This WACV 2021 Workshop paper is the Open Access version, provided by the Computer Vision Foundation. Except for this watermark, it is identical to the accepted version; the final published version of the proceedings is available on IEEE Xplore.

Geeks and guests: Estimating player's level of experience from board game behaviors

Feyisayo Olalere
Utrecht University
Dept. Information and Computing Sciences
f.e.olalere@students.uu.nl

Metehan Doyran
Utrecht University
Dept. Information and Computing Sciences
m.doyran@uu.nl

Ronald Poppe
Utrecht University
Dept. Information and Computing Sciences
r.w.poppe@uu.nl

Albert Ali Salah
[1]Utrecht University
Dept. Information and Computing Sciences
[2]Boğaziçi Univ., Dept. Computer Engineering
a.a.salah@uu.nl

Abstract

Board games have become promising tools for observing and studying social behaviors in multi-person settings. While traditional methods such as self-report questionnaires are used to analyze game-induced behaviors, there is a growing need to automate such analyses. In this paper, we focus on estimating the levels of board game experience by analyzing a player's confidence and anxiety from visual cues. We use a board game setting to induce relevant interactions, and investigate facial expressions during critical game events. For our analysis, we annotated the critical game events in a multiplayer cooperative board game, using the publicly available MUMBAI board game corpus. Using off-the-shelf tools, we encoded facial behavior in dyadic interactions and built classifiers to predict each player's level of experience. Our results show that considering the experience level of both parties involved in the interaction simultaneously improves the prediction results.

1. Introduction

Board games have become promising tools in stimulating and studying social and psychological behaviors in people (e.g., [8, 13]). Game rules can easily be adapted to evoke or suppress specific individual or interactive behaviors such as those relating to cooperation and competition, frustration or enjoyment, and winner or losing. Owing to this flexibility and their wide popularity, board games are being used to better study various aspects of human behavior.

When studying the behaviors displayed during board gameplay, researchers have predominantly used self-reported questionnaires. Much research has gone into creating and validating questionnaires for a wide array of social and cognitive constructs (e.g., [1, 18, 23]). Once a specific set of questions has been validated to measure the target variable, a questionnaire can easily be employed in a broad range of contexts [16]. However, self-report questionnaires have well-known disadvantages. First, filling in questionnaires takes time, which increases the burden of participants to enroll in scientific studies. Second, self-reporting is open to framing effects and can create biases such as the "response bias", where respondents try to present themselves in a more favorable light [16]. One of the solutions to these drawbacks is to use automated analysis methods for these behaviors. Automatic evaluation of behaviors has been shown to reveal preferences in a more objective way than using questionnaires [9]. Measuring behavior during gameplay can be a natural way of prompting rich cues to study psychological constructs [22].

In this paper, we estimate a person's confidence and anxiety levels from visual cues, in a board game setting. The player's level of board game experience is used as a proxy. We use a publicly available corpus of board game interactions, which contains four-player interactions in collaborative games. We enrich the dataset with a set of new annotations about critical game events.

We analyze how the display of specific facial expressions of players during gameplay interactions are affected by their personalities and their game experience. Based on these analyses, we build a classifier that uses the facial expressions of players during critical game events to predict the level of experience in the board game of the players. We make use of not just a single player's facial features for pre-

diction, but fuse facial features from dyadic interactions in a multi-task learning setting. We make all our annotations publicly available as an extension to the existing MUMBAI dataset. The following are our contributions:

1. We predict the self-reported level of experience of people in multi-person interaction settings using facial expression dynamics of player interactions during critical game events.

2. We explore the possible connections between critical game events and players' facial expressions with the self-reported game experience and personality questionnaires.

In the next section, we discuss the related work in multimodal behavioral analysis and focus on game-based interactions. In Section 3, we discuss the MUMBAI dataset used in this study, as well as the additional annotations to conduct our experiment. In Section 4, we describe the variables used for the correlation analysis and the methods used in the classifier training. We present and discuss the results of our analysis in Section 5. Finally, we conclude the paper with a discussion and implications for further analysis.

2. Related work

To better study human behaviors or skills, it is common to create settings that will elicit these behaviors and signals, and to collect multimodal observations from these settings. For dyadic and group interactions, meeting scenarios and survival tasks that prompt discussions are frequently studied [6, 20]. There is comparatively little research on the automated analysis of behavior in games and during playful interactions [19]. In this section we review some of the related work in group interactions and game settings.

Game settings are particularly good in eliciting some emotional expressions. Players often display happiness and excitement (i.e. positive valenced emotions) in fast paced and exuberant games. In more contemplative games, we see expressions of concentration, anxiety, relief, curiosity, and surprise. Depending on the game state, boredom, elation and frustration can be observed [11].

Giannakakis et al. described a number of facial cues related to anxiety and stress, and proposed an analysis pipeline to assess these emotional states via face analysis [12]. In their experimental analysis, they induce stress via videos or emotional recall tasks. They have reported reduced rhythmicity of lip movements and and increased mouth movements associated with increased levels of stress/anxiety. Their study is particularly interesting in the usage of facial cues, as in naturalistic settings and daily life scenarios, wearable sensors are more informative than facial analysis for stress assessment [5], but present a more intrusive data collection setup. For group settings, affective cues that go beyond the basic emotional expressions, such as anxiety, are particularly important [10]. In psychology, such cues are studied in the context of parent-child interactions, as well as out-of-group interactions.

A recent study to investigate multi-party interactions during conversations amongst a group of people introduced the Teams corpus [14]. This dataset was curated from teams of 3-4 people playing a cooperative board game called the Forbidden Island. There was a total of 63 teams and 47 hours of recordings. The dataset contains video, audio, transcripts, and questionnaire data, including self-report questionnaires on personality, cognitive style, and collective orientation. The post-game questionnaire contains questions to measure the perception of team processes such as cohesion, satisfaction, and potency/efficacy. The authors investigated several research questions using this dataset, including linguistic entrainment, which is a phenomenon where speakers in conversation start to use similar linguistic features during the conversation [4].

In a more recent study, the GAME-ON Dataset was created to study group cohesion [15]. A total of 151 participants participated in groups of three to play an escape game similar to Cluedo. Data on the participants' emotional state and their perceptions of leadership, warmth, and competence of their other group members were collected using the Group Environment Questionnaire [7]. The dataset consists of audio-visual recordings, manual annotation of participant perception of cohesion over time, and the responses to the questionnaires. The study leverages an existing theoretical framework for studying group cohesion in different dimensions [21], but focuses on task cohesion and social cohesion amongst friends.

Another recent game corpus is the MUMBAI dataset [11], which focuses on collaborative board game plays to observe non-verbal signals for affective displays and interactions. Four players are simultaneously recorded, and a number of pre-game and post-game questionnaires complement the recordings, including a self-reported personality questionnaire [1]. Since this corpus is publicly available and contains video recordings (as opposed to for instance the Teams corpus, which is audio-only), we use it in the present work. The dataset also contains affect annotations of negative and positive facial expressions for in-game events. Moreover, the players involved in the data collection have indicated their level of experience in playing board games, which serves as an interesting proxy in assessing player confidence and in-game anxiety. The ground truth on personality allows us to study how various profiles react when a critical game event occurs. We will also test whether the facial expressions of the players as a response to game events and other player interactions can be used to predict certain properties about each player, such as their level of experience in the board game.

Figure 1. Participants playing the Magic Maze game, from the three different camera views used in the recording setup.

3. The experimental setup

In this section, we discuss the experimental setup used in our study, including the dataset we have used and the annotations produced to address our research questions.

The MUMBAI dataset was created to allow for the automated analysis of multi-modal behaviors in a multi-player game. The dataset consists of video recordings of 62 game sessions along with manual annotation of affect, self-reported questionnaires on personality and game experience, automatically extracted facial features and body landmarks, and the game outcome (win or loss). Each game session consists of a group of four people playing a board game. The used games are Magic Maze, Kingdomino, Qwixx, Pandemic, King of Tokyo, and The Mind. These games are either cooperative (co-op) or competitive in nature, but most of the played games in the corpus are from cooperative games, which arguably create more interactions between the players, who need to pay attention to each other's actions and to coordinate their behavior with each other to win the game.

3.1. Player affect annotations

The MUMBAI dataset provides two sets of manually annotated player affect. The first set of annotations used in this study was done by two naive annotators. They obtained an inter-reliability score (i.e. Cohen's Kappa) of 0.735 for binary neutral class vs. the rest and a score of 0.669 for all the categories [11]. This set of annotations contains the expressive moments for each player. Seven labels (Positive, Small positive, Neutral, Small Negative, No label, Focus, Small Focus, Non-Game event) were used to indicate how much a player's facial expression differs from a neutral facial expression. The labels used represent positive expressions (e.g, laughter), negative expressions (e.g, frowning), neutral expression, focus (not negative or positive, but represents expressions like concentration depicted by narrowed eyes), and non-game events. The dataset is collected within a very natural game playing setting, with unobtrusive sensing, and includes naturally occurring non-game events such as players picking up a call during gameplay (see Figure 1).

The second set of annotations provided in the MUMBAI dataset focus on game-specific facial expressions. The expressive moment annotations from the first set were classified into four categories. However, this was not used in this study.

3.2. Questionnaires

Two questionnaires were filled in by the participants. The first questionnaire was given to capture the participant's personality traits, while the second questionnaire captured the player's in-game experience.

The HEXACO-60-PI-R (HEXACO-60) personality test was used for the personality-related questionnaire. This is a 60 question questionnaire that assesses personality based on six dimensions: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. Participants answered each question with a 1-5 scale system, where 1 means strongly disagree and 5 means strongly agree. This questionnaire is somewhat more detailed than the Big-Five questionnaire, and the additional Honesty-Humility dimension is relevant for game settings.

To capture the player's in-game experience, each player filled a Game Experience Questionnaire (GEQ) [18] after every game session. The in-game and social presence module of the GEQ was used to measure and evaluate both the participants' experience during the game, as well as their empathy, negative feelings, and behavioral involvement with other players in the game session.

3.3. Magic Maze game annotation

We focus on one of the board games in the MUMBAI dataset, namely the Magic Maze game. The Magic Maze games in the corpus contained 39 recorded video sessions. Each session had a total of four participants and across all sessions, there was a total of 57 distinct participants. Age ranged from 15 to 43 (see Figure 2 (left)), 31% of the participants were female and 69% were male.

All participants reported their level of experience on a scale of 0-4, where 0 is not experienced and 4 is very experienced. In this paper, we treat the experience prediction task as a binary classification problem. To this end, we consider players with a reported experience score of 0–1 as inexperienced and those with scores 2–4 as experienced. See Figure 2 (right) for the distribution.

Magic Maze is a cooperative game where the players jointly move four pawns to explore a maze and steal trea-

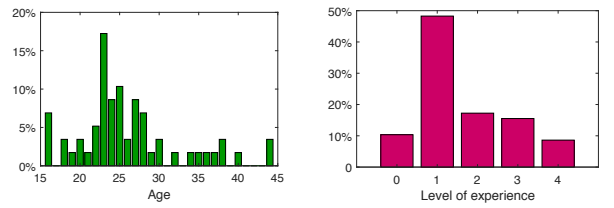

Figure 2. Distributions of participants' age and game experience.

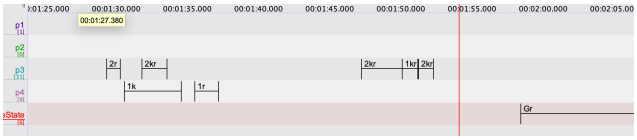

Figure 3. Critical game event annotation in ELAN software

sures in the maze. This game is time-bound, as the players have to complete the game before a hourglass runs out. While each player has a certain direction to move in and some special game functions based on their cards, players do not take turns and interact with game elements simultaneously. Moving a pawn can be done by any player at any point in time. Players are not allowed to communicate verbally except during certain game events. They can communicate non-verbally throughout the game by placing a red cone in front of the player they expect to make a move, or by other means such as staring or gesturing. The players win the game jointly if they successfully steal all treasures and get all the pawns out of the maze before the hourglass runs out. Otherwise they lose the game.

3.4. Annotation of critical game events

Placing the red cone in front of a player is the main source of communication in this game, hence we consider it a critical game event. We focus on red cone usage as a critical game event because we expect that facial expressions exhibited during interactions and outside interactions should vary. The ELAN software was used to annotate the video recordings of the game sessions (see Figure 3). For each player, we annotate the moment when they place a red cone in front of another player. For each game session, we annotated three game events: when a pawn is placed on an hourglass tile, when a verbal interaction is initiated (which happens when a green pawn opens a new tile of the maze) and the beginning of the second phase of the game. The second phase of the game starts after all the pawns have stolen their treasures and now have to make their escape from the maze. At this point, certain special functions in the game can no longer be used to make the escape harder, and this phase requires closer cooperation between the players.

We annotated different ways of using the red cone as different game events: placing a red cone in front of a player, knocking the red cone down on the surface in front of a player and knocking the table with hands in front of the player (see Table 1). While this event is specific to the Magic Maze game, we reason that similar game-specific interaction moments exist in most board games. Automatic processing of these moments via computer vision approaches will necessarily require some customization for each game, but the main processing tools, such as gaze detection, body skeleton detection and hand tracking, facial expression detection, are common to each scenario.

Label	Description
#r	The red cone was put in front of this player
#kr	A red cone was knocked down on the surface in front of this player
#k	A player knocks with their hand in front of this player
Hg	A pawn was placed on an hourglass tile hence the physical hourglass was reset and the players can talk
Gr	A green pawn opens up a new section of the maze and the players can now speak
S2	All the pawns have stolen their treasures and are now about to make their escape

Table 1. Game event labels. # is replaced with the player's index in the game session for each player 1-4.

4. Methodology

In this section, we discuss the different methods applied in this study (see figure 4). First, we explain how we test for correlations between the critical game events and the dimensions of the GEQ and HEXACO-60 questionnaires. Next, we explain the classification approach used to predict the level of experience of the players.

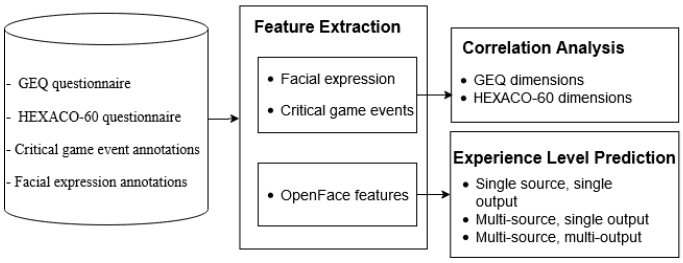

Figure 4. Methodology Pipeline.

4.1. Correlation Analysis

To see whether the facial expressions exhibited by players during critical game events correlates with their self-reported personalities and game experience, we combine our game event annotations with the facial expression annotations provided in the MUMBAI dataset. By combining the annotations, we are able to extract the counts of different facial expressions annotated for the players during

the Magic Maze games. The counts, average and standard deviation of annotated facial expression per player are the variables used to test for correlations against the game experience and HEXACO-60 personality questionnaires.

4.1.1 GEQ dimensions

The GEQ dimensions include Competence, Immersion, Flow, Tension, Challenge, Negative Affect, Positive Affect, Empathy, Negative feelings, and Involvement dimensions. We get the counts of annotated facial expression per player in each game session. We normalized counts by the session length, as the time of game play varied across sessions. Below are the descriptions for counts extracted and used for the correlation test against the GEQ dimensions:

1. **Count of critical game events:** We obtain the number of times the players initiated critical game events in each game session.

2. **Count of positive facial expressions:** By combining the game annotations (Table 1) with the existing facial expression annotations in the MUMBAI dataset, we can extract four features based on the positive facial expressions expressed by the players.

   (a) During critical game events: This refers to the count of positive facial expressions annotated within the period a critical game event was carried out by a player. Since expressions may not occur simultaneously with the event, a 3-second buffer was allowed for the expression, counted from the end of the critical game event.

   (b) Outside critical game events: Similar to the first item, but obtained outside critical game events.

   (c) In the first part of the game: We calculate the count of the positive facial expressions annotated for each player in the first part of the game.

   (d) In the second part of the game: The second part of the Magic Maze game is where we expect more interaction and coordination. We obtain the count of the positive facial expressions that occur in the second part of the game.

3. **Count of negative facial expressions:** Similar to the counts of positive facial expressions, we extract four features based on the negative facial expressions displayed during and outside critical game events.

4. **Count of focus facial expressions:** In the original MUMBAI study [11], the "focus expression" annotation depicts when a player pays full attention to the board game. Moments where participants had narrowed eyes or lower blink rate were labeled as focus expression. Using these labels, we extract the same four features as done with positive and negative facial expressions.

4.1.2 HEXACO-60 dimensions

The HEXACO-60 dimensions includes Emotionality, Honesty-Humility, Extraversion, Agreeableness, Conscientiousness, and Openness to Experience. Unlike the GEQ questionnaire, participants only fill in the HEXACO-60 personality questionnaire once, even though they are allowed to participate in multiple game sessions. We get similar facial expression counts as explained above, but now we take the average counts per players over all the game sessions they participated in.

First we get the average critical game event performed by each player and this serves as one of the variable we tested correlation for against the HEXACO-60 dimensions. Next we get the average counts based on the three possible facial expressions annotated for each player (positive, negative and focus). Alongside the average counts of facial expressions, we also calculate the standard deviation so as to capture the subtleties that might not get reflected in the average count. The following are the counts for each possible facial expression annotated:

1. **Facial expressions during critical game events:** We take the standard deviation and the average count of each facial expression annotated for each player during critical game events across all their game sessions.

2. **Facial expressions outside critical game events:** This consists of the standard deviation and average count of each facial expression that occurred for each player outside the critical game event regions.

3. **Facial expressions during first part of the game:** We take the standard deviation and average count of each facial expression displayed by each player during the first part of the game across all their game sessions.

4. **Facial expressions during second part of the game:** Similar to the previous item, for the second part of the game.

4.2. Decision tree classifier

In this section, we discuss the various methods applied in training our classifiers to predict the level of experience of the interactor and interactee during a critical game event.

We use the OpenFace 2.0 library [2] to obtain facial features from each player, which are summarized into 50 frame-long segments (1.667 seconds) using first and second order derivatives. An 50 frame window was preferred, as the shortest expressions in the dataset were about 50 frames

long, and the best inter-annotator agreement was achieved using the 50 frame window. These windows are shifted by 16 frames to capture overlapping regions. We used gaze direction, gaze angle, 2D facial landmark locations, and facial action units to summarize each player's facial features and used them as the input to our classifiers. We tested three different approaches in training the classifier:

1. **Single source, single output:** In this approach, we train two individual classifiers, one for the interactor, and one for the interactee. In both cases, the experience level of the person is predicted as the output. For each interactor in a game session, we get all the OpenFace features that were annotated within the same period that each game event annotation occurred. We also give an offset of three seconds to capture any delayed expressions related to the game event.

2. **Multi source, single output:** In this approach, we assume that the interacted party is providing relevant cues as well for the prediction task. We train two classifiers, but each receives input from both the interactor and the interactee. One predicts the experience level of the interactor, and the other, that of the interactee.

3. **Multi source, multi output:** We train just one classifier with this method. Similar to the multi source, single output method, we use the OpenFace features per segment for each interactor-interactee pair in each critical game moment for predictions. However, we predict the level of experience for the interactor and interactee together, subsequently, this is a multi-task classifier.

Since critical game events take longer than 50 frames (which is the feature extraction interval), all our classifiers predict multiple times for each region. We also evaluate how decision fusion performs when we combine the predictions at the 50 frames segment level up to critical game event region level. We apply majority voting for each critical game event region and select the most frequently predicted level of experience for the interactor and the interactee. After majority voting, we get as many data points as the number of critical game events that occurred in each game session. We present and discuss the results of this decision fusion in the next section (also see Table 4).

The classifier used for our predictions is a decision tree [3]. We did not run extensive experiments with many classifier types in order not to positively bias our results. We preferred the decision tree, as it can provide further insights into feature relevance after training.

We used an optimized version of the CART algorithm, provided in the scikit-learn library [17]. We used 5-fold cross-validation in splitting the annotation files into training and testing sets. We used the Gini impurity function to measure the quality of splits and the best strategy to choose what split to keep at each nodes.

As discussed in Section 4.1.2, we first make predictions using the OpenFace features of the interactor and interactee extracted per 50 frames segment within each critical game event moment. After this, we combine the predictions per segment for each critical game event moment by performing majority voting.

5. Baseline experiments and result

In this section, we present the results of the experiments. First, we summarize the most important correlations between the critical game events and self-reported game experience and personality. Second, we discuss the classification performance for the task of predicting the game player's level of experience using facial interactions at critical game events.

5.1. Correlation analysis

We start our analysis by checking if a participant's involvement in a critical game event correlates with their reported personality and game experience. We calculated the Spearman rank-order correlation for each extracted feature with each dimension of both questionnaires. Since we did not find a high positive or negative correlation coefficient in both questionnaires, we discuss significant results (p-value $\leq$0.05) with coefficient values greater than 0.3.

5.1.1 GEQ correlation results

Instead of reproducing the entire 25×10 correlation table for game experience questionnaire (GEQ) results, we summarize the most important findings. We observe a significant negative correlation (-0.447) between the number of times participants perform a critical game event and their immersion in the game. The immersion dimension of the GEQ seeks to measure how engaged players felt during the game, including experiences such as losing connection with the outside world and being imaginative during gameplay [18]. Our initial hypothesis was immersion would increase with player experience, because performing game events require some level of concentration. However, the relationship between these events and immersion is not so trivial. There are cases where players perform a critical game event that is followed by an inadequate response from another player. This can lead to reduced game immersion. For example, in the Magic Maze game, when a player uses the red cone to get another player's attention, the first player typically stops making further moves until the second player makes a move in the game.

The next significant negative correlation (-0.426) that occurs is between the number of negative facial expressions displayed by players outside of critical game moments

and the competence dimension. This dimension measures how good, skillful, or successful the players felt during the game [18]. We observe that the negative correlation is slightly stronger when we consider the negative facial expressions displayed by the participants during the second part of the game (-0.378) compared to the first part of the game (-0.360). Generally, we see that most of the features that are extracted when the players show a negative facial expression (see Section 4.1.1) have a negative correlation with the competence dimension of GEQ. This could mean that the more the negative facial expressions displayed by the players within and outside critical game events, the less competent they feel about the game. Lastly, we see a negative correlation between the number of negative facial expressions displayed by players outside of critical game events in the second part of the game and the positive affect dimension (-0.306). This dimension measures fun and enjoyment during gameplay [18]. The correlation could mean that the more fun the players have during the second part of the game, the less negative facial expressions they display.

The only positive correlation that occurs with the GEQ is between the number of positive facial expressions displayed by players outside of critical game events and the tension dimension (0.340). In the design of the game experience questionnaire, they note that the feeling of tension usually described by players is not the same as a negative affect [18]. This is easily noticed in this game, as it is a cooperative game and not a competitive game, so tension tends to arise when the players notice they are running out of time. While they are trying to figure out what to do, we see that most players are smiling or grinning. This relates to the nature of the game.

5.1.2 HEXACO-60 correlation result

In the HEXACO-60 personality questionnaire, we found a positive correlation between the average number of positive facial expressions displayed by players outside of critical game events and the facial expressiveness dimension (see Table 2 for a partial overview). We also see a positive correlation between the facial expressiveness dimension and the standard deviation of negative facial expression count displayed by the players during critical game events and in the second part of the game. The facial expressiveness dimension captures facial expressions such as fearfulness, anxiety, dependence, and sentimentality [1]. This could mean that the more pronounced the facial expression displayed by a player, the more expressive that player is.

The next important correlation is a negative correlation between the standard deviation of positive facial expression count displayed by players outside of a critical game event and the conscientiousness dimension. This correlation also holds for the standard deviation of focused moment count displayed by players both during and outside a critical game event. The conscientiousness dimension measures organization, diligence, perfectionism, and prudence of a player [1]. The correlation could indicate that the more variation of positive and focus count displayed by a player during and outside a critical game event, the less conscientious the player is, or the other way around.

We also observe a negative correlation between the average count of positive facial expression displayed by players during the second part of the game and the extraversion dimension. The extraversion dimension measures social self-esteem, social boldness, sociability, and liveliness of the players [1]. This seems somewhat counter-intuitive as we would expect that the more extrovert a person, the more positive expressions they would display. In contrast, the negative correlation could be an indicator of how tense the second part of the game is and we would expect players to display less positive expressions when they are tense. There also exists a negative correlation between the average count of negative facial expression displayed in the second part of the game and the openness to experience dimension. This dimension measures a player's aesthetic appreciation, inquisitiveness, creativity, and unconventionality [1].

Lastly, in the MUMBAI experiment, participants were asked to fill in their level of expertise when it comes to playing board games in general. We observe a positive correlation between the average number of time players were involved in a critical game event and their self-reported level of experience in board games. This could mean that the more experienced players tend to be more involved in the critical game events.

As we see that the facial expressions displayed during critical game events correlate with some of the dimensions in both questionnaires, we expect that the facial cues extracted at these points should carry useful signals about the player's game behavior. Based on this hypothesis, we proceed to predict the level of experience of players at each point where a critical game event occurs.

5.2. Player experience level prediction

We created two baselines to compare against the decision tree classifier's predictions. The first baseline is a random classifier. The second baseline is the majority baseline. We take the most frequent label in the training set and set it as the predicted label in the test set. This was done to account for class imbalance within our dataset.

We present major F1 scores, since our classes are not balanced, in Table 3. The three approaches using facial expressions perform much better than both of the baselines. From this table, we see that the Single Source, Single Output method predicts the interactor's experience better than the other approaches, with an F1 score of 0.660. We also notice that the performance difference compared to the other

	Self-reported Experience	Emotion-ality	Honesty-Humility	Extra-version	Agree-ableness	Conscien-tiousness	Openness to experience
Avg red cone count per player (pp)	0.402	-0.198	-0.180	0.237	-0.037	0.027	-0.253
Std positive emotion count pp in RZ	0.101	0.123	-0.020	0.027	-0.001	-0.366	-0.202
Std positive emotion count pp in NRZ	-0.025	0.112	-0.194	-0.062	0.078	**-0.458**	-0.238
Std negative emotion count pp in RZ	0.288	**0.440**	0.027	-0.031	0.0840	-0.110	0.147
Std negative emotion count pp in the NRZ	0.254	0.368	0.125	-0.327	0.131	-0.189	0.090
Std focus emotion count pp in RZ	-0.014	0.048	-0.158	0.052	0.333	**-0.408**	-0.230
Std focus emotion count pp in the NRZ	-0.224	0.166	-0.107	-0.022	0.054	**-0.414**	-0.047
Avg positive emotion in first part of the game	-0.029	0.313	-0.254	-0.205	0.118	0.059	-0.253
Avg positive emotion in second part of the game	0.153	0.381	-0.118	-0.398	0.256	-0.019	-0.265
Avg negative emotion in first part of the game	0.130	0.169	0.035	-0.228	-0.101	0.048	-0.189
Avg negative emotion in second part of the game	0.234	-0.052	-0.336	-0.238	-0.13	-0.278	**-0.408**
Std negative emotion in first part of the game	0.263	**0.434**	0.012	-0.160	0.044	-0.155	0.100
Std negative emotion in second part of the game	0.243	0.261	-0.168	-0.141	0.116	-0.350	-0.062

Table 2. Correlations between game events and self-reported level of experience and the six HEXACO-60 dimensions. RZ: red cone zone, NRZ: non-red cone zone.

two methods is smaller than 0.03. Only the facial cues of the interactor seem to be sufficient to predict the level of experience. However, this is not the case with predicting the interactee's level of experience. The classifier performs best at predicting the interactees' level of experience when the Multi Source, Multi Output method is applied. With this result, we see that by feeding the classifier facial cues of the interactor, the classifier is slightly better able to predict the interactee's level of experience. Although the improvement is small, we speculate that the initiator of the critical game event (interactor) can judge the interactee's level of experience and show facial expressions relative to their experience difference.

Method	Interactor	Interactee
Single Source Single Output	**0.660**	0.582
Multi Source Single Output	0.647	0.583
Multi Source Multi Output	0.637	**0.585**
Random baseline	0.477	0.493
Majority baseline	0.385	0.383

Table 3. Level of experience prediction performance (F1 scores) for each segment of critical game events.

Method	Interactor	Interactee
Single Source, Single Output	**0.665**	0.588
Multi Source, Single Output	0.658	0.592
Multi Source, Multi Output	0.623	**0.596**
Random baseline	0.480	0.489
Majority baseline	0.383	0.394

Table 4. Decision fusion performance (F1 scores) at each critical game event region.

Table 4 shows the performance of applying majority voting to the small segment (50 frames) predictions in each critical game event region. We observe similar results with Table 3. The fusion classifier performs best at predicting the level of experience for the interactor when the Single Source, Single Output method is used. It also shows that the classifier predicts the interactee's level of experience best when the Multi Source, Multi Output method is used.

6. Conclusions

We have explored the problem of predicting a player's level of experience during multiplayer board games using facial expressions of all players. We observed some statistically significant correlations between critical game events and game experience, as well as personality, as measured by self-report questionnaires. Additionally, we observed a positive correlation with the self-reported level of experience. We reported a classification experiment to predict the player's level of experience using the players' facial expressions. We have demonstrated that a straightforward decision tree classifier can predict the level of experience of both the interactor and interactee using facial cues. Including information about the interactor in the classifier resulted in better predictions of the interactee's level of experience.

The methods applied in this study can be extended to other board games and would be useful in extracting group and individual behavior in multi-person interactions. In addition to facial features, body movement features can be added to the analysis. Since we are using a limited dataset, we did not attempt to improve classification rates via more elaborate classifiers. However, we recognize that the features we used have limited power when processing temporal sequences. Applying temporal classifiers may yield better prediction accuracy, and potentially, better insights.

References

[1] Michael C Ashton and Kibeom Lee. The HEXACO-60: A short measure of the major dimensions of personality. *Journal of personality assessment*, 91(4):340–345, 2009.

[2] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. In *13th IEEE Int. Conf. on Automatic Face & Gesture Recognition*, pages 59–66, 2018.

[3] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[4] Susan E Brennan. Lexical entrainment in spontaneous dialog. *Proceedings of ISSD*, 96:41–44, 1996.

[5] Yekta Said Can, Bert Arnrich, and Cem Ersoy. Stress detection in daily life scenarios using smart phones and wearable sensors: A survey. *Journal of biomedical informatics*, 92:103139, 2019.

[6] Jean Carletta. Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.

[7] Albert V Carron, W Neil Widmeyer, and Lawrence R Brawley. The development of an instrument to assess cohesion in sport teams: The group environment questionnaire. *Journal of Sport and Exercise psychology*, 7(3):244–266, 1985.

[8] Ching Yue Chow, Reisya Rizki Riantiningtyas, Mie Bojer Kanstrup, Maria Papavasileiou, Gie Djin Liem, and Annemarie Olsen. Can games change children's eating behaviour? A review of gamification and serious games. *Food Quality and Preference*, 80:103823, 2020.

[9] Hamdi Dibeklioglu and Theo Gevers. Automatic estimation of taste liking through facial expression dynamics. *IEEE Transactions on Affective Computing*, 11(1):151–163, 2020.

[10] Sidney D'Mello and Rafael A Calvo. Beyond the basic emotions: What should affective computing compute? In *CHI'13 extended abstracts on human factors in computing systems*, pages 2287–2294. 2013.

[11] Metehan Doyran, Arjan Schimmel, Pınar Baki, Kübra Ergin, Batıkan Türkmen, Almıla Akdağ Salah, Sander Bakkes, Heysem Kaya Kaya, Ronald Poppe, and Albert Ali Salah. Mumbai: Multi-person, multimodal board game affect and interaction analysis dataset. *Journal on Multimodal User Interfaces*, to appear.

[12] G Giannakakis, Matthew Pediaditis, Dimitris Manousos, Eleni Kazantzaki, Franco Chiarugi, Panagiotis G Simos, Kostas Marias, and Manolis Tsiknakis. Stress and anxiety detection using facial cues from videos. *Biomedical Signal Processing and Control*, 31:89–101, 2017.

[13] Martijn Jansen and Tilde Bekker. Swinxsbee: A shared interactive play object to stimulate children's social play behaviour and physical exercise. In *Proc. INTETAIN*, pages 90–101, 2009.

[14] Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. The teams corpus and entrainment in multi-party spoken dialogues. In *Proc. EMNLP*, pages 1421–1431, 2016.

[15] Lucien Maman, Eleonora Ceccaldi, Nale Lehmann-Willenbrock, Laurence Likforman-Sulem, Mohamed Chetouani, Gualtiero Volpe, and Giovanna Varni. Game-on: A multimodal dataset for cohesion and group analysis. *IEEE Access*, 8:124185–124203, 2020.

[16] Jennifer Dodorico McDonald. Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire*, 1(1):1–19, 2008.

[17] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] K Poels, YAW de Kort, and WA Ijsselsteijn. D3.3: Game experience questionnaire, 2007.

[19] Albert Ali Salah, Ben AM Schouten, Stefan Göbel, and Bert Arnrich. Playful interactions and serious games. *Journal of Ambient Intelligence and Smart Environments*, 6(3):259–262, 2014.

[20] Dairazalia Sanchez-Cortes, Oya Aran, and Daniel Gatica-Perez. An audio visual corpus for emergent leader analysis. In *Workshop on multimodal corpora for machine learning: taking stock and road mapping the future*, 2011.

[21] Jamie B Severt and Armando X Estrada. On the function and structure of group cohesion. In *Team cohesion: Advances in psychological theory, methods and practice*. Emerald Group publishing limited, 2015.

[22] Shoshannah Tekofsky, Jaap Van Den Herik, Pieter Spronck, and Aske Plaat. Psyops: Personality assessment through gaming behavior. In *Proc. Int. Conf. on the Foundations of Digital Games*, 2013.

[23] Jane Wardle, Carol Ann Guthrie, Saskia Sanderson, and Lorna Rapoport. Development of the children's eating behaviour questionnaire. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 42(7):963–970, 2001.