Multiplayer Tension In the Wild: A Hearthstone Case

Paris Mavromoustakos-Blom
Tilburg University
Tilburg, The Netherlands
p.mavromoustakosblom@uvt.nl

David Melhart
Malta University
Msida, Malta
david.melhart@um.edu.mt

Antonios Liapis
Malta University
Msida, Malta
antonios.liapis@um.edu.mt

Georgios N. Yannakakis
Malta University
Msida, Malta
georgios.yannakakis@um.edu.mt

Sander Bakkes
Utrecht University
Utrecht, The Netherlands
S.C.J.Bakkes@uu.nl

Pieter Spronck
Tilburg University
Tilburg, The Netherlands
p.spronck@uvt.nl

ABSTRACT

Games are designed to elicit strong emotions during game play, especially when players are competing against each other. Artificial Intelligence applied to predict a player's emotions has mainly been tested on single-player experiences in low-stakes settings and short-term interactions. How do players experience and manifest affect in high-stakes competitions, and which modalities can capture this? This paper reports a first experiment in this line of research, using a competition of the video game *Hearthstone* where both competing players' game play and facial expressions were recorded over the course of the entire match which could span up to 41 minutes. Using two experts' annotations of tension using a continuous video affect annotation tool, we attempt to predict tension from the webcam footage of the players alone. Treating both the input and the tension output in a relative fashion, our best models reach 66.3% average accuracy (up to 79.2% at the best fold) in the challenging leave-one-participant out cross-validation task. This initial experiment shows a way forward for affect annotation in games "in the wild" in high-stakes, real-world competitive settings.

CCS CONCEPTS

• **Applied computing** → **Computer games**; • **Computing methodologies** → *Computer vision*; • **Human-centered computing** → *Field studies*.

KEYWORDS

Player affect, player modeling, facial expression analysis, competitive games

ACM Reference Format:
Paris Mavromoustakos-Blom, David Melhart, Antonios Liapis, Georgios N. Yannakakis, Sander Bakkes, and Pieter Spronck. 2023. Multiplayer Tension In the Wild: A Hearthstone Case. In *Foundations of Digital Games 2023 (FDG 2023), April 12–14, 2023, Lisbon, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3582437.3582440

FDG 2023, April 12–14, 2023, Lisbon, Portugal

ACM ISBN 978-1-4503-9855-8/23/04…$15.00
https://doi.org/10.1145/3582437.3582440

1 INTRODUCTION

Video games are considered one of the leading interactive entertainment media, because of their capability to elicit emotions from the participants [21]. Recent studies have employed games in order to analyze the affective state of players [24] as well as the audience of online video game streams [29]. Moreover, analyzing player affect can play a major role during the design of games [46]. The analysis of player affect can enable the implementation of personalized video games, which adapt to players' affective state during game play [28].

Currently, most studies regarding player affective state analysis consider small-scale single-player casual (video) games. However, if performed in a competitive multiplayer environment, the study of player affect can expand beyond examining players' affective responses to game events; it enables the study of facets such as affective player interaction and emotion contagion [7]. Furthermore, the study of player affect "in the wild" [17], although challenging, encapsulates the spontaneity of players' affective responses and boosts generalizability towards real-life game play sessions. Accurately modeling a player's affective state during game play may facilitate not only computers but also humans to map their opponent's affective responses to (hidden) in-game information.

In the present paper, we analyze player affect in the wild during a multiplayer digital card game competition of *Hearthstone* (Blizzard, 2014). Our goal is to predict player tension by analyzing both competitors' facial expressions during game play. We consider player facial expressions a rich manifestation of underlying affect, capable of being unobtrusively collected through a computer's webcam. Furthermore, we select tension as the target affective state as we expect the time-pressured, multi-layered decision making aspect of competitive *Hearthstone* games to elicit this particular feeling onto the competitors. Similar studies have previously attempted to analyze player affect during card games; however, they were conducted under strictly controlled experimental conditions [44].

This paper presents a dataset consisting of over 26 hours of "in the wild" game play and player facial camera recordings, collected during an in-person *Hearthstone* competition. In total, 17 participants played 78 games, which resulted in 156 videos containing the *Hearthstone* game board and players' face cameras. The video footage was processed by two expert players, who annotated the tension of each player in a time-continuous and unbounded fashion for every game played. At the same time, we extracted per-frame estimations of facial action unit intensity values, as well as eye gaze

and head pose estimations using a facial expression analysis toolkit. We treated the extracted dataset in an ordinal, multiplayer fashion and built a machine learning model which predicts temporal changes in player tension by analyzing temporal changes in both participating players' facial expressions. Through this approach, we aim to investigate whether the shared social context (live interactions in a real-world competition) can serve as a predictor of one player's tension, through its influence on both opponents' facial expressions.

We consider this work novel in terms of (a) player affect analysis in a competitive, multiplayer, "in the wild" setting; (b) treatment of this complex dataset through ordinal affect signal processing and leave-one-participant-out cross-validation, and (c) integration of both opponents' facial expressions as predictors for one player's changes in tension. Ultimately, this research may contribute towards the interpretation of players' affective responses with respect to hidden in-game information. The proposed method can be employed for real-time player tension estimation in broadcast e-sports tournaments, the majority of which contain a frontal player face camera. Moreover, we make the curated data (Facial Action Units, tension annotation) available[1] for further research in this testbed.

2 RELATED WORK

Player affect has been recognized as an important factor for the design of engaging (video) games [13]. Games have been employed by researchers to not only elicit emotions, but also to evaluate, express and synthesize them [46]. A central goal of the study of emotion in games is to connect players' emotions to their in-game experiences [46]. The study of human affect through digital means is dubbed *affective computing*, i.e. computing that relates to, arises from, or influences emotions [35].

Facial expressions are one of the most studied modalities in the domain of affective computing, as the face is regarded as the body's most expressive part [33]. The study of the human face does not only consider facial expressions as manifestations of emotion, but may also include eye gaze [20] and head pose tracking [28]. Ekman and Friesen introduced the Facial Action Coding System [9], a system that identifies groups of muscles which are activated in order to form a facial expression. These muscle groups are also referred to as Action Units (AUs). Numerous studies have been conducted in order to associate the activation and intensity of specific AU movements to the underlying human emotions [39]. However, researchers have argued that facial expressions are not always directly correlated to genuine emotions [6]. Several studies have been focused on players' facial expressions during game play [28]. Using multiple modalities (including player facial expressions), Doyran *et al.* [7] collected a rich dataset containing annotations of player affect and interaction analysis during board game sessions. Similarly to the present study, Piton *et al.* [36] used game play sessions of *Hearthstone* to validate the peak-end memory effect in the context of a digital card game; players' retrospective memory of game play sessions was mainly represented by the most emotionally arousing moments or the final moments of the game. Despite the fact that players' self-reports validated the peak-end effect, an analysis of their facial expressions did not yield the same results. In the present study, we will explore whether *Hearthstone* players' facial expressions provide any descriptive information about their feeling of tension. Importantly, we track tension and facial expressions throughout the session, and derive models regarding relative escalation or de-escalation of tension for the same player within the same session.

Assessing the affective state of (video) game players has been central to several studies, varying from clinical play sessions [8] to creating a virtual chess opponent that shows emotion [19]. To human players, emotion recognition can be highly relevant for game play. For example, poker players base their decisions on the assessment of opponents' emotions (manifested mostly through their facial expressions and motor actions) as much as on their own strategic knowledge of the game [14]. This can be generalized over most card games, where players attempt to extract hidden information from their opponents' speech, body motion and facial expression patterns [14]. Specifically, Slepian *et al.* [42] discovered a positive correlation between upper limb motion smoothness and a poker players' hand quality. In similar fashion, Vinkmeier *et al.* [44] were able to predict poker hand folds by analyzing players' facial expressions using the AUs. Apart from facial expression analysis, games have also been used as a medium to recognize, monitor, or even manipulate player stress levels [12]. In this paper, we analyze player facial expressions in order to estimate tension—an affective state that has been related to stress [15].

The link between user affective states and biophysical data under controlled laboratory settings has been established [40]. Laboratory-collected corpora such as RECOLA [38] aim to enable the investigation of human affective interactions in collaborative tasks. Within the context of games, the Pow-Wow dataset [47] incorporates annotations of human communication types during a multiplayer collaborative game. However, new technologies have been developed to enable affect measurement in the wild (i.e. non-controlled experimental conditions) [17]. In this context, extracting a user affect baseline is a highly challenging task; to that end, researchers have employed participant self-report [40] and human expert annotation [1] mechanisms. Various datasets discussing affect measurement in the wild have been published, such as AffectNet [32], SFEW [4], AFEW [5] and FG-Emotions [22]. Regarding facial expression analysis in the wild, the most prominent affect recognition methods are machine learning [43] and deep learning [22]. In this paper, we collect player facial expression measurements in the wild (with respect to changes in illumination, social interactions between players and face occlusion during recording). We employ machine learning methods to estimate player tension and use human expert annotations as ground truth.

3 COMPETITION DATASET AND ANNOTATION

To capture manifestations of emotion in high-stakes settings, a *Hearthstone* competition with commercial rewards was conducted at Tilburg University on May 4th, 2019. This section details the game, participants, protocol, and data collected during the competition, as well as how tension was annotated after the competition.

[1] https://surfdrive.surf.nl/files/index.php/s/3Paqan93nALhJe3

Figure 1: Snapshot of a video recording. Webcam feed is captured at the bottom left and game screen is captured at the top right of the recording video file.

3.1 Hearthstone Competition

We chose *Hearthstone* as the competitive game for this study (see Fig. 1). *Hearthstone* is a one vs. one digital collectible card game and one of the world's leading e-sports (competitive games) titles. In *Hearthstone*, players build decks of 30 cards from a large card collection and compete against their opponent's decks. The goal of the game is to bring the opponent's health points (HP) to zero (HP start at 30). The game is played in turns; on each turn, the acting player draws a card and can use a limited "mana" resource to play cards (monsters or spells) from their hand onto the game board. When a player's HP reaches zero, the game ends.

The competition was set up in best-of-three match format; the first player to win two games would win a match. A player who lost two matches in total was eliminated from the competition. This means that even players who did not win any matches would play at least four games during the competition. Moreover, the semi-finals and finals of the competition were played in best-of-five format (first player to win three games wins the match). After one match was discarded due to webcam recording failure, a total of 31 matches (78 games) were retained in the dataset.

Commercial rewards were offered to the players finishing in the top three positions. A gaming mouse and a handmade souvenir *Hearthstone* card was offered to the player ranked first. Players finishing in 2nd and 3rd positions received a Blizzard gift card. Our experimental protocol included rewards to emphasize the competitive nature of the tournament and to motivate players to win.

Before each match, every player declared three (in best-of-three matches) or four (in best-of-five matches) different decks that they had prepared. Each player was allowed to ban one of their opponent's decks before each match. If a player won a game with one of their decks, the winning deck was not allowed to be re-used in that match.

3.2 Test Environment

The competition was held in a university computer lab area. The lab consists of two rows of five computers facing each other. For each match, the two opponent players were seated on opposing PCs. To ensure eye contact between opponents, the computer monitors were set at the minimum height configuration. A webcam was mounted on top of each monitor, recording player facial expressions at 30 frames per second. Alongside the webcam feed, Open Broadcaster Software (OBS) was used to record the game screen. The Game Lab's computers are equipped with high-end processors and GPUs, as well as a Gigabit cable Internet connection, ensuring minimum latency for recordings and game play.

Only currently competing players were allowed in the lab area, to prevent distractions from the audience. Before their first match, players signed an informed consent form which explained this study's goals and data collection processes. Participants were instructed to remain seated for the entire duration of their match and to refrain from looking at adjacent players' monitors. The latter ensured that (a) players would face their webcam for as long as possible and (b) players would not gather information about their prospective opponents' decks. No other restrictions were applied, resulting in loosely controlled experimental conditions. For example, players could interact with their opponents during matches. We imposed as few restrictions as possible to create suitable conditions for capturing players' spontaneous facial reactions.

3.3 Process of Data Collection in the Wild

Before their first match, participants signed an informed consent form and filled in a short questionnaire regarding their gender, age, estimated hours per week spent on *Hearthstone*, and a subjective scoring on experience with *Hearthstone* on a 1 to 5 scale.

Players' webcam feed and game screens were captured in a single video file, through the OBS software. The webcam feed was positioned at the bottom left of the recording file, while the game screen was positioned at the top right (see Fig. 1). The overlap between the two captures does not occlude any important game features. For each computer, recording started 15 minutes before the launch of the competition, and ended after the final game was played. The start and end timestamps of each game, as well as the relevant participant IDs were manually annotated afterwards by the authors. The start of a game was defined as the moment when the player classes (decks) are announced and shown on players' screens; the end of a game was defined as the moment when the result is shown on the game screen.

While we considered monitoring alternative input modalities (e.g. physiological signals or audio recordings), we chose to follow the data collection protocol described in this section due to the non-invasive and unobtrusive nature of webcam facial recordings. Even though wearable physiological sensors could provide a more reliable measurement of affect, we considered such a solution highly invasive within the context of in-the-wild data collection.

Furthermore, as shown in Figure 1, the player webcam feed may capture players or tournament organizers walking in the background. While the webcam recording could have been setup to zoom in on player's faces only, this would lead to loss of information when players change their posture; therefore, background noise could not be entirely avoided. However, the tournament organizers ensured that players were not allowed to look at their competitors' screens and instructed them to leave the room as soon as their matches were completed. Regarding audio recordings, since *Hearthstone* does not require verbal communication between opponents, we considered that modality irrelevant within the scope of this study.

Figure 2: A snapshot of third-person tension annotation using the PAGAN RankTrace annotation method.

3.4 Participants and Video Dataset

For this study, 17 players (all male, $\mu = 22.7$ years, $\sigma^2 = 3.6$ years of age) signed up and participated in the competition. The average experience of players was $\mu = 3.4$ ($\sigma^2 = 1.2$) out of 5 and the average reported hours per week playing *Hearthstone* were $\mu = 9.4$, ($\sigma^2 = 6.8$). A unique participant ID was assigned to each player, for anonymity purposes.

Data collection resulted in 156 separate game play videos (78 games × 2 players). The total duration of the recordings was 26.5 hours, with an average per-game duration of 10.4 minutes. Each participant played 9.1 games on average. Table 1 summarizes the dataset's main properties.

3.5 Third-person Annotation of Tension

Tension was manually annotated by two expert *Hearthstone* players using the PAGAN continuous annotation tool [30]. The annotation was conducted in June 2020, 13 months after the competition took place. It is important to note that the annotators had no previous experience in appraising human facial expressions. Hence, we expect them to base their annotations mostly on gameplay features. Using PAGAN's RankTrace annotation method [23], we defined *tension* as an unbounded continuous variable. RankTrace allows users to define the degree of change of an affect dimension in an unbounded fashion, while showing the users the entirety of this session's annotation so far (see Fig. 2). Annotators were instructed to annotate "the level of tension in the video" for each game video in the dataset, which means that the annotators considered both the state of the game and the players' facial reactions while annotating. Tension was defined as "a feeling of excitement, exhilaration, and suspense, or frustration and nervousness". Each game was annotated by both annotators separately.

It is important to note that besides tension, the two experts also annotated a player's winning advantage through a per-frame discrete variable, namely +1 (likely to win), 0 (tie) or −1 (likely to lose). However, analysis or prediction of players' winning probability falls outside the scope of this paper.

Table 1: Properties of the Hearthstone competition dataset.

Properties	Dataset
# Participants	17
# Games	78
# Videos	156
Video database size	26.5 hours
Video duration (minutes)	mean: 10.4, min 4.2, max: 40.6
# videos per participant	mean: 9.1, min: 4, max: 21
Annotation Perspective	Third-person
Annotation Type	Continuous unbounded
Affective Labels	Tension

4 TENSION PREDICTION METHODOLOGY

With the video dataset and third-person annotation of tension, in this paper we attempt to model the player's tension in a relative manner (as changes from its previous state) based on one or both players' facial expressions. Video data is captured at 30Hz and the tension annotation is captured at 4Hz; such a granularity is not meaningful when it comes to modeling affect since temporary glitches or annotator lag can introduce noise to the data. Following extensive literature on processing time-continuous annotations [2, 23, 30], we split the data sequences of both input (video dataset) and output (tension) into time windows of 3 seconds, and we shift the tension annotation trace by 1 second when aligning it with the video data [26]. The following sections clarify how we process the video data used as input to our classification task (Section 4.1), how we aggregate the two experts' tension annotations and convert them into class labels (Section 4.2) and which algorithm we use for the classification task (Section 4.3).

4.1 Input

This paper explores how facial expressions and similar visual features from the players' captured feeds can be used as predictors of tension, and the process for extracting such features is presented in Section 4.1.1. Moreover, we evaluate how the opposing player's expressions may also sufficiently capture a player's tension, and we describe three treatments for our experiments in Section 4.1.2.

4.1.1 Extracting Data from Video. While the video data annotated by the two experts include both the game context and the face of the player whose tension is being assessed (see Fig. 1), in this paper we focus only on features extracted from the webcam feed. The OpenFace [3] library is used to extract per-frame estimations of facial Action Unit (AU) presence and intensity, as well as head pose and eye gaze estimations. In total, 17 AU, 2 gaze and 3 pose features are computed on a per-frame basis (the video has 30 frames per second). Table 2 shows the 22 features calculated in this way.

We discard video frames where the OpenFace confidence level was below 75%. This threshold was empirically chosen, as we observed that confidence below 75% introduced inaccuracies either due to face occlusion or out of bounds movement of the head. Removing frames in this fashion is a necessity when the source videos

Table 2: List of features that were extracted from video processing and tension annotation.

Modality	Features
AU intensity	AU01 (Inner brow raiser), AU02 (Outer brow raiser), AU04 (Brow lowerer), AU05 (Upper lid raiser), AU06 (Cheek raiser), AU07 (Lid tightener), AU09 (Nose wrinkler), AU10 (Upper lip raiser), AU12 (Lip corner puller), AU14 (Dimpler), AU15 (Lip corner depressor), AU17 (Chin raiser), AU20 (Lip stretcher), AU23 (Lip tightener), AU25 (Lips part), AU26 (Jaw drop), AU45 (Blink)
Gaze	gazeX, gazeY
Head Pose	poseX, poseY, poseZ

are collected in-the-wild, as users were often moving around, talking, or forgetting about the presence of the webcam. This cleanup step removes only 4.4% of the total frames across all sessions.

Since we need to aggregate these 22 features in 3-sec time windows ($\tilde{9}0$ frames), we calculate both the mean and maximum value of that feature within those 3 seconds. Moreover, since we are interested in predicting the relative change of tension (see Section 4.2), we also calculate the difference of the mean or max feature from the previous time window. With these four ways of processing each of the 22 per-frame features (two absolute as mean and max of a time window, and two relative as the differences between the current and previous time window) we collect 88 inputs per window.

4.1.2 Multiplayer Data as Input. A key research question in this work is how effective the multiplayer perspective is as a predictor for one player's tension changes. Here we introduce three different ways of preparing the input for the classification task described in Section 4.3.

As our baseline, we assume that the player's facial expression reveals their tension levels. This view aligns with a general view within affective computing, where modeling a user's affect from their facial expressions is almost ubiquitous [7, 16, 18]. We identify this treatment as *player in perspective* (PiP), and use the 88 features extracted from the current video that the annotators have assessed in terms of this player's tension. No videos of this player are included in the training set (leave-one-player-out cross-validation as described in Section 4.3).

To assess whether the context of one player's tension can be captured in the other player's facial expression, we create the *opponent player* (OP) dataset which includes the 88 features extracted from the video of the opposing player which captures the same game. Our hypothesis here is that two people share the same context: a competitive game where one event may increase the feeling of tension for both players, but ultimately bring one player closer to winning and the other closer to losing. Therefore, in the majority of cases, the player's reaction to such an event should be the opposite of the opponent's. Moreover, since the experimental setup allowed both players to see each other, the social context may also play a role (e.g. a smirk or laugh at the opponent) in triggering

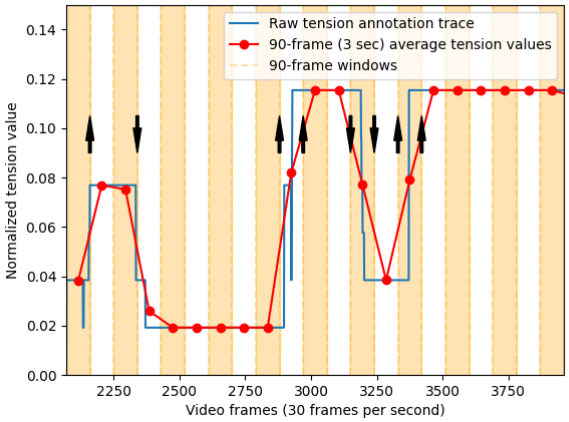

Figure 3: Graphical illustration of the tension ordinal ranking label calculation ($\epsilon = 0.01$). The ordinal ranking labels between successive 90-frame average tension values are illustrated as an upwards arrow (escalating tension) or downwards arrow (de-escalating tension).

facial expressions from both players. Since the video being annotated only displays the webcam of the player in perspective, the opponent player's facial data are extracted from their respective recording. This means that the opponent's facial expression data has been included in the training set, but is only trained to predict that opponent's tension (not the player's in perspective). There is thus no data leakage during leave-one-player-out cross-validation.

Finally, we combine both the features of the player in perspective and the opponent to produce the *mutliplayer* (MP) dataset, which includes 176 inputs from the two videos. Our hypothesis here is that the two players who share the same context (both in terms of game state and real-world social cues) can more accurately capture the reasons why a player's tension level changes. Once again, in leave-one-player-out crossvalidation only the footage and tension traces of the player in perspective are omitted, with no data leakage.

4.2 Output

Our goal is to build robust predictors of a player's tension from their facial expressions during a real-world, high-stakes competition. Here we discuss how we treat tension in a relative fashion to derive class labels (Section 4.2.1) and how we aggregate the assessments of the two experts (Section 4.2.2).

4.2.1 Labels for Relative Tension Changes. We follow [45] and treat emotions in an ordinal fashion. Therefore, we are not interested in the absolute value of tension—which would mean little anyway since RankTrace annotations are unbounded—but rather on how *tension changes over time*. As a first step, we normalize each annotation trace (i.e. in one video and by one annotator) to a value range within $[0, 1]$ via min-max normalization. Since tension annotations are collected at 4Hz, we average the tension values within a 3-sec time window (n) to derive that window's tension (t_n). We then calculate the difference between mean tension in the current time

window and the previous one: $\Delta t = t_n - t_{n-1}$. This relative value represents the point-by-point change (escalating, de-escalating or stable) in each feature's time series. To remove noise, we define a decision threshold ϵ for what constitutes a valid change in tension between consecutive time windows. Using this threshold, we label data points where $\Delta t > \epsilon$ (positive values) as "escalating tension" and $\Delta t < -\epsilon$ (negative values) as "de-escalating tension" (see Fig. 3). All values within $[-\epsilon, \epsilon]$ are considered ambiguous, as the direction of change for tension is not distinct enough, and are discarded for the purposes of classification experiments in Section 5.

We have selected four threshold values in order to incorporate a sensitivity analysis on our labeling algorithm: $\epsilon = \{0.1, 0.05, 0.01, 10^{-6}\}$. At high thresholds, we follow an aggressive approach and remove many instances where the annotation does not dramatically shift between consecutive 3-second intervals. This approach results in a "cleaner" dataset which only contains significant data points, but consequently lowers the volume of training and test data; moreover, there is the risk that some patterns in the data will be missed. At lower thresholds, we are more lenient and consider even smaller changes in the tension annotation to qualify as valid, resulting in a larger but potentially less clear-cut dataset. While past work has often found that the best models perform with a ranking threshold of 0.1 or above [25, 31], we include $\epsilon = 10^{-6}$ in our tests as a baseline that essentially considers all but the smallest changes to be valid tension rankings.

4.2.2 Annotator Trace Aggregation. In order to aggregate the two annotator's tension traces in an ordinal fashion [45], we use an ambiguity-aware (AMBER) aggregation method [41]. After having normalized and labeled both annotators' tension traces, we discard all data points where the annotators' tension labels showed absolute disagreement (escalating vs. de-escalating tension or vice-versa). This means that the data points where the two annotators showed mild disagreement (i.e. where one annotator was ambiguous while the other was not) were still included in the data set, labeled as escalating or de-escalating based on the annotator that was certain. Note that applying AMBER in this way counteracts some of the dangers of high ϵ values which may render much of the dataset as ambiguous, since only one annotator needs to be certain for the data point to be retained for training and testing. In this way, AMBER manages to augment the dataset. Indicatively, for $\epsilon = 0.01$, annotator A's ordinal tension rankings had a per-fold average of 9,900 data points, and annotator B produced a per-fold average of 7,700 data points. The AMBER aggregation resulted in a per-fold average of 13,700 data points. Out of these, an average of 12,900 data points constituted the training set and an average of 800 data points were used for testing.

4.3 Classification Algorithm

Through the processes described above, we collect a dataset to run a classification task, with the "escalating" and "de-escalating" tension labels as target classes. This classification task aims to explore whether, and to what extent, players' facial expressions can be a good predictor of tension during a competitive high-stakes game.

A Random Forest (RF) was the preferred classifier, as it was shown to perform well in the classification of player facial expressions within the context of games [28]. RFs are decision-tree based learning algorithms that can be employed in both classification and regression tasks. We implemented RF classifiers with 100 estimators and entropy as a splitting criterion, using the `scikit-learn` Python library [34].

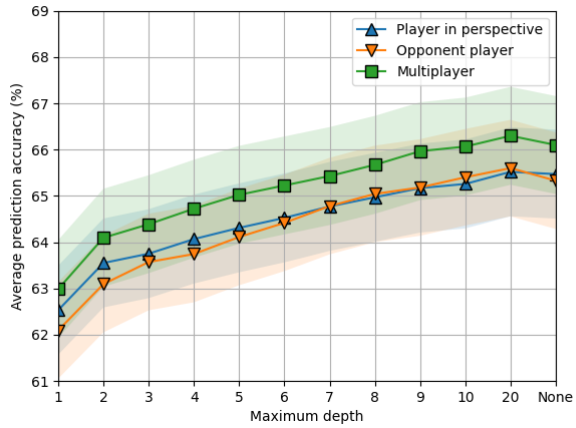

Figure 4: Test classification accuracy of the RF classifier using various maximum decision tree depth settings, using data labeled with $\epsilon = 0.01$. Results are averaged from 5 independent runs, with the 95% confidence interval as shaded area.

The classification task was conducted using a leave-one-player-out method, meaning that the classification task was 17-fold; each fold used a separate player's facial expression data as a validation set, and the predictive model was trained on the remaining 16 players' data. Every classification fold was run 5 times and the average score was calculated. To ensure class balance before classification, two mirrored observations from each pair of consecutive ordinal tension rankings were extracted, following [31]. Specifically, if t_n and t_{n-1} are two consecutive average tension values (at time windows n and $n-1$ respectively) and $y = 1$ (escalating tension) is the label derived from $t_n - t_{n-1}$, we then also include the mirrored data point $y' = -1$ (de-escalating tension, deriving from $t_{n-1} - t_n$) in the dataset. This over-sampling method was applied both to the training set and the test set, ensuring a random guesswork accuracy of 50% for both sets.

5 RESULTS

The main dimension we wish to explore in this paper is the impact of facial data originating from the player in perspective (PiP), the opponent player (OP), or both (MP), in predicting tension changes. However, parameters in the learning process that may affect the predictive power of the Random Forest (RF) include the maximum depth of its trees, and the ambiguity threshold (ϵ) which affects the size and noise of the training and testing set. We conducted leave-one-player-out cross-validation experiments (repeating each experiment five times) with different maximum tree depth parameters, and chose the best maximum depth for each treatment (PiP, OP, MP) and each threshold value. Having a high maximum depth in RFs may lead to over-fitting. As shown for an indicative threshold

Table 3: RF test classification accuracy (%), using different labeling thresholds (ϵ) and the best max depth per setup. Both mean and best fold results are averaged from 5 independent runs, with 95% confidence interval in parentheses.

	Player in perspective		Multiplayer		Opponent player	
ϵ	Mean	Best fold	Mean	Best fold	Mean	Best fold
0.1	61.2 (±1.0)	71.4	61.5 (±0.9)	69.2	60.7 (±1.0)	70.9
0.05	64.3 (±0.9)	74.4	65.3 (±1.0)	76.3	64.6 (±1.1)	75.4
0.01	65.5 (±1.0)	77.9	66.3 (±1.1)	79.2	65.6 (±1.1)	78.0
10^{-6}	65.7 (±1.0)	77.3	66.5 (±1.1)	78.0	65.8 (±1.1)	77.4

$\epsilon = 0.01$ in Fig. 4, higher max depth values lead to more accurate models but allowing any depth (i.e. None) in the RF leads to over-fitting and a drop in test accuracy.

Using the best max depth parameter setup, Table 3 shows the classification accuracy of the three treatments across different labeling thresholds. Although the three different models achieved moderate accuracy versus a random guess baseline of 50%, we observe that the MP model consistently outperforms the PiP model. While there are no significant differences in test accuracies when averaging across all folds, it is more relevant to compare model accuracies on the same fold (i.e. the same unseen player) through a paired Wilcoxon signed-rank test. Except for $\epsilon = 0.1$, for all other thresholds the MP model fold accuracies were significantly higher in the paired test than both PiP and OP models ($p < 0.05$). Indicatively, for $\epsilon = 0.01$ the MP model had strictly higher average test accuracies (from five independent runs) than the PiP model in 15 of 17 folds, and the OP model in 13 of 17 folds. Out of these, looking at the distribution of results per fold across the five independent runs via one-tailed Mann Whitney U tests ($p < 0.05$, 5 samples), the MP model significantly outperforms the PiP model in 12 of 17 folds and the OP model in 10 of 17 runs.

Worth mentioning is that a player's escalating or de-escalating tension seems to be predicted with equal accuracy by their own, or by their opponent's facial expressions separately. Our initial hypothesis was that the tension of the opposing player (which was in the training data) somehow matched the tension of the player in perspective. To provide some insight, we calculated a per-game correlation coefficient of the two participating players' ordinal tension rankings, to examine the similarity of the produced signals. The average coefficient value extracted over all games was 0.08 ($\sigma^2 = 0.12$), indicating no significant correlation between the player's own tension shifts and their opponent's. Generally, we conclude that within the context of this experiment, a multiplayer approach yields more accurate results than a single-player approach.

Regarding the labeling threshold, we observe that $\epsilon = 10^{-6}$ marginally produces the most accurate classifications. However, this threshold may incorporate noise since nearly all shifts in tension are considered; instead, with a threshold of $\epsilon = 0.01$ we achieve comparable accuracies (with a higher accuracy on the best fold) with a less noisy ground truth. We choose therefore to focus on $\epsilon = 0.01$ in our analysis, as evidenced already in Fig. 4. It is evident that aggressive data cleanup with thresholds of $\epsilon = 0.1$ does not result in particularly accurate models even with the MP treatment. Indicatively, the total number of valid tension changes with $\epsilon = 10^{-6}$ is 15,000, dropping to 13,700 with $\epsilon = 0.01$ (8.7% drop); with $\epsilon = 0.1$ only 5,100 valid tension changes remain, almost a third that with the most lenient threshold.

In additional experiments, we investigated whether the two expert annotators' individual traces could lead to different prediction accuracies. At $\epsilon = 0.01$, the best scoring model using only annotator A's traces produced an average prediction accuracy of 64.7% (77% at the best fold). Using annotator B's annotation traces, we reached an average prediction accuracy of 66.5% (79% at the best fold). While accuracies with data from annotator B is comparable to those reported in Table 3, the aggregation method based on inter-rater agreement described in Section 4.2 results in larger datasets and a more reliable ground truth.

6 DISCUSSION

In this paper, we studied the relationship between players' facial expressions and the perception of players' tension via third-person annotations. Using a large dataset of over 26 hours and 78 complete games of *Hearthstone*, we attempt to derive models of tension from facial expressions alone. When predicting a players' ordinal tension rankings through their recorded facial expressions, our best model reached an average accuracy of 65.5%. Moreover, by including the opponent player's facial expressions as inputs, we consistently outperformed models trained on expressions of the player in perspective. Our best multiplayer model reached an average accuracy of 66.3% (79.2% at the best fold). This shows that even though our models were not based on the shared context (e.g. in-game data), opponents' affective states are likely to be connected and synchronously manifested through their facial expressions. Furthermore, after a sensitivity analysis we concluded that a lenient threshold of what constitutes tension change is preferred, since it increases the robustness of the model by incorporating a larger dataset to learn from. The study argues that real-world instances of game play where the tension level may be higher due to high-stakes competition can offer more insight regarding the impact of game play on affect and its manifestations (e.g. via facial expressions). The few studies using real-world footage of high-stakes games have so far focused on poker tournaments [42]. With this study we contribute a novel treatment of real-world high-stakes data that combines facial expressions and game play footage and provide a first exploration for how this dataset can be processed to derive computational models of tension.

While the dataset is rich in both breadth of games (including e.g. short games and very long games) and in modalities captured (game footage, webcam feeds), it should be noted that all participants in the tournament were white males. This bias is a byproduct of the uncontrolled, in-the-wild nature of the event. Participation to the competition was voluntary and open for all ages and genders; participants were not manually selected. Earlier studies on players' posture and expression have shown that different genders can manifest emotion while playing differently [27]. Future experiments should strive for a broader diversity in the players captured. Additionally, future experiments could explore alternative inputs, such as player posture and gestures. Such modalities, while still extracted through non-invasive webcam recordings, may prove to

be accurate descriptors of player affective states. On the contrary, while wearable wristbands could provide reliable physiological measurements of affect, we consider these types of sensors invasive and believe that their measurements are not transferable from strictly controlled experimental conditions to in-the-wild data collection.

Another limitation in our current work is that the third-person annotation of tension proved to be challenging as evidenced by the different levels of accuracy achieved from each annotator's traces. We hypothesize that a first-person annotation trace, e.g. by the player watching their own game play footage after the game, could better capture the ground truth of their affective states. However, such annotations may add more noise as each player can assess only their own footage and annotate idiosyncratically. Moreover, asking players to annotate 10-minute videos during a competition could hinder their emotional commitment to the high-stakes, "real-world" nature of the eliciting event. Lastly, recent studies have shown a connection between player facial expressions and level of expertise [10, 11]. This may indicate that expert players deliberately regulate their facial expressions to hide their emotions about the current game state from the opponent. As a consequence, detecting tension from these players' facial expressions becomes an increasingly challenging task.

This first exploration of the relationship between players' expressions and tension can be expanded with additional modalities. Since *Hearthstone* is a fairly long game, additional experiments could explore alternative treatments of the tension shifts rather than the current ordinal rankings. Potential alternatives could explore longer time windows when averaging both features and tension values (e.g. using each player's game turn as time window), or finding tension shifts with dynamic (and longer) time intervals rather than the current constant (and short) window of 3 seconds. An obvious next step is to include the raw pixels of the webcam footage as additional input to the processed features from OpenFace. More ambitiously, the entirety of the captured video, including the game play footage, could better capture the context for both the player's facial expression and the annotators' perceived tension. Since annotators were shown both game play footage and webcam feed (see Fig. 1), it is possible that one annotator may have focused on the webcam feed while another may have focused on the game play context to assess tension. Through the use of deep learning and multimodal fusion [37] it is possible to capture the patterns both in the pixels of game play [24] and the pixels of the player's webcam feed and thus better predict emotion. Lastly, we expect deep learning models to potentially detect players' non-facial reactions, such as hand gestures, head pose and eye gaze movements.

Beyond the perceived tension of players, future work will explore how facial expressions or even the predicted tension can be used as predictors of game play outcomes, such as the final winner of the match or the experts' per-frame winning advantage annotation. We believe that such predictions, if extracted through explainable AI, could be leveraged by competitors in order to extract hidden in-game information (e.g. an opposing player's evaluation of the current game state) from their opponents' affective signals.

7 CONCLUSION

This paper studies how players' facial expressions in a real-world and high-stakes competition of a popular (video) game can be used as predictors of the players' tension levels. The specifics of the dataset, collected in 2019 during a *Hearthstone* competition and combining webcam feed with in-game footage, pose interesting challenges as discussed in the paper. The long (and inconsistent) duration of the videos, the imbalance of the dataset in terms of how often certain participants appear, and the need for third-person annotation make the problem more difficult. By treating both inputs and tension outputs in a relative fashion, and by observing the fluctuation in tension levels through a multiplayer perspective, this paper offers a way forward for processing game play footage collected "in the wild". The takeaways and key next steps that emerge from this study can help drive research in this challenging affect modeling task.

ACKNOWLEDGMENTS

The authors would like to thank the Tilburg University Esports Association "Link" for their collaboration. Antonios Liapis and Georgios N. Yannakakis were supported by the European Union's H2020 research and innovation programme (Grant Agreement No. 951911).

REFERENCES

[1] Egils Avots, Tomasz Sapiński, Maie Bachmann, and Dorota Kamińska. 2019. Audiovisual emotion recognition in wild. *Machine Vision and Applications* 30, 5 (2019), 975–985.

[2] Değer Ayata, Yusuf Yaslan, and Mustafa Kamaşak. 2016. Emotion recognition via random forest and galvanic skin response: Comparison of time based feature sets, window sizes and wavelet approaches. In *Proc. of the Medical Technologies National Congress.*

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *Proc. of the IEEE Intl. Conf. on Automatic Face & Gesture Recognition.* 59–66.

[4] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2011. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *Proc. of the Intl. Conf. on Computer Vision Workshops.* 2106–2112.

[5] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. 2012. A Semi-Automatic Method for Collecting Richly Labelled Large Facial Expression Databases from Movies. *IEEE Multimedia* 19, 3 (2012), 01–34.

[6] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. 2012. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. In *Proc. of the European Conf. on Computer Vision.* Springer, 525–538.

[7] Metehan Doyran, Arjan Schimmel, Pınar Baki, Kübra Ergin, Batıkan Türkmen, Almıla Akdağ Salah, Sander CJ Bakkes, Heysem Kaya, Ronald Poppe, and Albert Ali Salah. 2021. MUMBAI: multi-person, multimodal board game affect and interaction analysis dataset. *Journal on Multimodal User Interfaces* 15, 4 (2021).

[8] Metehan Doyran, Batıkan Türkmen, Eda Aydın Oktay, Sibel Halfon, and Albert Ali Salah. 2019. Video and Text-Based Affect Analysis of Children in Play Therapy. In *Proc. of the Intl. Conf. on Multimodal Interaction.* 26–34.

[9] Paul Ekman and Wallace V Friesen. 1978. *Manual for the facial action coding system.* Consulting Psychologists Press.

[10] Gianluca Guglielmo, Paris Mavromoustakos Blom, Michal Klincewicz, Boris Čule, and Pieter Spronck. 2022. Face in the game: Using facial action units to track expertise in competitive video game play. In *Proc. of the IEEE Conference on Games.* 112–118.

[11] Gianluca Guglielmo, Paris Mavromoustakos Blom, Michal Klincewicz, Elisabeth Huis in't Veld, and Pieter Spronck. 2022. Blink to win: Blink patterns of video game players are connected to expertise. In *Proc. of the International Conf. on the Foundations of Digital Games.*

[12] Christoffer Holmgård, Georgios N Yannakakis, Héctor P Martínez, Karen-Inge Karstoft, and Henrik Steen Andersen. 2015. Multimodal PTSD characterization via the startlemart game. *Journal on Multimodal User Interfaces* 9, 1 (2015), 3–15.

[13] Eva Hudlicka. 2008. Affective computing for game design. In *Proc. of the North American Conf. on Intelligent Games & Simulation.* 5–12.

[14] Heidi Johansen-Berg and Vincent Walsh. 2001. Cognitive neuroscience: who to play at poker. *Current Biology* 11, 7 (2001), R261–R263.

[15] John H Kerr, Hakuei Fujiyama, and Jessica Campano. 2002. Emotion and stress in serious and hedonistic leisure sport activities. *Journal of Leisure Research* 34, 3 (2002), 272–289.

[16] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2012. DEAP: A database for emotion analysis using physiological signals. *IEEE Trans. on Affective Computing* 3, 1 (2012), 18–31.

[17] Dimitrios Kollias, Mihalis Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. 2017. Recognition of affect in the wild using deep neural networks. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition Workshops*. 1972–1979.

[18] Jean Kossaifi, Robert Walecki, Yannis Panagakis, Jie Shen, Maximilian Schmitt, Fabien Ringeval, Jing Han, Vedhas Pandit, Antoine Toisoul, Bjoern W Schuller, et al. 2019. SEWA DB: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 43, 3 (2019), 1022–1040.

[19] Gy Kovács, Zs Ruttkay, and A Fazekas. 2007. Virtual chess player with emotions. In *Proc. of the Hungarian Conf. on Computer Graphics and Geometry*. 182–188.

[20] Michael Lankes, Bernhard Maurer, and Barbara Stiglbauer. 2016. An eye for an eye: Gaze input in competitive online games and its effects on social presence. In *Proc. of the Intl. Conf. on Advances in Computer Entertainment Technology*.

[21] Nicole Lazzaro. 2009. Why we play: affect and the fun of games. *Human-computer interaction: Designing for diverse users and domains* 155 (2009), 679–700.

[22] Liqian Liang, Congyan Lang, Yidong Li, Songhe Feng, and Jian Zhao. 2020. Fine-grained facial expression recognition in the wild. *IEEE Trans. on Information Forensics and Security* 16 (2020), 482–494.

[23] Phil Lopes, Georgios N Yannakakis, and Antonios Liapis. 2017. Ranktrace: Relative and unbounded affect annotation. In *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*. 158–163.

[24] Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. 2019. From Pixels to Affect: A Study on Games and Player Experience. In *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*.

[25] Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. 2021. The Pixels and Sounds of Emotion: General-Purpose Representations of Arousal in Games. *IEEE Trans. of Affective Computing* (2021). accepted.

[26] Soroosh Mariooryad and Carlos Busso. 2014. Correcting time-continuous emotional labels by modeling the reaction lag of evaluators. *IEEE Trans. on Affective Computing* 6, 2 (2014), 97–108.

[27] Rosa Mikeal Martey, Jennifer Stromer-Galley, Jaime Banks, Jingsi Wu, and Mia Consalvo. 2014. The strategic female: gender-switching and player behavior in online games. *Information, Communication & Society* 17, 3 (2014), 286–300.

[28] Paris Mavromoustakos Blom, Stefan Methorst, Sander Bakkes, and Pieter Spronck. 2020. Modeling and adjusting in-game difficulty based on facial expression analysis. *Entertainment Computing* 33 (2020).

[29] David Melhart, Daniele Gravina, and Georgios N Yannakakis. 2020. Moment-to-moment Engagement Prediction through the Eyes of the Observer: PUBG Streaming on Twitch. In *Proc. of the Intl. Conf. on the Foundations of Digital Games*.

[30] David Melhart, Antonios Liapis, and Georgios N Yannakakis. 2019. PAGAN: Video affect annotation made easy. In *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*. 130–136.

[31] David Melhart, Antonios Liapis, and Georgios N. Yannakakis. 2021. Towards General Models of Player Experience: A Study Within Genres. In *Proc. of the IEEE Conf. on Games*.

[32] A. Mollahosseini, B. Hasani, and M. H. Mahoor. 2019. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. on Affective Computing* 10, 1 (2019), 18–31.

[33] Fatemeh Noroozi, Ciprian Adrian Corneanu, Dorota Kamińska, Tomasz Sapiński, Sergio Escalera, and Gholamreza Anbarjafari. 2021. Survey on Emotional Body Gesture Recognition. *IEEE Trans. on Affective Computing* 12, 2 (2021), 505–523.

[34] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[35] Rosalind W Picard. 2000. *Affective computing*. MIT press.

[36] Agner Piton, Paris Mavromoustakos Blom, and Pieter Spronck. 2020. Exploring Peak-End Effects in Player Affect through Hearthstone. In *Proc. of GAME-ON*. 51–56.

[37] Dhanesh Ramachandram and Graham Taylor. 2017. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Processing Magazine* 34 (2017), 96–108.

[38] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *Proc. of the IEEE Intl. Conf. and workshops on automatic face and gesture recognition*.

[39] Klaus R Scherer, Heiner Ellgring, Anja Dieckmann, Matthias Unfried, and Marcello Mortillaro. 2019. Dynamic facial expression of emotion and observer inference. *Frontiers in psychology* 10 (2019), 508.

[40] Philip Schmidt, Robert Dürichen, Attila Reiss, Kristof Van Laerhoven, and Thomas Plötz. 2019. Multi-Target Affect Detection in the Wild: An Exploratory Study. In *Proc. of the Intl. Symposium on Wearable Computers*. 211–219.

[41] Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth Narayanan. 2019. The Ambiguous World of Emotion Representation. https://doi.org/10.48550/ARXIV.1909.00360

[42] Michael L Slepian, Steven G Young, Abraham M Rutchick, and Nalini Ambady. 2013. Quality of professional players' poker hands is perceived accurately from arm motions. *Psychological science* 24, 11 (2013), 2335–2338.

[43] Thai Son Ly, Nhu-Tai Do, Soo-Hyung Kim, Hyung-Jeong Yang, and Guee-Sang Lee. 2019. A novel 2D and 3D multimodal approach for in-the-wild facial expression recognition. *Image and Vision Computing* 92 (2019).

[44] Doratha Vinkemeier, Michel Valstar, and Jonathan Gratch. 2018. Predicting folds in poker using action unit detectors and decision trees. In *Proc. of the Intl. Conf. on Automatic Face & Gesture Recognition*. 504–511.

[45] Georgios N. Yannakakis, Roddy Cowie, and Carlos Busso. 2021. The ordinal nature of emotions: An emerging approach. *IEEE Trans. on Affective Computing* 12, 1 (2021), 16–35.

[46] Georgios N Yannakakis and Ana Paiva. 2014. Emotion in games. *Handbook on affective computing* 1, 1 (2014), 459–471.

[47] Takuma Yoneda, Matthew R Walter, and Jason Naradowsky. 2020. Pow-Wow: A Dataset and Study on Collaborative Communication in Pommerman. In *Proc. of the ICML Workshop on Language in Reinforcement Learning*.