates two gaze angles, in the horizontal and vertical directions, respectively. Similar to head movement, we calculated the mean and variance of the velocity and acceleration per segment. This gives eight features for the gaze.
- **Facial Action Units:** OpenFace provides us with a subset of facial action units (AU), used to describe facial movements, as well as an intensity value between 1 and 5. The AUs that are used are shown in Table 5. The mean and variance of the intensity are calculated for each AU. Additionally, we consider AU-45 which corresponds to the closing of the eye. Simple thresholding of the smoothed intensity of AU-45 as a function of time provides us with the number of eye blinks. In total, this gives 19 features.

High-level features capture contextual details such as affective facial expressions, mouth movements, and categorical gaze direction:

- **Affective facial expressions:** Following the mappings from AUs to compound facial expressions [20], we combined AUs by averaging their intensities: One happiness, two sadness, three surprise/fear, seven fear, and three anger expressions are coded using different combinations of AUs from OpenFace. We applied mean and variance over segments to summarize these 16 expressions into 32 features.
- **Mouth movements:** To capture mouth openness and movements we used the upper and lower lip landmarks. We summarized the distance change between these landmarks using mean and variance over video segments. These features indicate the segments where players talk and where they keep their mouths open. The latter can be a signal of shock, focus, or laughter.
- **Categorical gaze direction:** Gaze angle output of OpenFace is not informative for a classification system without any additional knowledge such as the height and the relative location of a participant. We processed these features to compute one categorical gaze direction variable with three categories: looking at the board, at another player, or elsewhere. We summed up the gaze and head orientations of a player and applied thresholds based on their height and seating position. Summarizing this feature over a segment is done by majority voting.

4.3.2 Pose features

We extracted body keypoint locations with OpenPose, an open-source multi-person human body keypoint detector [11]. The detector provided us 25 two dimensional keypoints, of which we used 9 for each detected upper body. Lower body keypoints were typically not visible and OpenFace already provides head keypoints. We did not use the hand and face keypoint detectors. Different from OpenFace's single detection confidence score, each body keypoints detection in OpenPose has its own confidence score. After analysing the detections throughout the videos we chose to apply distance thresholding between neck and nose keypoints instead of applying average confidence thresholding over all the keypoints. This technique of eliminating outlier detections performed well for our dataset since our participants are seated facing towards the cameras.

Similar to the face detections, we assigned all body detections over time to the same person using the same procedure and smoothing. We chose neck points to represent the location of the detections for k-means.

We chose the same segment length and stride to summarize the OpenPose features per segment. Our low-level

OpenPose features consist of summarizing 9 keypoints and their first- and second-order derivatives using mean and variance. This results in 108 features: considering both x and y dimensions 18 values for keypoint means, 18 more for variance; 36 for mean (18) and variance (18) of first-order derivatives and 36 more for mean (18) and variance (18) of second-order derivatives. High-level OpenPose features are mean and variance summarization of left and right shoulder distance and hands-to-face distance, 6 features in total. We noticed that people tend to touch their faces and hold their chin when they are focused. Shoulder distance is calculated to inform the classifiers of the proximity of the players to the camera since they face towards the camera most of the frames. Lower shoulder distance means the player is seated normally and higher shoulder distance means the player is bending over the game board towards the camera.

5 Baseline experiments and results

In this section, we present classification results for two tasks that can be addressed using the MUMBAI dataset: expressive moment detection and emotional expression classification, respectively. We provide both qualitative and quantitative results. As such, these experiments provide insights into the challenges that can be addressed and the particular richness of the data.

In all of our experiments, we follow the same experimental protocol. We randomly divided the whole dataset into 70% training set and 30% test set based on the game sessions. Counting all the small segments in these sessions, a total of 132,352 training and 45,770 test segments are used for expressive moment detection. For emotional expression detection, we have 11,244 segments for training and 6,632 segments for testing. All the hyper-parameter tuning was done on the training set using 3-fold cross-validation. Feature set selection for each classifier was also performed during the cross-validation. Our benchmark results are presented on the test set using both classical machine learning methods such as K-Nearest Neighbours (KNN), Decision Trees (DT) [63], Random Forests (RF) [9], Extreme Learning Machines (ELM) [31] as well as Long-Short Term Memory networks (LSTM) [30] which are responsible for many state-of-the-art results in temporal analysis problems [28]. We chose F1 score, which is the harmonic mean of precision and recall, to be our evaluation metric. Precision and recall scores are also presented in our benchmark tables.

We excluded non-game event segments from our experiments since they contain irrelevant facial and bodily expressions. 0.92% of all the segments are removed by excluding non-game events.

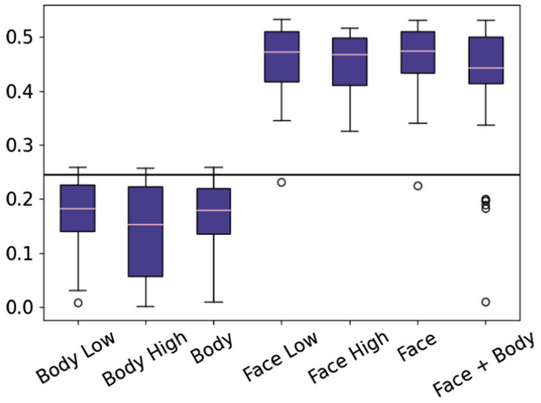


Fig. 12 Feature set comparison for binary expressive moment detection using cross-validation. Black line shows F1 score for the expressive (minority) class predictor

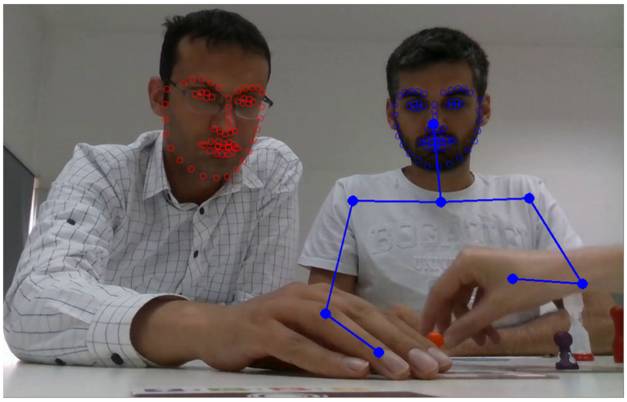


Fig. 13 OpenPose misses the player on the left and incorrectly detects the hands of the player on the right as other players' hands. OpenFace detections are correct

5.1 Expressive moment detection

We perform expressive moment detection on Set A annotations as a binary classification problem, into expressive or neutral categories. The distribution of the classes in Set A annotations is highly imbalanced. 86.05% of all the video segments belong to the neutral class. The expression classes are combined into a single class called 'expressive' for this experiment. The results on the test set are presented in Table 6.

ELM and LSTM classifiers have the best F1 scores on the test set for expressive moment detection. We perform a feature set selection as well as hyperparameter tuning via cross-validation. The 'Features' column shows the selected feature set for the specified classifier on cross-validation. The hyper-parameters chosen for each classifier in Table 6 are as follows: 31 nearest neighbours for KNN; maximum depth of 5 for DT; 50 trees and maximum depth of 50 for RF; 100 hidden units with radial basis function with 0.01 width as the activation function for ELM; 25 hidden units LSTM layer

Table 6 Expressive moment detection binary classification results using Set A annotations

Classifier	Features	F1	Precision	Recall
KNN	Face high	0.319	0.596	0.218
DT	Face + body	0.463	0.584	0.385
RF	Face	0.379	0.611	0.275
ELM	Face low	**0.538**	0.501	0.582
LSTM	Face	0.521	0.414	0.703
All expressive		0.183	0.111	1.000

Table 7 Emotional expression classification results using emotion labels of Set A annotations

Classifier	Features	F1	Precision	Recall
KNN	Face high	0.359	0.406	0.359
DT	Face high	0.380	0.390	0.392
RF	Face	**0.467**	0.463	0.482
ELM	Face high	0.431	0.456	0.415
LSTM	Face	0.453	0.442	0.490
Random		0.218	0.250	0.248

followed by 100 units dense layer with the rectified linear unit and 2 units dense layer with softmax for LSTM. We trained our LSTM for 10 epochs with a batch size of 512. We chose Adam optimizer [38] with 0.001 learning rate and applied dropout regularization. To overcome the data imbalance (86.05% neutral versus 13.95% expressive), inversely proportional class weights are used for DT, RF, and LSTM. When training ELM, we chose to upsample our minority class instead of utilizing Weighted Kernel ELM [80], which is another method for overcoming data imbalance.

We compare the feature sets in Fig. 12 by creating box plots of the results acquired by all classifiers with all hyper-parameter settings we used during cross-validation for expressive moment detection. The black line shows F1 score for the expressive (minority) class predictor. It is clear that feature sets including face features are doing better in expressive moment classification than using the body modality. Each frame in the dataset contains two persons, and OpenFace detects 1.99 faces on the average, whereas OpenPose detects bodies with a lower rate, 0.86 on the average before eliminating outliers. The problem is partly caused by the fact that the lower body is often not in view, and the hands of the players are confused with each other. Especially in games where excessive hand movements are involved, such as the Magic Maze, this is the case (Fig. 13).

Figure 14 shows the affective interaction of four players in five Magic Maze game sessions. We combined Set A annotations into three categories: the focus classes are combined into neutral and the affect classes are combined into negative and positive classes. In these representative visualizations, we see several patterns. Other than a few instances, expressive moments of a player are mostly followed or accompanied by other players' expressive moments, not unlike turn-taking behaviour. Black rectangles show emotional contagion in mostly positive expressions. We observe for example that a player initiates a chain reaction by making a funny face to hint possible moves to other players. A less common but still notable pattern is marked with green ellipses, showing negative emotional contagion. In our dataset, we observed that negative expressions are less contagious during gameplay, and do not reach all players like positive expressions usually do. A third pattern is marked with purple pentagons. These are the moments that a player makes a (silly) mistake and displays a negative expression, which results in positive expression bursts from other players, such as laughter.

In order to evaluate the degree to which emotional contagion can be processed automatically, we trained a 4-layer convolutional neural network that takes the binary expressiveness annotations as an input. The network predicted each player's annotations using the annotations of other three players of a ten-segment window centered on the current segment. That way we provide temporal information to the network. The network achieved an F1 score of 0.407. A baseline classifier predicting always expressive achieves an F1 score of 0.234, which is higher than a random predictor (0.21). This experiment illustrates that emotion contagion can be automatically assessed using facial expressions to some extent.

5.2 Emotional expression classification

We next investigated the degree with which we can automatically detect and classify emotional expressions via standard classifiers. We first performed a test using all the segments that have emotion labels from Set A annotations, which uses four classes of positive and negative affect. Table 7 shows the test results. The selected hyper-parameters which are different than the first benchmark's are the following: five neighbours for KNN; maximum depth of 15 for DT; 25 trees with maximum depths of 10 for RF; 500 hidden units for ELM; 100 LSTM units for the first layer for LSTM. In this benchmark, RF achieves a better F1 score than the rest.

We present our last benchmark on Set B annotations, which are four groups of game-related emotions (i.e. Anxious, Bored, Confused, Delighted). Classification results on the test set are shown in Table 8. The optimized parameters for the selected classifiers show some differences: three nearest neighbours for KNN; maximum depth of 15 for DT; 100 trees with maximum depth of 10 for RF; 100 hidden units with hyperbolic tangent function as the activation function for ELM; 250 hidden units LSTM layer for LSTM. Similar to the previous benchmark, RF achieves a better F1 score than other tested classifiers for this task.

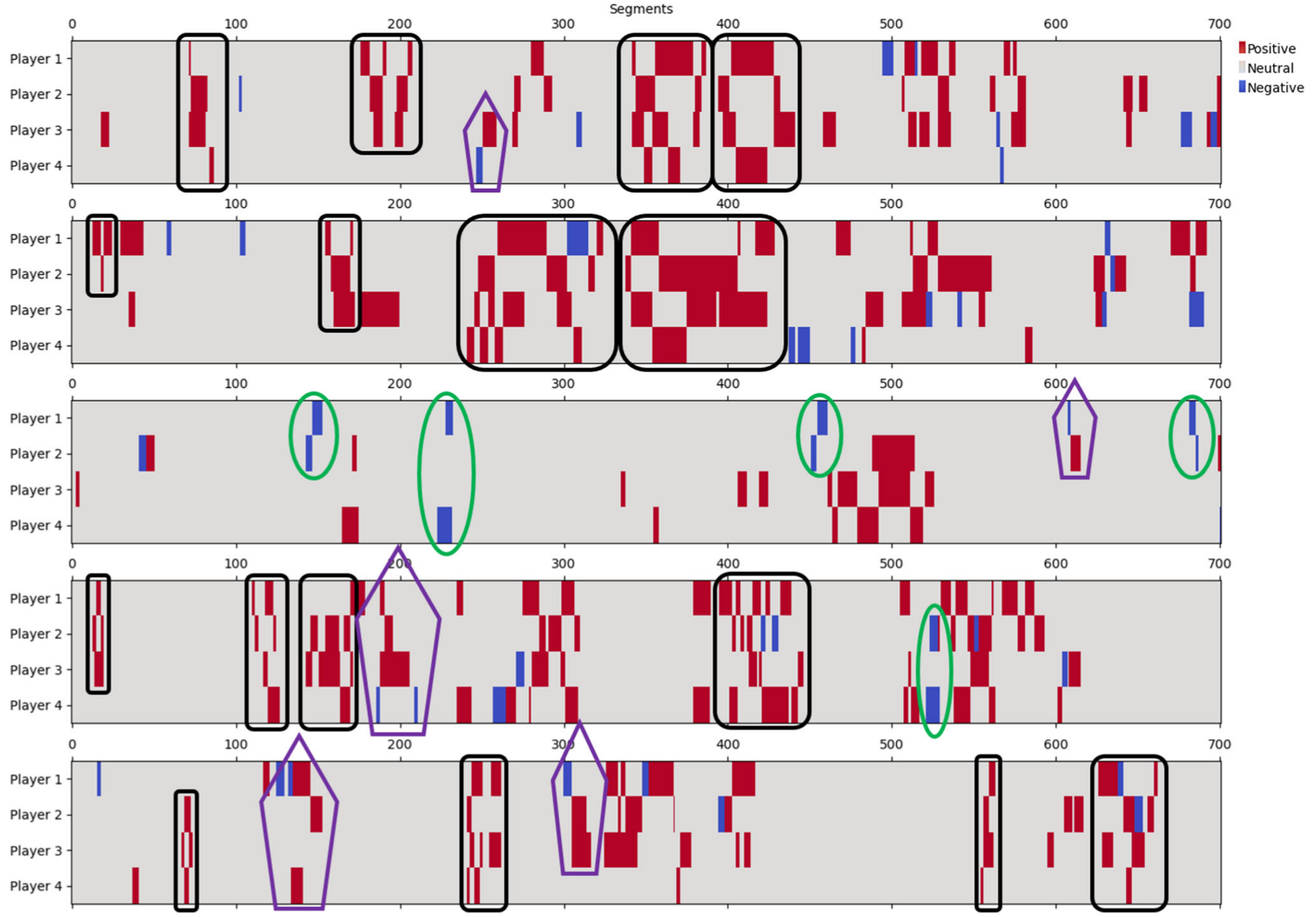
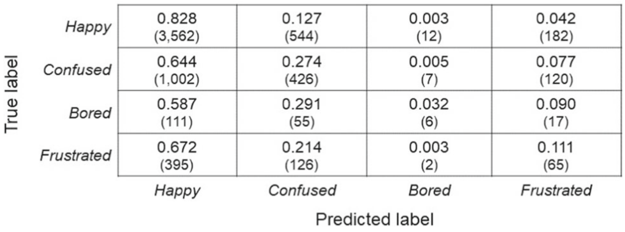


Fig. 14 Expressiveness of players throughout five Magic Maze games. Each segment is 0.5 s long. Best viewed in color

Table 8 Game emotion classification results on annotation set B, with four labels of anxious, bored, confused, and delighted

Classifier	Features	F1	Precision	Recall
KNN	Face high	0.272	0.297	0.277
DT	Face all	0.282	0.292	0.282
RF	Face low	**0.316**	0.366	0.311
ELM	Face all	0.293	0.307	0.299
LSTM	Face all	0.309	0.308	0.323
All delighted		0.213	0.186	0.250

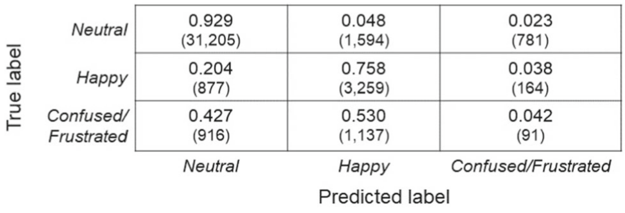


True label \ Predicted label	Happy	Confused	Bored	Frustrated
Happy	0.828 (3,562)	0.127 (544)	0.003 (12)	0.042 (182)
Confused	0.644 (1,002)	0.274 (426)	0.005 (7)	0.077 (120)
Bored	0.587 (111)	0.291 (55)	0.032 (6)	0.090 (17)
Frustrated	0.672 (395)	0.214 (126)	0.003 (2)	0.111 (65)

Fig. 15 Confusion matrix for emotional expression classification task, given as percentage (raw count of samples)

The overall emotion classification results illustrate the difficulty of detecting the often subtle emotional displays. From the confusion matrix (Fig. 15), we see that the bored class is too rare for automatic detection, and confusion and frustrated classes could be fused for further analysis. We keep these labels in the database for providing a resource for further studies. Focusing on basic emotional expressions allows one to have larger sets of samples for training models, as these expressions have been studied for a long time. Conversely, using dimensional models may hide some of these issues related to specific expressions.

As an extra experiment, we have trained our best classifier on a different set of labels, where we have added the neutral class from annotation Set A, removed the bored class (too few samples) and combined the "confused" and "frustrated" classes. The confusion matrix for the three-class classification problem is provided in Fig. 16, illustrating that while "happy" is separable from the neutral class, the detection of the confused/frustrated state automatically is still very challenging.



True label \ Predicted label	Neutral	Happy	Confused/Frustrated
Neutral	0.929 (31,205)	0.048 (1,594)	0.023 (781)
Happy	0.204 (877)	0.758 (3,259)	0.038 (164)
Confused/ Frustrated	0.427 (916)	0.530 (1,137)	0.042 (91)

Fig. 16 Confusion matrix for the revisited emotional expression classification task, given as percentage (raw count of samples). The class imbalance is evident from the raw counts

Our preliminary investigations into game enjoyment (as measured by GEQ - the game experience questionnaire) did not reveal clear links between emotional displays and game experience. [55] added a new set of annotations to the MUMBAI dataset and showed that anxiety and self-reported player experience can be predicted by analysing facial expressions during critical game events.

6 Conclusions

We have introduced the Multi-Person, Multimodal Board Game Affect and Interaction Analysis Dataset, MUMBAI, consisting of video recordings of players playing different types of board games engaging in multi-person interactions. Two sets of manual expression and emotion annotations, accompanied by self-reported personality tests of all the participants and self-reported game experience questionnaires filled after every game sessions make this dataset open to various research directions on multi-person interaction and affect analysis. We also extracted face and body features of each player that we make available online with our dataset.

The unique setup of the dataset creates some specific challenges for automatic body analysis, which can be considered to be typical for such game observation settings. The body features are not easy to capture, as the lower body is typically hidden (Fig. 12). Furthermore, the orientation of the cameras creates an unusual perspective, where whenever players interact with the board, the visible hand images are over-sized as they get closer to the cameras (Fig. 13). Facial expressions, on the other hand, are easier to spot and analyse automatically, but game-specific emotions like frustration and confusion occur rarely, and more research is needed to deal with in-the-wild conditions and severe class imbalance [16].

We presented three experiments for expression detection and emotion classification to serve as a benchmark. MUMBAI is also challenging for affective analysis because four-person interactions during a highly engaging board-game play create both strong and subtle expressions and actions. This dynamic range suggests the use of multiple levels of annotation, and we provided two different sets of annotations to cope with this issue.

Our benchmark experiments consist of one dynamic collaborative game, but the dataset includes both collaborative and competitive game settings to provide room for comparative analyses. Our analysis of personality traits and game experience indicated some expected results (such as the link between emotionality and the self-reported game affect), but also some unexpected pointers such as the negative correlation between extraversion and immersion that deserve further exploration.

References

1. Abeele VV, Spiel K, Nacke L, Johnson D, Gerling K (2020) Development and validation of the player experience inventory: a scale to measure player experiences at the level of functional and psychosocial consequences. Int J Hum Comput Stud 135:102370
2. Argyle M (2013) Bodily communication. Routledge, London
3. Ashton MC, Lee K (2009) The hexaco-60: A short measure of the major dimensions of personality. J Pers Assess 91(4):340–345
4. Aung M, Bonometti V, Drachen A, Cowling P, Kokkinakis AV, Yoder C, Wade A (2018) Predicting skill learning in a large, longitudinal moba dataset. In: 2018 IEEE conference on computational intelligence and games (CIG). IEEE, pp 1–7
5. Baltrušaitis T, Robinson P, Morency LP (2016) Openface: an open source facial behavior analysis toolkit. In: IEEE winter conference on applications of computer vision (WACV). IEEE
6. Baltrusaitis T, Zadeh A, Lim YC, Morency LP (2018) Openface 2.0: Facial behavior analysis toolkit. In: 13th IEEE international conference on automatic face & gesture recognition. IEEE, pp 59–66
7. Blom PM, Bakkes S, Spronck P (2019) Towards multi-modal stress response modelling in competitive league of legends. In: 2019 IEEE conference on games (CoG). IEEE, pp 1–4
8. Bonny JW, Castaneda LM, Swanson T (2016) Using an international gaming tournament to study individual differences in MOBA expertise and cognitive skills. In: Proceedings of the 2016 CHI conference on human factors in computing systems, pp 3473–3484
9. Breiman L (2001) Random forests. Mach Learn 45(1):5–32
10. Bull PE (2016) Posture & gesture, vol 16. Elsevier, Dordrecht
11. Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA (2019) OpenPose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 43:172–186
12. Cohen J (1960) A coefficient of agreement for nominal scales. Educ Psychol Measur 20(1):37–46

13. Corneanu C, Noroozi F, Kaminska D, Sapinski T, Escalera S, Anbarjafari G (2018) Survey on emotional body gesture recognition. IEEE Trans Affect Comput
14. Csikszentmihalyi M (1990) Flow: the psychology of optimal experience, vol 1990. Harper & Row, New York
15. Desmet P (2003) Measuring emotion: Development and application of an instrument to measure emotional responses to products. In: Funology. Springer, pp 111–123
16. Dhall A, Goecke R, Ghosh S, Gedeon T (2019) Emotiw 2019: automatic emotion, engagement and cohesion prediction tasks. ACM international conference on multimodal interaction. ACM, New York, NY, USA, pp 546–550
17. Dibeklioğlu H, Salah AA, Gevers T (2012) Are you really smiling at me? spontaneous versus posed enjoyment smiles. In: European conference on computer vision. Springer, pp 525–538
18. Doyran M, Türkmen B, Oktay EA, Halfon S, Salah AA (2019) Video and text-based affect analysis of children in play therapy. In: 2019 international conference on multimodal interaction (ICMI '19). ACM, New York, NY, USA, pp 26–34
19. Doyran M, Türkmen B, Oktay EA, Halfon S, Salah AA (2020) Multimodal affect analysis of psychodynamic play therapy. Psychother Res
20. Du S, Tao Y, Martinez AM (2014) Compound facial expressions of emotion. Proc Natl Acad Sci 111(15):E1454–E1462
21. Ekman P, Friesen WV (1978) Manual for the facial action coding system. Consulting Psychologists Press, Berkeley
22. Ekman P, Friesen WV, Hager JC (2002) Facial action coding system: the manual on CD ROM. A Human Face, Salt Lake City, pp 77–254
23. Escalante HJ, Kaya H, Salah AA, Escalera S, Güçlütürk Y, Güçlü U, Baró X, Guyon I, Jacques JCS, Madadi M, Ayache S, Viegas E, Gurpinar F, Wicaksana AS, Liem C, Van Gerven MAJ, Van Lier R (2020) Modeling, recognizing, and explaining apparent personality from videos. IEEE Tran Affect Comput. https://doi.org/10.1109/TAFFC.2020.2973984
24. Filntisis PP, Efthymiou N, Koutras P, Potamianos G, Maragos P (2019) Fusing body posture with facial expressions for joint recognition of affect in child-robot interaction. arXiv:1901.01805
25. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378
26. Frey D (1986) Recent research on selective exposure to information. In: Advances in experimental social psychology, vol 19. Elsevier, pp 41–80
27. Gardner RA (1986) The psychotherapeutic techniques of Richard A. Gardner, Creative Therapeutics Cresskill, NJ
28. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J (2017) LSTM: a search space odyssey. IEEE Trans Neural Netw Learn Syst 28(10):2222–2232
29. Güçlütürk Y, Güçlü U, Baró X, Escalante HJ, Guyon I, Escalera S, van Gerven MAJ, van Lier R (2018) Multimodal first impression analysis with deep residual networks. IEEE Trans Affect Comput 9(3):316–329
30. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
31. Huang GB, Wang DH, Lan Y (2011) Extreme learning machines: a survey. Int J Mach Learn Cybernet 2(2):107–122
32. Hung H, Chittaranjan G (2010) The idiap wolf corpus: Exploring group behaviour in a competitive role-playing game. In: Proceedings of 18th ACM international conference on multimedia (MM '10). Association for Computing Machinery, New York, NY, USA, pp 879–882
33. Hung JC, Lin ZQ, Huang CH, Lin KC (2019) The research of applying affective computing based on deep learning for eSports training. In: International conference on frontier computing. Springer, pp 122–129
34. Johansen-Berg H, Walsh V (2001) Cognitive neuroscience: who to play at poker. Curr Biol 11(7):R261–R263
35. Joo H, Simon T, Cikara M, Sheikh Y (2019) Towards social artificial intelligence: nonverbal social signal prediction in a triadic interaction. In: CVPR
36. Kaya H, Gürpınar F, Salah AA (2017) Video-based emotion recognition in the wild using deep transfer learning and score fusion. Image Vis Comput 65:66–75. https://doi.org/10.1016/j.imavis.2017.01.012
37. Khan RA, Crenn A, Meyer A, Bouakaz S (2019) A novel database of children's spontaneous facial expressions (LIRIS-CSE). Image Vis Comput 83:61–69
38. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
39. Kleinginna PR, Kleinginna AM (1981) A categorized list of emotion definitions, with suggestions for a consensual definition. Motivation Emot 5(4):345–379
40. Kleinsmith A, Bianchi-Berthouze N (2012) Affective body expression perception and recognition: a survey. IEEE Trans Affective Comput 4(1):15–33
41. Korotin A, Khromov N, Stepanov A, Lange A, Burnaev E, Somov A (2019) Towards understanding of esports athletes' potentialities: the sensing system for data collection and analysis. arXiv:1908.06403
42. Law ELC, Brühlmann F, Mekler ED (2018) Systematic review and validation of the game experience questionnaire (GEQ)-implications for citation and reporting practice. In: Proceedings of the annual symposium on computer–human interaction in play, pp 257–270
43. Lemaignan S, Edmunds CER, Senft E, Belpaeme T (2018) The PInSoRo dataset: supporting the data-driven study of child–child and child–robot social dynamics. PLoS ONE 13(10):1–19
44. Littlewort GC, Bartlett MS, Salamanca LP, Reilly J (2011) Automated measurement of children's facial expressions during problem solving tasks. In: Face and gesture 2011. IEEE, pp 30–35
45. Lucey S, Goecke R, Dhall A, Gedeon T (2012) Collecting large, richly annotated facial-expression databases from movies. IEEE Multimed 19(03):34–41
46. Mackinnon L, Bacon L, Cortellessa G, Cesta A (2013) Using emotional intelligence in training crisis managers: the pandora approach. Int J Distance Educ Technol 11:66–95. https://doi.org/10.4018/jdet.2013040104
47. Maman L, Ceccaldi E, Lehmann-Willenbrock N, Likforman-Sulem L, Chetouani M, Volpe G, Varni G (2020) Game-on: a multimodal dataset for cohesion and group analysis. IEEE Access 8:124,185–124,203
48. Martin A, Guéguen N (2012) Mimicry in social interaction: its effect on learning. Springer, Boston, pp 2275–2277
49. Matorin AI, McNamara JR (1996) Using board games in therapy with children. Int J Play Therapy 5(2):3–16
50. Mavromoustakos-Blom P, Bakkes S, Spronck P (2019) Modeling and adjusting in-game difficulty based on facial expression analysis. Entertain Comput 31(100):307
51. Mavromoustakos-Blom P, Kosta M, Spronck P, Bakkes S (2020) Player facial expression analysis in competitive hearthstone. In: Proceedings of the 2020 IEEE conference on games (CoG)
52. Mulligan K, Scherer KR (2012) Toward a working definition of emotion. Emot Rev 4(4):345–357
53. Nickerson ET, O'Laughlin KB (1980) It's fun-but will it work? The use of games as a therapeutic medium for children and adolescents. J Clin Child Psychol 9
54. Noroozi F, Kaminska D, Corneanu C, Sapinski T, Escalera S, Anbarjafari G (2018) Survey on emotional body gesture recognition. IEEE Trans Affect Comput
55. Olalere F, Doyran M, Salah AA, Poppe R (2021) Geeks and guests: Estimating player's level of experience from board game behav-

iors. In: International workshop on human behavior understanding. Springer

56. Picard RW (2000) Affective computing. MIT press, Boston
57. Poels K, de Kort Y, IJsselsteijn W (2007) D3.3 : Game Experience Questionnaire: development of a self-report measure to assess the psychological impact of digital games. Technische Universiteit Eindhoven
58. Poppe R (2017) Automatic analysis of bodily social signals. In: Burgoon JK, Magnenat-Thalmann N, Pantic M, Vinciarelli A (eds) Social signal processing. Cambridge University Press, Cambridge, pp 155–167
59. Press WH, Teukolsky SA (1990) Savitzky-golay smoothing filters. Comput Phys 4(6):669–672
60. Psaltis A, Kaza K, Stefanidis K, Thermos S, Apostolakis KC, Dimitropoulos K, Daras P (2016) Multimodal affective state recognition in serious games applications. In: IEEE international conference on imaging systems and techniques (IST). IEEE, pp 435–439
61. Rehg J, Abowd G, Rozga A, Romero M, Clements M, Sclaroff S, Essa I, Ousley O, Li Y, Kim C et al. (2013) Decoding children's social behavior. In: Proceedings of CVPR, pp 3414–3421
62. Rouast PV, Adam M, Chiong R (2019) Deep learning for human affect recognition: Insights and new developments. IEEE Trans Affective Comput
63. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21(3):660–674
64. Salah AA, Gevers T (2011) Computer analysis of human behavior. Springer, Berlin
65. Salah AA, Gevers T, Sebe N, Vinciarelli A (2010) Challenges of human behavior understanding. In: International workshop on human behavior understanding. Springer, pp 1–12
66. Salen K, Zimmerman E (2004) Rules of play: game design fundamentals. MIT press, Cambridge
67. Salter DA, Tamrakar A, Siddiquie B, Amer MR, Divakaran A, Lande B, Mehri D (2015) The tower game dataset: A multimodal dataset for analyzing social interaction predicates. In: International conference on affective computing and intelligent interaction (ACII), pp 656–662
68. Schaefer CE, Reid S (1986) Game play. Wiley, New York
69. Schimmel A, Doyran M, Baki P, Ergin K, Türkmen B, Salah AA, Bakkes S, Kaya H, Poppe R, Salah AA (2019) MP-BGAAD: multi-person board game affect analysis dataset. In: Proceedings eNTERFACE, 15th international summer workshop on multimodal interfaces, pp 1–11
70. Schirmer A, Adolphs R (2017) Emotion perception from face, voice, and touch: comparisons and convergence. Trends Cognit Sci 21(3):216–228
71. Schwarz J, Marais CC, Leyvand T, Hudson SE, Mankoff J (2014) Combining body pose, gaze, and gesture to determine intention to interact in vision-based interfaces. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 3443–3452
72. Shapiro EG, Hughes SJ, August GJ, Bloomquist ML (1993) Processing of emotional information in children with attention-deficit hyperactivity disorder. Dev Neuropsychol 9(3–4):207–224. https://doi.org/10.1080/87565649309540553
73. Shouse E (2005) Feeling, emotion, affect. M/c J 8(6):26
74. Smith P, Shah M, da Vitoria LN (2003) Determining driver visual attention with one camera. IEEE Trans Intell Transp Syst 4(4):205–218
75. Stafford T, Devlin S, Sifa R, Drachen A (2017) Exploration and skill acquisition in a major online game. In: The 39th annual meeting of the Cognitive Science Society (CogSci). York

76. Stathopoulou IO, Tsihrintzis GA (2011) Emotion recognition from body movements and gestures. In: Intelligent interactive multimedia systems and services. Springer, pp 295–303
77. Sun X, Lichtenauer J, Valstar M, Nijholt A, Pantic M (2011) A multimodal database for mimicry analysis. In: D'Mello S, Graesser A, Schuller B, Martin JC (eds) Affective Comput Intell Interact. Springer, Berlin Heidelberg, pp 367–376
78. Wulvik AS, Dybvik H, Steinert M (2020) Investigating the relationship between mental state (workload and affect) and physiology in a control room setting (ship bridge simulator). Cognit Technol Work 22(1):95–108
79. Zagal JP, Rick J, Hsi I (2006) Collaborative games: lessons learned from board games. Simul Gaming 37(1):24–40
80. Zong W, Huang GB, Chen Y (2013) Weighted extreme learning machine for imbalance learning. Neurocomputing 101:229–242

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.