Inter-Player Data for the Prediction of Emotional Intensity in a Multiplayer Game

Alexander Brooke
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, England
0009-0009-5907-9044

Matthew Crossley
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, England
0000-0001-5965-8147

Huw Lloyd
Department of Computing and Mathematics
Manchester Metropolitan University
Manchester, England
0000-0001-6537-4036

Stuart Cunningham
School of Computer and Engineering Sciences
University of Chester
Chester, England
0000-0002-5348-7700

***Abstract*—This work assesses the feasibility of predicting emotional intensities for a given player in a testbed multiplayer game, using facial expression data collected from other players in the multiplayer group. Whilst there is significant literature on the utilisation of affect detection to build models of player experience, little research considers the additional data provided from other players in a multiplayer setting, despite the inherently shared experiences that they provide. A dataset describing 24 participants is collected, detailing ten levels of a testbed game, *Colour Rush*, with data collected describing facial expression activity and responses to the Discrete Emotions Questionnaire. The viability of modelling uncaptured player experiences is tested using artificial neural networks trained on facial expression data from target players, non-target players and a combination of both. Findings indicate that multiplayer data can be beneficial in the prediction of a target player's emotional responses, although this holds true only in a minority of cases, and for specific groups of players.**

***Index Terms*—Emotion, multiplayer games, player experience modelling**

I. Introduction

Player modelling describes the creation of an abstract representation of a player, from which preferences, behaviour, or experience can be predicted [1]. Using a player model, content can be adapted, selected, or generated, creating level layouts [2], soundscapes [3], or narratives [4], or can be used to control non-player characters [5], better fitting the targeted aspects of an individual player. Research into multiplayer games then shows promise for experience modelling, with previous work by Tan et al., describing a greater level of expressivity shown by players in multiplayer settings [6]. Predicting player experiences utilising data leveraged specifically from a multiplayer setting is however largely under-researched, despite the vast quantity of multiplayer games available in the market [7].

Previous work describes the creation of player experience models based on various modalities of data, often tied to gameplay itself [2, 5, 8], emotion recognition via affect detection [9, 10], or tied directly to self-report [11, 12]. Combinations of these modalities has been shown to create models more accurately describing player experience [13], with fusion of methods of affect detection [14, 15], or gameplay events and physiological signals [16, 17], being common avenues for investigation. Exploration into use of the additional modalities provided by multiplayer settings is therefore of interest, whether to optimise models of individual or group experiences.

Modalities of data collected in parallel to a target player may also present benefits for player modelling, with one aim being that of modelling 'uncaptured' experiences, in which the experience of a target player is not captured using any form of physiological measurement. Under various conditions, a game seeking to capture player experience data may be unable to do so, whether due to a player's lack of hardware, technical problems, or the player's unwillingness to be directly monitored. In these situations, it may be of use to model a player's experience based on other methods, should they show approval of this.

This work first attempts to predict players' responses as a validated measure of discrete emotional experience, using data collected from other players in a multiplayer scenario, emulating a situation in which the target player's data is uncaptured. This is compared with attempts to predict a player's responses using their own data. Further to this, predictive models are built using data from both the target and ancillary players, to understand the effect of combining these modalities.

It should be noted that the authors expect the genre of a game, and each group of players' background to have a great impact on the relationship between player's experiences. A cooperative game for example may create similar experiences between players, whilst a competitive game may create polarising experiences. This work therefore seeks to provide initial exploration into the feasibility of modelling uncaptured experiences through inter-player data, as opposed to highlighting relationships expected across all multiplayer games. With this in mind, this work poses an initial response

to two research questions:

RQ1 To what extent can self-report measures of affect for one player be predicted using measures of affect collected from other players during a shared gaming experience?

RQ2 To what extent can predictions of self-reported experience using one player's data be improved through the addition of data collected from others who shared the gaming experience?

For this, participants were tasked with playing through a series of levels in a testbed multiplayer game, with emotional response collected in the form of facial expression activity, and self-report, captured for post-hoc analysis.

II. Related Work

Work by Mavromoustakos-Blom et al., [18] presents one example of experience modelling using data collected from a player other than the one targeted. In their study, players competed in pairs, with models built using random forests to classify escalating and de-escalating levels of tension between frames of facial expression data from a target player, their opponent, and both players together. Across all tests, small but consistent improvements to model accuracy were made using input describing both players' facial expressions, despite no correlation being found between the tension observed for each. In response to these findings, this work seeks to utilise similar input data to estimate player experience over the course of entire game levels, as measured using self-report, following methods similar to previous work [10], whilst investigating this approach in a larger group setting.

Similar to the work of Mavromoustakos-Blom et al., [18], various studies find relation between the emotions of individuals in dyads [19, 20], although these often focus on communication and generalised applications, rather than feasibility in gaming contexts. Of this, much work relates to the effect of emotion contagion (the effect by which an individual's emotional expression will affect the experience of others [21]), which has been shown to result in increased emotional responses between participants with closer relationships [22]. Game related works considering emotion contagion often relate to its simulation in virtual systems [23], whilst work considering the effect in further fields of research often describe its use in tasks such as crowd control [24], overall group emotion recognition [25], or prediction using physiological modalities inappropriate for active deployment in contemporary game design efforts [26] (focusing on physiological synchronicity).

Unbounded by the logistical complexity of studying multiplayer games, many approaches to experience modelling in singleplayer games make use of physiological indicators of affect, such as facial expression, skin conductance, heart rate and brain activity [27], relating these readings to self-report measures of experience, such as the Game Experience Questionnaire [28]. The comparative lack of studies taking advantage of the additional data provided by a multiplayer setting however, leaves open questions related to its use.

Of the modalities discussed, facial expression recognition (FER) is selected due to its ready deployment, and the comparative unobtrusiveness and feasibility of its use discussed in previous work [6, 29]. FER also aligns well with an understanding that facial expression is a key factor in emotion contagion and synchronicity [30], although further insight is expected of similarly applicable work using audio data to consider conversation tone and content.

The Discrete Emotions Questionnaire (DEQ) [31] is selected as a measure of discrete emotional intensity, expected to be directly applicable to the data collected via facial expression analysis, and having previously been shown to adequately describe gameplay experiences [32]. This also follows exploratory work utilising facial expression data to predict the emotional intensity of individual players, as measured by the DEQ [33].

III. Method

In response to the posed research questions, a study was designed in which data collected from players of an original testbed game was used to predict emotional response in other players. This section describes the participant sample, the emotional stimulus used, and the feature groups compiled for further study. This is followed by an overview of the analysis conducted.

A. Participants

Following institutional ethical approval, forty participants were recruited to provide data for this and further studies, each taking part in a randomly allocated group of four (allowing for availability). Participants were recruited using convenience sampling, with the study being advertised on social media and in university shared areas. Demographic information collected ahead of the study describe a predominantly white British group aged between 18 and 34 (35 aged 18-24 and 5 aged 25-34). The majority of participants were male (27 male, 10 female, 2 other, 1 preferred not to say) and all bar two reported playing video games for three or more hours per week. An information sheet and consent form were provided ahead of the study, with eligible respondents (aged 18 or over with a Windows machine, webcam and microphone) being selected for participation. All participants took part in the study remotely from one another, playing the game online on their own hardware.

B. Colour Rush

The emotional stimulus used for the study was original testbed game *Colour Rush*.

Colour Rush is a four player semi-cooperative top-down 2D game, developed in the Unity game engine [34], in which players are tasked with collecting coins in a series of ten procedurally generated levels. Coins are scattered throughout each level, but can also be obtained through the completion of colour mixing tasks, which are more easily completed by working as part of a team, promoting cooperation. Players are also provided with a flamethrower than can be used to further exploration or engage other players in combat, promoting competition. A section of a level from *Colour Rush* is displayed in Figure 1.

Further description of the gameplay in *Colour Rush* is given in previous work [33].

Fig. 1. The starting area in a level of *Colour Rush*. Pictured are: a player, the level's drop-off point for colour mixing and splitting tasks, barrels containing coins, paint blobs, and environmental tools for mixing and splitting the paint blobs.

Ahead of starting the game, participants were asked to rate how well they knew the other participants in their group, rating this on a five-point Likert scale, from "Not at all" to "Very well". After each level played, participants then completed the DEQ, responding to 32 emotional descriptors on a seven-point Likert scale. Responses to each descriptor were totalled into the DEQ's eight subscales of anger, anxiety, desire, disgust, fear, happiness, relaxation and sadness, resulting in 0-24 ratings for each emotion. Low variability in responses for all emotions but desire (SD=5.015), happiness ($SD = 5.278$) and relaxation ($SD = 5.290$) led to them being removed from analysis. The remaining three emotions were normalised to a 0-1 scale using min-max normalisation, and form the focus of this study, with meaningful insights for the removed emotions expected to be drawn more clearly from data collected in other genres of games and multiplayer settings.

C. Feature Groups

During each level, visual data was collected from participants' webcams. Footage was processed using OpenFace 2.2 [35], creating 17 facial Action Unit (AU) intensities per frame. Reported confidence for each frame was used to remove 0.59% of the data collected, using a threshold of 0.75, as in work by Mavromoustakos-Blom et al., [18]. Remaining intensities were smoothed in a 0.5 second window using median filtering [36], with any remaining gaps caused by confidence culling filled using linear interpolation, as in Doyran et al., [37].

AU intensities were transformed into intensities for the basic emotions proposed by Ekman [38], averaging the intensity of prototypical AUs [39] for each, as in previous work [10].

Of the sessions recorded, incomplete data for four participants led to their respective groups being removed from the study. Data is therefore present for a total of 240 player-levels collected from the remaining six groups, each with complete DEQ responses and facial expression recognition data.

Mean intensity and interquartile range (IQR) for each basic emotion over the course of each level was calculated for each player, creating twelve facial expression metrics per player-level. This was then used to form three distinct feature groups, each with target data describing the three target DEQ emotions.

1) Self: 240 sets of facial expression metrics paired with each series of DEQ results from the same player.

2) Group: The 240 sets of DEQ results, paired with the mean of each facial expression metric from all non-target players in the same group.

3) Paired: The 240 sets of DEQ results, paired with the facial expression metrics for each other player in the same group, creating a total of 720 pairs of input and output data.

D. Correlation Analysis and Predictive Modelling

Response to RQ1 was provided through correlation analysis of each feature group, and predictive modelling using feed-forward artificial neural networks (ANNs) seeking to test the predictive power of the data collected.

Expanding on the methodology applied by Pedersen, Togelius and Yannakakis [2], we utilise a nested cross validation approach. For this, the dataset was separated into six folds in a leave-one-group-out (LOGO) approach, ensuring each fold contained all of the data for a single group to ensure no data leakage from the same participant or group between folds.

Sequential Forward Selection (SFS) was chosen as an efficient algorithm to select features at each outer fold, allowing for consideration of each feature group at varying subset sizes. The SFS algorithm incrementally built feature sets by testing all combinations of a retained feature subset and each remaining available feature, retaining features at each iteration that maximised overall performance using a selected performance metric. For this, we use the performance of single and multi-layer perceptrons (SLPs and MLPs), emulating the approach seen in work by Pedersen et al., [2] to determine the utility of each feature subset. Performance estimations for each subset at each stage of the SFS algorithm were calculated as the average performance of the most successful models seen across each fold of an inner LOGO cross validation approach (utilising the five test groups from each outer fold). Performance for each model was calculated as the mean squared error (MSE) between predictions and their true values.

Feature subsets were built at all sizes, ranging from the single most successful predictor of affect, up to all 12 features per feature group to allow for comparison between each at each subset size. Additionally, we include models trained using singular zero values as input data, thereby utilising zero features from each feature group, and providing a baseline at which only the distribution of results from each fold was used to train the resulting models. Through this, SFS was used to create a total of 65 feature subsets per feature group, per emotion, per outer fold, and across five independent repetitions of the entire experiment.

The optimal network topology for each feature subset was then determined through grid search of hidden layer neuron counts from 1-30 and 1-10 for each of the two hidden layers provided to each MLP, further following the methodology applied in Pedersen et al., [2], and again training and testing using LOGO cross validation internal to each outer fold. Final performance values for each feature subset were then calculated as the MSE of predictions made by models trained

on each outer fold's entire test set (data from five groups) and tested on data from the left out group. Final performance values for each feature subset therefore describe the performance of a model tested on completely unseen data.

IV. Results and Discussion

This section describes results from initial correlation analysis of the collected data, and predictive modelling of players' emotional intensity as measured using the DEQ.

A. Correlation Analysis

Correlation analysis was conducted for each feature group, testing for linear relationships between each player's DEQ responses for target emotions of desire, happiness and relaxation, and their own facial expression statistics, or those of the other players. Results of this analysis are summarised in Table I, in which all correlation coefficients significant at $p < 0.05$ are given for each feature group and target emotion.

TABLE I
Features Significantly Correlated with Targeted DEQ Emotions at $\alpha = 0.05$

Features	n	Desire (r)	Happiness (r)	Relaxation (r)
Self	240	-	Happiness IQR (0.275) Disgust Mean (-0.202) Sadness Mean (-0.160) Fear Mean (-0.156)	Disgust Mean (-0.205) Sadness IQR (-0.173) Surprise Mean (-0.171) Fear IQR (-0.156)
Group (Other Players)	240	-	Disgust Mean (0.381) Sadness IQR (0.246) Sadness Mean (0.246) Fear IQR (0.246) Surprise Mean (0.243) Fear Mean (0.239) Surprise IQR (0.231) Happiness IQR (0.188) Happiness Mean (0.179) Disgust IQR (0.170) Anger Mean (0.130) Anger IQR (0.129)	-
Paired (Other Player)	720	Happiness IQR (0.084)	Disgust Mean (0.248) Sadness IQR (0.157) Fear IQR (0.154) Surprise Mean (0.145) Surprise IQR (0.139) Happiness IQR (0.136) Sadness Mean (0.136) Fear Mean (0.130) Disgust IQR (0.109) Happiness Mean (0.103)	-

Correlations are generally weak, although clearly describe a relationship between player's facial activity and responses to the DEQ items contributing to the happiness subscale, even between players. The strongest correlation, for example, is found between felt happiness according to the DEQ, and the mean intensity of disgust seen across all other players. Consideration of these results prompted further manual inspection of the collected footage, breaking each emotion back down to its prototypical AUs [39]. Whilst other emotions were considered well represented, high intensity moments of disgust were often observed in situations where all of the requisite AUs (9, 10 and 17) were activated during moments of intense laughter. In contrast, few moments of true disgust were observed by the authors, with the aesthetic of the game not conforming to its elicitation. Expressions of disgust as observed whilst playing *Colour Rush* are therefore expected to more closely align to moments of malicious laughter as described by Nikopoulos [40], which both fits with the apparent relationship between disgust and the happiness of others, and the competitive aspects of the game. This also aligns with the negative correlation seen between AU10 (and therefore Disgust) and DEQ relaxation, with relaxation expected to relate to less competitive moments.

The relationship between metrics describing each basic emotion expression for other players and the target player's DEQ responses for the happiness subscale suggest a relationship between happiness and any facial movement in players from the group, with all significant relationships between DEQ happiness and inter-player data being positive. In contrast, player's own facial expressions are more specifically related to their emotional experience, with the positive and negative linear relationships found aligning with expectations. This is seen also in relationships between relaxation and the self data, with the mean of further semantically negative expressions of sadness and fear also relating negatively to the emotion of relaxation. The lack of significantly related features from each feature group and the emotion of desire is unexpected, although consideration that the desire subscale may have been the most ambiguous in a gaming scenario may explain noise in the data collected. This suggests a potential limitation in the use of the DEQ in gaming contexts, although validation of the DEQ-VG in later work may suggest opportunities for the tool's adaptation to gaming scenarios [32].

B. Predicting Uncaptured Experience

Following the predictive modelling described in section III, performance results were analysed for statistical differences. Figure 2 describes the MSE of predictive models predicting desire, for each feature count and feature group, with metrics from the five independent tests and six groups aggregated for clarity. From this, a clear underperformance of models trained using paired data can be seen, with closer similarity observed between the self and group models. Despite this similarity, models using group data only outperformed models using each player's own data at the one and two feature count level, suggesting promising initial improvement in the use of group data, but overall better performance when using a player's own facial expressions. The group models' improvement over predictions utilising zero features, or those using paired data does however suggest that utilisation of facial expressions from the remaining players in a group does indeed provide usable information around which to predict the intensity of emotion felt by a player for which no data was captured. Across models using each feature group, improvement is generally seen with the addition of each new feature, although this benefit depreciates, as would be expected, given that each additional feature was deemed less useful by SFS.

A similar trend is observed in Figure 3, in which the performance of models predicting player's responses to the DEQ Happiness subscale are described. Paired data models again underperform in comparison to both the group and self models. In the prediction of happiness, against what linear relationships would suggest, group models routinely perform more poorly than self models, with the relationship between even single features, such as the previously discussed mean of disgust expressions across the group, providing less

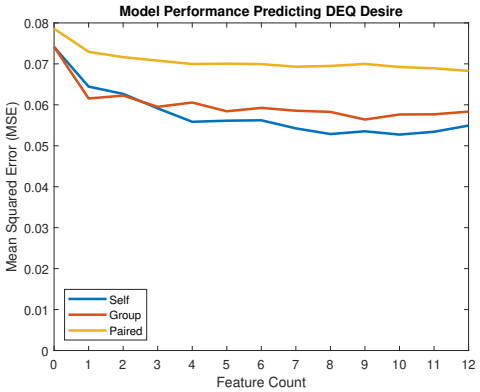

Fig. 2. The average performance (MSE) of models predicting DEQ desire, when using 0-12 features selected using sequential forward selection on data from a player's own facial expression data (Self), the average of other players in the group's facial expressions (Group) and the facial expressions of other players in the group individually (Paired).

useable data in the general prediction of the subscale. This suggests a disparity in the relationships seen, when considered between each group of participants, and therefore folds of data used in cross validation. This disparity is however lessened when considering each participant's own data, leading to more consistently improved models, despite fewer significantly correlated features.

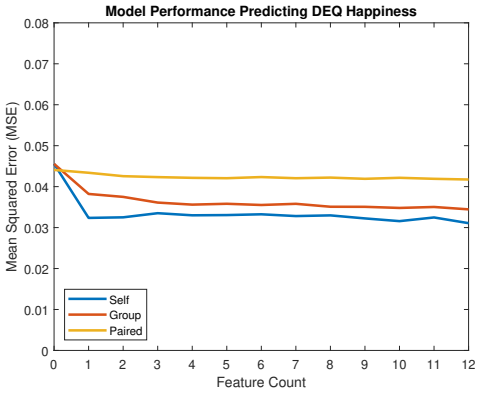

Fig. 3. The average performance (MSE) of models predicting DEQ happiness, when using 0-12 features selected using sequential forward selection on data from a player's own facial expression data (Self), the average of other players in the group's facial expressions (Group) and the facial expressions of other players in the group individually (Paired).

The same pattern is again seen in prediction of relaxation, as described by Figure 4. Here, group models performs more similarly to paired models, again suggesting incomparability between the relationships seen across groups.

The general underperformance of models using paired data across each emotion is expected, with the difference in feature-to-output relationships suggested between groups expected to be further prevalent in the paired dataset. A difference in group dynamics is expected to have impacted this, with difference in how players cooperated and competed across levels expected to impact the relationship between expressions seen in individual pairings of players. One player may have worked cooperatively with the target player for example, whilst another may have worked against them, resulting in an overall difference in the relationships seen, and leading

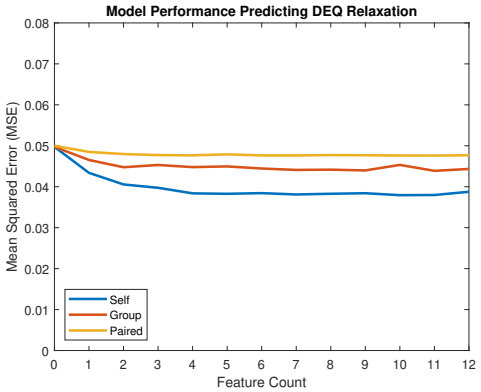

Fig. 4. The average performance (MSE) of models predicting DEQ relaxation, when using 0-12 features selected using sequential forward selection on data from a player's own facial expression data (Self), the average of other players in the group's facial expressions (Group) and the facial expressions of other players in the group individually (Paired).

to noise in the dataset. Further work considering inter-player data may therefore consider differences between different multiplayer game modes, to understand whether differences in how players compete and cooperate impact results.

C. Combining Inter-Player Data

Given the relative success of models using group data over paired data, we focus further work utilising inter-player data in response to RQ2 on the group dataset. Regarding this, further tests were conducted using the same methodology, selecting features using SFS internally to an outer cross validation loop, this time from a dataset combining the self and group average feature groups, for a total of 24 features, that we hereby refer to as the combined dataset for brevity. From this, we collect model performances for models using between 0 and 12 features, for comparability with models trained on the self data alone. Figure 5 describes the MSE achieved by models using features from the combined dataset for the prediction of each target emotion, as well as the MSE of models using self data for comparative purposes.

In multiple cases, the incomparability between folds expected to contribute to poor performance with the group dataset appears to also impact models using the combined dataset, with features selected for high performance on each fold's training data performing more poorly than self models when tested on the fold's unseen data. Generally, however, these models perform similarly, suggesting little benefit in the use of inter-player group data in addition to target player affect.

D. Statistical Analysis

Responses to both RQ1 and RQ2 are given following substantiation of the observed trends through statistical analysis. For this, we apply various non-parametric statistical tests, consistent with previous work [18], and the violated assumptions necessary for parametric testing (with Shapiro Wilk tests on various subsets of the data suggesting significant deviation from a normal distribution). All statistical results are deemed significant at $p < 0.05$. We first consider the difference between each feature group (self, group, paired and combined)

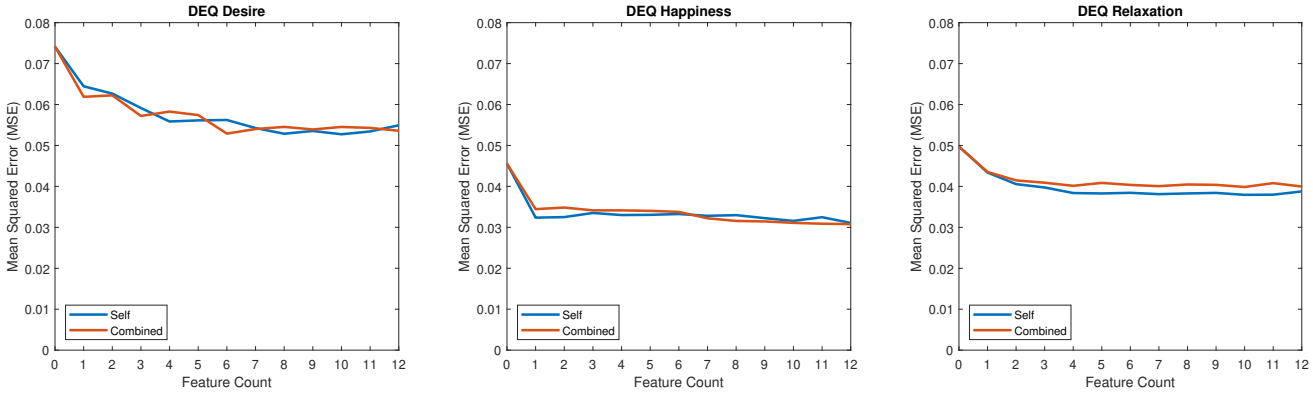

Fig. 5. The average performance (MSE) of models utilising facial expression features from the targeted player, with (Combined) and without (Self) features describing mean expression features for the other players.

in terms of averages across the five independent tests, and six folds of cross validation, using a Friedman test for each target emotion. Results indicate that the MSE of models trained on the four series of data do deviate significantly from one another for each emotion, with follow-up pairwise testing using Wilcoxon signed-rank tests showing significant difference between all pairs of groups, other than self and combined models in prediction of desire and happiness, as can be seen in Table II.

TABLE II
PAIRWISE WILCOXON SIGNED-RANK TEST RESULTS (p) FOR EACH PAIR OF MODEL GROUPS

Feature Group 1	Feature Group 2	DEQ Emotion Desire	Happiness	Relaxation
Self	Group	**0.0009***	**0.0005***	**0.0005***
Self	Paired	**0.0002***	**0.0005***	**0.0002***
Self	Combined	0.8501	0.6221	**0.0005***
Group	Paired	**0.0002***	**0.0005***	**0.0002***
Group	Combined	**0.0015***	**0.0005***	**0.0005***
Paired	Combined	**0.0002***	**0.0005***	**0.0002***

* Significant at $\alpha = 0.05$

Further following the methodology applied by Mavromoustakos-Blom et al., [18], we consider the difference between models trained and tested on the same series of data, conducting further pairwise Wilcoxon signed-rank tests between the four feature groups at the per-fold level. For this we describe only general trends in results for brevity. Results again describe various significant differences between the four feature groups, but as expected, show some inconsistency between folds. Interestingly, model performances for the second and third folds of data more commonly show the combined dataset outperforming models trained on self data alone. The best example of this comes from pairwise testing for the third fold of predictors of desire, which describes similarity between the self and paired data (which otherwise under-performs at all other folds), and similarity between the group and combined models, both of which significantly outperform models using the self feature group, averaging greater performances across the five independent tests at every feature count. This is seen again in prediction of happiness, in which predictors for the second and third group's responses again show the greatest accuracy when utilising the group and combined feature groups, over the target player's own data. This is described in Figure 6, in which model performances for each fold of the data are shown, averaging across the five independent tests predicting happiness.

Figure 6 also describes improvement over self models using the combined dataset for the sixth group, again suggesting some benefit in the use of inter-player data in a minority of cases. The difference between predictive performance at each fold however highlights the greater impact of target data distribution on prediction accuracy, over the use of each feature group. Predictors for the third fold of DEQ happiness for example, routinely attained lower MSE than those predicting responses in fold five, no matter the training group.

Confirmation of the observed trends is further provided following the analytical methodology described by Mavromoustakos-Blom et al., [18], considering the five independent tests' performance data from models using each feature group at the per-fold per-feature count level, via two-tailed Mann Whitney-U tests. Results from these tests again show difference in the direction of effect in significantly different model performances when considering fold three and the other folds in various cases. In almost all instances, when predicting desire, group and combined models significantly outperform paired data models, with this being true for models trained on self data in almost all cases other than those from fold three. Similarly, in a smaller series of cases, self models significantly outperform group models when predicting desire, although this is inverted consistently for fold three, as it is when also comparing the self and combined models.

Considering the differences shown between the third group of participants and the rest of the groups, further exploration of the entire dataset collected suggests that the utility of group data is greatly impacted by how well participants knew each other. Whilst the majority of participants responded to how well they knew the other members of their group with "Not at all", every participant in group three responded with "Very well". Implication that how well participants knew each other effected the utility of inter-player data is not unexpected, with previous work suggesting this may be the case [22]. Further work is required to substantiate this further, allowing for a

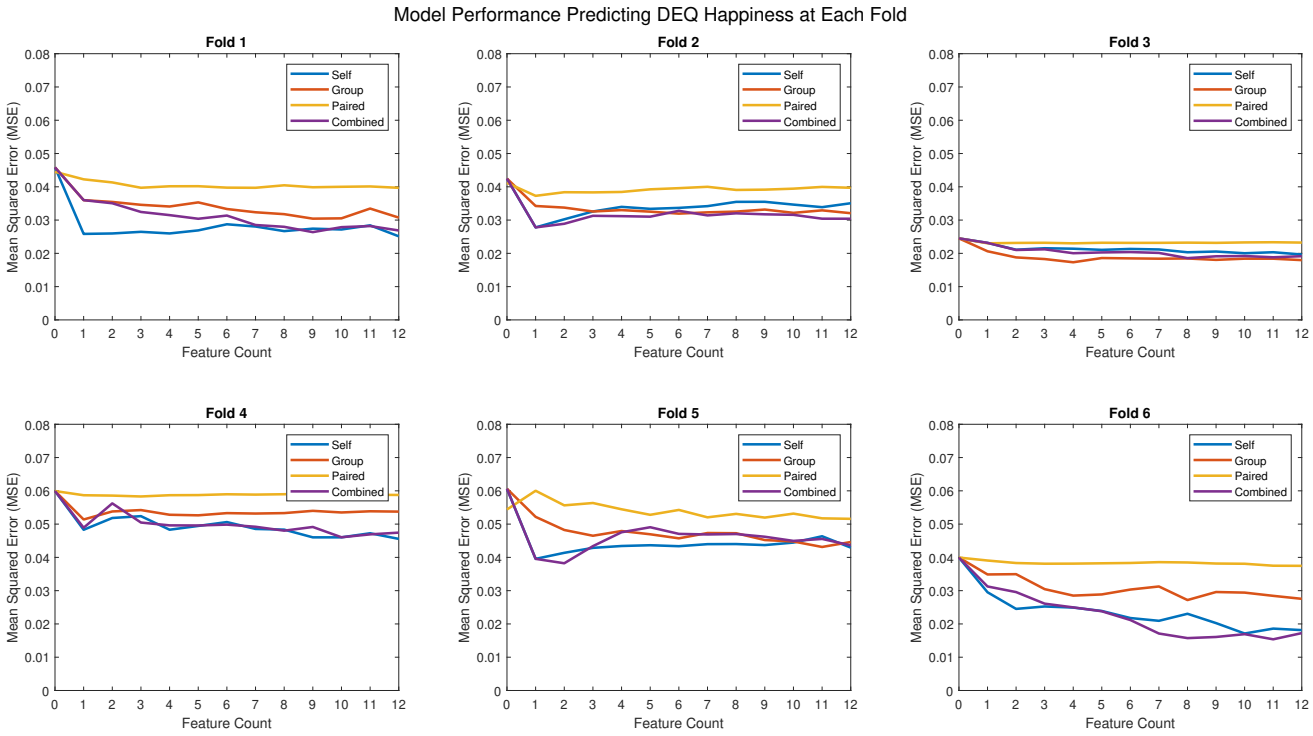

Fig. 6. The average performance (MSE) of models utilising facial expression features from the targeted player, with (Combined) and without (Self) features describing mean expression features for the other players, in the prediction of DEQ Happiness, at each fold (group of players).

greater sample of participants that know each other, for direct comparison to those who do not.

V. Conclusions

In response to RQ1 (To what extent can self-report measures of affect for one player be predicted using measures of affect collected from other players during a shared gaming experience?), this study has shown that inter-player facial expression data used in the prediction of uncaptured emotional intensities for desire, happiness and relaxation (as measured using the DEQ) does not adequately do so, to the same extent to which a player's own facial expression data can. Of the two inter-player feature groups tested, facial expression metrics averaged across all ancillary players significantly outperforms those collected individually, with dynamics between individually paired players expected to add noise to this dataset. Predictive power seen in models using group data does outperform baseline predictions and those using data from the paired dataset however, suggesting that this method does provide some potential in the prediction of uncaptured experience.

In response to RQ2 (To what extent can predictions of self-reported experience using one player's data be improved through the addition of data collected from others who shared the gaming experience?), this study has then shown that the consideration of both a player's own and the average of other players' facial expression metrics during feature selection often has little effect on the performance of models predicting the intensity of the target emotions of desire, happiness and relaxation. Comparison between the performances of models using these feature groups, averaged across independent tests, per fold, or per feature count, resulted in few significant differences being found, with those found often suggesting that the additional group data was detrimental to the generalisability of models. Inverted direction of effect seen in significant differences for the third fold of cross validation suggest that the third group of participants differed significantly from the other groups, in that models making use of inter-player data significantly outperformed those using even the participant's own facial expressions. Consideration of further data collected from participants suggests that this may be due to the greater pre-existing relationship between participants from this group, whilst further work is required to confirm this hypothesis.

Limitations of the study relate to sample size, and use of a single form of emotional stimulus and detection. It is expected that game genre and multiplayer setup play large roles in the findings of this study. With this in-mind, we suggest various avenues for future work, such as the use of individual facial AUs and their relatedness to other players' experiences; research into team-based games and the relationship between specifically cooperative and competitive player experiences; and further work making use of any of the many methods previously shown capable of capturing measures of affect and gameplay experience, such as wearable technology or audio capture, in a multiplayer setting. Further work exploring the use of different genres and multiplayer settings may also show more promise for the elicitation of emotions that were not explored as a part of this study, whilst a greater focus on the use of laboratory testing conditions may alleviate noise created through use of participants' own hardware.

A greater sample of participants, or samples from varying

demographics may also highlight cultural differences in the applicability of the methods utilised, with previous work on the universality of facial expressions suggesting this may be the case [41].

Further work may also wish to consider the ethical implications of inter-player data. This study has shown that modelling player responses using inter-player data may have merit in some cases, but implementation in a real-world setting should ensure that players being profiled through the interpretation of inter-player data based models, are still able to withdraw from profiling, as they would under conventional means.

REFERENCES

[1] S. C. Bakkes, P. H. Spronck, and G. van Lankveld, "Player behavioural modelling for video games," *Entertainment Computing*, vol. 3, no. 3, pp. 71–79, 2012.

[2] C. Pedersen, J. Togelius, and G. N. Yannakakis, "Modeling player experience for content creation," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 2, no. 1, pp. 54–67, 2010.

[3] P. Lopes, A. Liapis, and G. N. Yannakakis, "Sonancia: Sonification of procedurally generated game levels." ICCC, 2015.

[4] D. Thue, V. Bulitko, and M. Spetch, "Player modeling for interactive storytelling: a practical approach," *AI Game Programming Wisdom*, vol. 4, pp. 633–646, 2008.

[5] C. Holmgård, A. Liapis, J. Togelius, and G. N. Yannakakis, "Evolving personas for player decision modeling," in *2014 IEEE Conference on Computational Intelligence and Games*. IEEE, 2014, pp. 1–8.

[6] C. T. Tan, S. Bakkes, and Y. Pisan, "Inferring player experiences using facial expressions analysis," in *Proceedings of the 2014 Conference on Interactive Entertainment*, 2014, pp. 1–8.

[7] J. Zhu and S. Ontañón, "Experience management in multi-player games," in *2019 IEEE Conference on Games (CoG)*. IEEE, 2019, pp. 1–6.

[8] H. Xie, "A generic data representation for predicting player behaviours," Ph.D. dissertation, University of York, 2017.

[9] D. Gábana Arellano, L. Tokarchuk, and H. Gunes, "Measuring affective, physiological and behavioural differences in solo, competitive and collaborative games," in *Intelligent Technologies for Interactive Entertainment: 8th International Conference, INTETAIN 2016, Utrecht, The Netherlands, June 28–30, 2016, Revised Selected Papers*. Springer, 2017, pp. 184–193.

[10] P. Mavromoustakos-Blom, M. Kosa, S. Bakkes, and P. Spronck, "Correlating facial expressions and subjective player experiences in competitive hearthstone," in *Proceedings of the 16th International Conference on the Foundations of Digital Games*, 2021, pp. 1–5.

[11] L. Cardamone, D. Loiacono, and P. L. Lanzi, "Interactive evolution for the procedural generation of tracks in a high-end racing game," in *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, 2011, pp. 395–402.

[12] P. T. Ølsted, B. Ma, and S. Risi, "Interactive evolution of levels for a competitive multiplayer fps," in *2015 IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2015, pp. 1527–1534.

[13] S. K. D'mello and J. Kory, "A review and meta-analysis of multimodal affect detection systems," *ACM computing surveys (CSUR)*, vol. 47, no. 3, pp. 1–36, 2015.

[14] A. Nijholt and R. Poppe, "Facial and bodily expressions for control and adaptation of games (ecag'11)," in *2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG)*. IEEE, 2011, pp. 765–765.

[15] L. Müller, A. Bernin, A. Kamenz, S. Ghose, K. von Luck, C. Grecos, Q. Wang, and F. Vogt, "Emotional journey for an emotion provoking cycling exergame," in *2017 IEEE 4th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. IEEE, 2017, pp. 104–108.

[16] N. Shaker, S. Asteriadis, G. N. Yannakakis, and K. Karpouzis, "Fusing visual and behavioral cues for modeling user experience in games," *IEEE transactions on cybernetics*, vol. 43, no. 6, pp. 1519–1531, 2013.

[17] E. S. Siqueira, M. C. Fleury, M. V. Lamar, A. Drachen, C. D. Castanho, and R. P. Jacobi, "An automated approach to estimate player experience in game events from psychophysiological data," *Multimedia Tools and Applications*, vol. 82, no. 13, pp. 19 189–19 220, 2023.

[18] P. Mavromoustakos-Blom, D. Melhart, A. Liapis, G. N. Yannakakis, S. Bakkes, and P. Spronck, "Multiplayer tension in the wild: A hearthstone case," in *Proceedings of the 18th International Conference on the Foundations of Digital Games*, 2023, pp. 1–9.

[19] J. Quan, Y. Miyake, and T. Nozawa, "Incorporating interpersonal synchronization features for automatic emotion recognition from visual and audio data during communication," *Sensors*, vol. 21, no. 16, p. 5317, 2021.

[20] S. Mariooryad and C. Busso, "Exploring cross-modality affective reactions for audiovisual emotion recognition," *IEEE Transactions on affective computing*, vol. 4, no. 2, pp. 183–196, 2013.

[21] C. Von Scheve and S. Ismer, "Towards a theory of collective emotions," *Emotion review*, vol. 5, no. 4, pp. 406–413, 2013.

[22] M. Wróbel, "I can see that you're happy but you're not my friend: Relationship closeness and affect contagion," *Journal of Social and Personal Relationships*, vol. 35, no. 10, pp. 1301–1318, 2018. [Online]. Available: https://doi.org/10.1177/0265407517710820

[23] G. Pereira, J. Dimas, R. Prada, P. A. Santos, and A. Paiva, "A game prototype with emotional contagion," in *Affective Computing and Intelligent Interaction*, S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 315–316.

[24] J. Z. Huayan Shang, Panpan Feng and H. Chu, "Calm or panic? a game-based method of emotion contagion for crowd evacuation," *Transportmetrica A: Transport Science*, vol. 19, no. 1, p. 1995529, 2023. [Online]. Available: https://doi.org/10.1080/23249935.2021.1995529

[25] A. Dhall, A. Kaur, R. Goecke, and T. Gedeon, "Emotiw 2018: Audio-video, student engagement and group-level affect prediction," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 653–656.

[26] P. Bota, T. Zhang, A. El Ali, A. Fred, H. P. da Silva, and P. Cesar, "Group synchrony for emotion recognition using physiological signals," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 2614–2625, 2023.

[27] G. N. Yannakakis and A. Paiva, "Emotion in games," *Handbook on affective computing*, vol. 2014, pp. 459–471, 2014.

[28] W. A. IJsselsteijn, Y. A. De Kort, and K. Poels, "The game experience questionnaire," 2013.

[29] C. T. Tan, D. Rosser, S. Bakkes, and Y. Pisan, "A feasibility study in using facial expressions analysis to evaluate player experiences," in *Proceedings of The 8th Australasian Conference on Interactive Entertainment: Playing the System*, 2012, pp. 1–10.

[30] E. Hatfield, "Primitive emotional contagion," *Review of personality and social psychology: Emotion and social behavior/Sage*, 1992.

[31] C. Harmon-Jones, B. Bastian, and E. Harmon-Jones, "The discrete emotions questionnaire: A new tool for measuring state self-reported emotions," *PLOS ONE*, vol. 11, no. 8, pp. 1–25, 08 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0159915

[32] D. Bonk and J. Kim, "Factorial and construct validity of the discrete emotions questionnaire for videogames (deq-vg)," *Entertainment Computing*, vol. 42, pp. 100 488, 8, 2022.

[33] A. Brooke, M. Crossley, H. Lloyd, and S. Cunningham, "Towards predicting player experience as discrete emotion intensity using gameplay and visual data in a multiplayer game," 2024. [Online]. Available: https://doi.org/10.36227/techrxiv.172651567.77028268/v1

[34] Unity Technologies, "Unity," 2005.

[35] T. Baltrušaitis, P. Robinson, and L.-P. Morency, "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[36] S. Roohi, J. Takatalo, J. M. Kivikangas, and P. Hämäläinen, "Neural network based facial expression analysis of gameevents: a cautionary tale," in *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play*, 2018, pp. 429–437.

[37] M. Doyran, A. Schimmel, P. Baki, K. Ergin, B. Türkmen, A. A. Salah, S. C. Bakkes, H. Kaya, R. Poppe, and A. A. Salah, "Mumbai: multi-person, multimodal board game affect and interaction analysis dataset," *Journal on Multimodal User Interfaces*, vol. 15, no. 4, pp. 373–391, 2021.

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

[39] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the national academy of sciences*, vol. 111, no. 15, 2014.

[40] J. Nikopoulos, "The stability of laughter," *Humor*, vol. 30, no. 1, pp. 1–21, 2017.

[41] R. E. Jack, R. Caldara, and P. G. Schyns, "Internal representations reveal cultural diversity in expectations of facial expressions of emotion." *Journal of Experimental Psychology: General*, vol. 141, no. 1, p. 19, 2012.