# Modelling Player Preferences in AR Mobile Games

Vivek R. Warriar, John R. Woodward and Laurissa Tokarchuk
*School of Electronic Engineering and Computer Science*
*Queen Mary University of London*
London, UK
{v.r.warriar, j.woodward, laurissa.tokarchuk}@qmul.ac.uk

***Abstract*—In this paper, we use preference learning techniques to model players' emotional preferences in an AR mobile game. This exploratory study uses player behaviour to make these preference predictions. The described techniques successfully predict players' frustration and challenge levels with high accuracy while all other preferences tested (boredom, excitement and fun) perform better than random chance. This paper describes the AR treasure hunt game we developed, the user study conducted to collect player preference data, analysis performed, and preference learning techniques applied to model this data. This work is motivated to personalize players' experiences by using these computational models to optimize content creation and game balancing systems in these environments. The generality of our technique, limitations, and usability as a tool for personalization of AR mobile games is discussed.**

***Index Terms*—Augmented reality, mobile games, player preference modelling, content creation, player experience, preference learning, linear classifiers.**

## I. Introduction

Augmented Reality (AR) experiences have grown in popularity recently. The most popular medium for AR games are mobile devices, possibly due to the increasing simplicity of building and deploying mobile AR experiences. Popular AR SDKs such as ARKit and ARCore use the camera and inertial sensors to extract the device's position and orientation [1]. This information is then used to overlay digital content in the space around the device, which can be viewed through the device. This work focuses on AR games that involve players physical exploration of their local space. We focus on physical AR games for 2 reasons. First, they follow trends in mobile AR games that are more relatable to past experiences players have had, such as playing *Pokemon GO* (PoGo), *Ingress*[1] or *dARk*[2]. These games use narrative and game design to engage people with their surroundings (albeit in a more complex manner than our study game). Second, the potential health benefits of physical AR games make a strong case to personalize these environments to promote healthy behaviour among players. We use player movement, which is the time-series data of the device's position and rotation during the game. In turn, this data is treated as player behaviour in our work.

[1]PoGo and Ingress are both location-based AR games developed by Niantic

[2]dARk is a short story horror experience developed by Combo studio

978-1-7281-1884-0/19/$31.00 ©2019 IEEE

Following similar studies in player modelling [2], [3], we use player behaviour features (PBFs) and controllable game features (CGFs) to predict a player's emotional preferences. For example, does the player find level A more fun than level B or vice versa? Effective prediction of such preferences will enable procedural content generation (PCG) or game balancing systems to be optimized to a player's ideal emotional preferences. Our approach is novel, in that it uses movement data to model players' preferences. Ground truth is established through data collected from self-reported questionnaires. Exploring this domain in the context of subjective preferences is useful for personalizing game experiences.

Since data from popular AR mobile games is unavailable, we developed an AR Treasure-hunt game that is similar to existing games by incorporating reward systems, exploration of local space, and a narrative that motivates these rewards (described in section III). Before introducing the game, a background of relevant work is provided in the next section. We conducted a user study to collect player behaviour and preference data, described in section IV. We then perform inferential statistics and machine learning on PBFs and CGFs in order to understand the relationship between these features and players' preferences (methodological details are described in section V and the results from the analysis are described in section VI). As shown in [2], [3], preference models can be learned using different computational methods. In this paper, we show that a combination of CGFs and PBFs can be used to accurately predict a number of dimensions of emotional preferences. For AR mobile games, we propose a different model obtained with Support Vector Machines (SVM) which shows better performance than others tested. We also make feature recommendations from the set of PBFs and CGFs tested. Finally, the implications of these results, limitations, and future work are discussed in section VII.

## II. Background

Research in AR games has explored techniques to enhance PX, through PCG [4] and automatic game balancing [5], [6]. Despite this, we have observed a lack of studies that model PX in AR. Exploring this area will complement existing research to enhance personalization. Theoretical frameworks such as Experience-Driven Procedural Content Generation [7] offer directions to personalize game content. These approaches are driven by computational models of PX. This has been suc-

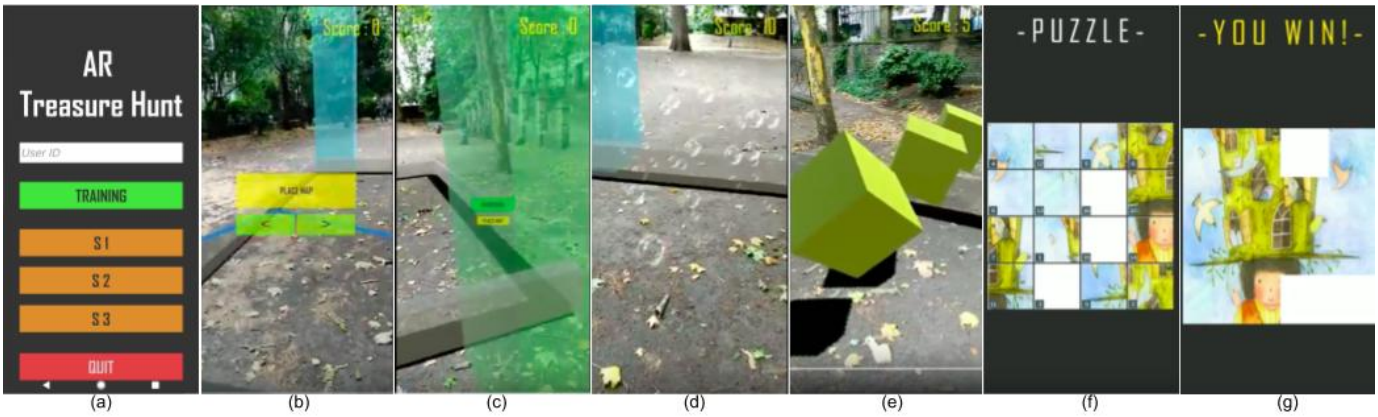


Fig. 1. Fig. [a-g] show the flow of a single round of the game. [a]: Shows the screen to select an experiment session. [b]: The options for a user to place a game map. [c]: The (green) start button that the player must tap in order to begin the game. [d]: The bubbles in the game indicating treasure in close by. [e]: Treasure that appears which the player collects. [f]: The 2D puzzle presented to the player in the unsolved form. The white squares indicate treasure pieces that were not collected. This screen is presented to the player once the exit area is entered (seen in blue in fig [b] and [d]). [g]: The solved 2D puzzle.

cessfully applied to traditional digital games [2] and physical interfaces [5]. AR mobile experiences are 'hand-held video see-through experiences' [1] that offer a variety of sensor data such as time, location, and movement. This is a more complex environment for the creation of computational models of PX. We use player movement to predict preferences. Related work has been conducted in [8] and [9] in detecting emotions from movement data of people playing Nintendo Wii games. In [9], a different approach to predicting affective states[3] is followed. Our approach is similar to [8], which predicts distinct emotional states (Triumph, Concentration, Defeat and Frustration were the states explored in the study). The 2 studies are similar, as ground truth for predictions is multiple observer agreement. Our approach differs: ground truth is established from self-reported questionnaire data from players.

## III. AR Treasure-hunt: Test-bed Platform Game

The aim of the game is to collect all hidden treasure from a constrained space. This treasure has been randomly distributed within the level area. The treasure pieces are parts of a picture. In order to win, the player will have to put together this picture (like a 2D puzzle). Figure 1 shows the screens during a single round of the game. Levels of the game vary across the number of treasure pieces and the size of the game area. This game can be played in parks and other open spaces.

The player places the AR level in the world before starting the game. The game begins once the player finalizes this placement. While a player is exploring the level, the boundary of the AR level and an exit (to the 2D puzzle) is visible to them. They will not be aware of how many treasure pieces are hidden in the level. These pieces are invisible to the player by default. If a player is close to a treasure piece they will receive an audio-visual clue. The clues are implemented using a particle system that is designed to look like a bubble emitter with an appropriate sound effect. Bubbles are emitted around the position of the treasure piece. Treasure appears only if the player is in close proximity to it. Once it appears the player can collect the item by moving the mobile device into it. Collecting treasure increases the players score by +1. The player is instructed to explore the level until they believe they have discovered all the hidden treasure and then move to the exit. Only when the player moves into the exit square, they are shown the 2D puzzle (in the shuffled order) and all the treasure pieces that were not collected appear to them as white squares. This is when the player will get confirmation if all pieces were collected or not. This design decision ensures that players are motivated by the exploration and discovery of dynamic content within the game space. Informing the player beforehand of the maximum number of pieces hidden in that level would reduce the sense of exploration and discovery, which is an important aspect of these real-world games.

The player wins the round once the puzzle is completed. The game has been designed in such a way that all game levels can be completed. However, the reward for the player varies depending on the amount of treasure they collect, which corresponds to the amount of the picture they get to appreciate at the end of the game round. The game interactions were designed to be simple so as to keep the cognitive load from the UI on the players low. The game was developed in Unity, using their experimental AR interface to handle the device and environment tracking for the game. The Unity asset store was used for game assets used in the game. The mobile device used for development and testing was the Google Pixel 2 XL.

## IV. Experimental Settings

The study design was informed by previous studies that model PX for content creation [2], [3], [5]. With this protocol, we build a data-set of player movement data and corresponding emotional preferences in AR game sessions. 42 volunteers (17 female and 25 male) aged 18-44 (51% were 18-24, 21% were 25-29, 15% were 30-34, 10% were 35-39, 3% were 40-44) took part in this study. When asked about prior experience

[3]Affective states is an approach to measuring and contextualizing emotions according to some dimensional space, usually: Valence and Arousal

playing AR games 45% of subjects had no prior experience. In the remaining 55% of subjects: 29% reported having only one experience in the past, 24% played a few times before, and 2% of participants played AR games regularly.

### A. Experimental Protocol

The study consisted of a number of sessions of the same format. In each session, participants played 2 rounds of the game with different CGFs in each round (resulting in varying levels that created a spectrum of emotional responses from players). Participants were not given any constraints on how to hold the phone (in portrait or landscape). They could use either hand to hold the phone depending on comfort. Most participants (except for 2) preferred portrait mode and tended to prefer their dominant hand. As we are interested in modelling a player's emotional preference, pilot studies were conducted to identify appropriate game features that could create a diverse range of emotional responses from players. The 2 chosen CGFs were:

- *The Area of the Level ($G_A$):* 2 sizes of levels are compared. *Large Area (LA)* levels are $\approx 30m \times 30m$ and *Small Area (SA)* levels are $\approx 5m \times 5m$. $G_A \in \{LA, SA\}$
- *Treasure in Level ($G_T$):* 2 amounts of treasure are compared: Low Treasure (LT) with 9 and High Treasure (HT) with 16 pieces respectively. $G_T \in \{LT, HT\}$

Using 2 CGFs has resulted in $4(2 \times 2)$ levels being compared: (1) $LA \times HT$ (2) $SA \times HT$ (3) $LA \times LT$ (4) $SA \times LT$. Since games are played in pairs, the total number of game pair combinations is 6. In this study we have focused testing on 3(out of 6) of the game pairs:

(1) $LA \times HT$ vs (4) $SA \times LT$
(2) $SA \times HT$ vs (3) $LA \times LT$
(1) $LA \times HT$ vs (3) $LA \times LT$

We do not explore the complete comparison space because real-world optimization for player preferences would rely on similar incomplete data-sets. The current choice of 2 binary variables as CGFs is adequate for the purposes of our exploratory study. It would easily become unfeasible to collect pairwise preferences of the complete comparison space if a more complex set of CGFs is used.

At the end of each game pair, the participants were given a 4-AFC protocol. This is a questionnaire that ranks the 2 games according to different dimensions. This study focused on *Boredom, Challenge, Excitement, Frustration* and *Fun* [10] as dimensions to measure player preference; since previous research has shown that these states are relevant to digital game-play. Following shows the 4-AFC protocol measuring the dimension of *Fun*:

Please select 1 of the following options

1) Game 1 felt more *Fun* than Game 2
2) Game 2 felt more *Fun* than Game 1
3) Game 1 and Game 2 felt equally *Fun*
4) Neither of the two games felt *Fun*

The same format is used to measure each dimension of preference. This data is used as ground truth for players' preferences between pairs of games. The study began with a briefing for each participant which included a training session on the game and the structure of each session of the study. This was followed by a trial session in the described format; data from this session is discarded. For the trial session, 2 levels were designed with different areas (*LA, SA*) containing 4 treasure pieces each. The trial allowed participants to familiarize themselves with the game and study format. The trial was followed by 3 experiment sessions; participants were given a minute's break between each. They were then debriefed and the study was concluded. The order of sessions was randomized and the order of game pairs was counterbalanced to minimize ordering effects in the data collected.

### B. Data Collection

During the study, player behaviour and preference data were collected. As each session consisted of comparing 1 game pair, each subject contributed 3 game pairs of preferences resulting in 126 games pairs (252 individual games). However, due to some technical crashes, only 117 game pairs were successfully recorded and used in the data analysis. The following data was collected from each pair:

*a) Player Behaviour Data ($PB$):* measured from player movement in game sessions. The mobile IMU sensors record position and rotation of the device during the game. This data is recorded at a frequency of *64 Hz* following guidelines from Preece et al. [11] and the discrete-time signals are stored as a 6-dimensional vector: $\alpha \in \{P_X, P_Y, P_Z, R_X, R_Y, R_Z\}$ for position and rotation. This sampling frequency is high enough to pick up both large (arm/hand movements, walking) and small movements (hand tremors, body jerks). The player's score(*S*), which increases as the treasure pieces are collected, is recorded at the same frequency.

*b) Emotional Preferences Data ($PF$):* The 4-AFC collects preference data between game pairs along dimensions of *Boredom, Challenge, Excitement, Frustration* and *Fun*.

## V. Methods

The resulting data-set from the study was used to explore preference learning approaches to modelling players' preferences. The data is pre-processed and PBFs are extracted. These features along with the CGFs were used to model players' preferences. In order to better understand the effects of each feature, and to explore to what extent noise from ordering effects have biased the data, we first conduct statistical analysis on the features.

### A. Data pre-processing

Pre-processing reduces noise and redundancy in the data. This stage is adapted from [12] and is broken into these steps: Data Segmentation, Low Pass Filtering, Coordinate Difference, and Dimensionality Reduction.

*a) Data Segmentation:* As we are only interested in game behaviour, movement data from when players were interacting with the mobile device before gameplay (e.g., while confirming the placement of the AR level in the physical space) was discarded.

*b) Low Pass Filtering:* The segmented data may be noisy and contain unwanted high-frequency components. In order to reduce this, we use a Gaussian filter with coefficients from [12]: $h = \frac{1}{16}[1, 4, 6, 4, 1]$. The filter is a 1D convolution of the Gaussian filter and each column of the raw data in $\alpha$ (details in section IV) given by the following equation:

$$y(n) = \sum_{t=-\infty}^{\infty} x(t)h(n-t) = x(n) * h(n) \tag{1}$$

In eq 1, *x* is a column of the vector $\alpha$ (raw data), *h* is the Gaussian filter and $*$ the convolution operation.

*c) Coordinate Difference:* Movement qualities such as velocity($\dot{\alpha}$), acceleration ($\ddot{\alpha}$) and jerk[4] ($\dddot{\alpha}$) are extracted for each of the columns of the vector $\alpha$ [13].

$$\dot{\alpha}(t) = x(t) - x(t-1) \tag{2}$$

$$\ddot{\alpha}(t) = \dot{\alpha}(t) - \dot{\alpha}(t-1) \tag{3}$$

$$\dddot{\alpha}(t) = \ddot{\alpha}(t) - \ddot{\alpha}(t-1) \tag{4}$$

This step outputs 3 6-D vectors for velocity ($\dot{\alpha}$), acceleration ($\ddot{\alpha}$) and jerk ($\dddot{\alpha}$). These 3 vectors contain data for velocity, acceleration, jerk, angular velocity, angular acceleration and angular jerk along the x, y and z axis. Analyzing the quality of movement in this way minimizes the impact of inter-participant differences in holding the phone on the preference predictions.

*d) Dimensionality Reduction:* In order to reduce the dimensionality of the feature space, we use the Euclidean norm of the x, y and z axis for velocity, acceleration, jerk, angular velocity, angular acceleration and angular jerk. This output 6-D vector along with the score is the final output of the data pre-processing phase for each game, given by $\beta \in \{V, A, J, RV, RA, RJ, S\}$ at $64Hz$.

### B. Feature Extraction

PBFs are extracted from the pre-processed movement data: $V, A, J, RV, RA, RJ$ (first 6 dimensions of the time series vector $\beta$), table I shows the 10 features that are extracted for each dimension resulting in 60 movement features. The $S$ signal (last dimension of $\beta$) is used to compute 2 features:

- Completion ($C$): The fraction of the score at the end of game divided by maximum possible score from game, given by: $C = \frac{Score\ at\ end\ Game}{Max\ Game\ score}$
- Score Rate ($SR$): The fraction of the score at the end of the game and the time taken to reach it (note: this is different from total time of the game), given by: $SR = \frac{Score\ at\ end\ Game}{Time\ to\ reach\ score}$

The final feature considered is the time of the game in seconds ($T$), resulting in 63 PBFs. These, along with CGFs (65 features in total) were used in the following analysis.

[4]Jerk is the derivative of acceleration

TABLE I
EXTRACTED MOVEMENT FEATURES

| **Feature** $X$ | Description |
|---|---|
| $X_m$ | Mean |
| $X_{std}$ | Standard Deviation |
| $X_{sk}$ | Skew |
| $X_{kur}$ | Kurtosis |
| $X_{min}$ | Minimum value |
| $X_{max}$ | Maximum value |
| $X_D$ | Max - Min |
| $X_{tMin}$ | Time of Minimum value |
| $X_{tMax}$ | Time of Maximum value |
| $X_{tD}$ | Time of Max - Time of Min |

### C. Statistical Analysis

Statistical analysis was conducted to check for ordering effects in the data and to understand the relationship between features (PBFs, CGFs) and emotional preferences. The Chi-square test is used to check for ordering effects in preference data, which is based on the number of times subjects expressed a preference for the first or the second game in the pair. Chi-square test is also used to check for statistically significant effects of the 2 CGFs on preferences as these are binary categorical features. The Wilcoxon signed-rank test was used to check for significant effects of the PBFs on preferences as these are continuous features. All tests for significant effects use a p-value $< 1\%$.

We followed the method to compute correlation coefficients from [5], given by $c(z) = \sum_{i=1}^{N_s}\{z_i/N_s\}$ where $N_s$ is number of pairs where subjects expressed clear preferences for one of the two games (picking the first 2 options of the 4-AFC), and $z_i = 1$ when the subject preferred the game with the larger value of the examined feature, and $z_i = -1$ when the subject chooses the other game. From the 117 game pairs that were analyzed $N_s$ is 59, 105, 89, 81, 106 for *Boredom, Challenge, Excitement, Frustration* and *Fun* respectively. Variance in $N_s$ shows that subjects find it difficult to express a clear emotional preference between game variants.

### D. Preference Learning

Preference learning techniques are applied to explore to what extent PBFs and CGFs can be used to predict players' preferences. We use the large margin algorithm [14], which was originally developed for a driving route recommendation system. This technique has been previously applied in similar studies of modelling PX [2], [3], [5]. We have also applied feature selection techniques to improve model performance.

*a) Large Margin Algorithm:* This method aims to model features of interest through a linear combination of a weighted vector that binds preferences to features. This is given by $P(F) = FW^T$, where $P(F)$ is the subjects preference, F is the extracted feature vector and W is the weights to be optimized. As we are predicting pairwise preference, we would like to predict $P(F_A) > P(F_B)$ if a subject has a preference for Game A over Game B. Here $F_A$ and $F_B$ are PBFs and CGFs extracted from each of the games A and B

respectively. This inequality can be expressed through a linear combination $F_A W^T > F_B W^T$, which is further rewritten as $(F_A - F_B)W^T > 0$ or $F_D W^T > 0$, where $F_D$ is the feature difference vector for the preference. The problem is thus reformulated as a linear classification of estimating $W$, where the input features are the feature difference between the original feature space of the 2 game pairs being compared.

Previous research explores this as a binary classification problem where $F_D W^T \in [0, 1]$: 0 is assigned to an instance where the subject has a preference for Game A over Game B and 1 for the opposite preference. This is accomplished by either filtering the data to remove instances with no preferences [14] or by forcing choice onto subjects via the 2-AFC protocol [3]. We believe that for this exploratory study, it would be beneficial to investigate both binary (via data filtering) and ternary classifications where $F_D W^T \in [0, 1, 2]$: 2 is the class assigned to instances where the subject had no preference (options 3 and 4 in the 4-AFC protocol). We investigate this approach as it matches the format of the data collected without filtering.

As we are interested in optimizing the weight vector W, which can be used to linearly combine the feature space $F_D$ in order to predict preferences, we check the performance of three linear classifiers for this problem: logistic regression, linear discriminant analysis (LDA) and support vector machines (SVM). LDA has been used in a related study [3]. However, it is unclear from previous work if other linear models can outperform this approach. In this study, we evaluate model performance using the sample accuracy and standard deviation from 10-fold cross-validation (CV).

*b) Feature Selection:* Previous studies observe that model performance improves through feature selection techniques [2], [3]. While there are a large number of approaches, we use sequential forward selection (SFS) and sequential floating forward selection (SFFS) in this study as they are often used in similar work [2], [3]. In [2], SFS and SFFS outperformed other techniques tested. SFS is a bottom-up search algorithm that tries to find the best performing feature set. It starts with the best performing single feature and adds new features from the remaining set such that model performance of the new set generates the best possible overall performance over other potential features for addition. SFFS is similar to SFS except that when a forward step is performed, the algorithm also checks if a feature from the existing set can be excluded in order to improve overall model performance.

## VI. Results

This section presents results from analyzing each dimension of emotional preference with both statistical analysis and preference learning techniques. This section concludes with feature recommendations based on the results.

### A. Statistical Analysis

This subsection describes results from order testing (to check if the ordering of the game has created noise in the preferences) and the correlation analysis of statistically significant features.

TABLE II
Statistically Significant (p-value $< 1\%$) Correlation Coefficients for Boredom.

| **Feature** | $c(z)$ |
|---|---|
| Controllable Level Features | |
| Area of level ($G_A$) | 0.339 |
| Treasure in level ($G_T$) | -0.322 |
| Player Behaviour Features | |
| Acceleration Mean ($A_m$) | 0.390 |
| Acceleration Standard Deviation ($A_{std}$) | 0.390 |
| Maximum Acceleration ($A_{max}$) | 0.458 |
| Max-Min Acceleration ($A_D$) | 0.458 |
| Jerk Mean ($J_m$) | 0.424 |
| Jerk Standard Deviation ($J_{std}$) | 0.458 |
| Maximum Jerk ($J_{max}$) | 0.424 |
| Max-Min Jerk ($J_D$) | 0.424 |
| Velocity Mean ($V_m$) | 0.390 |
| Time ($T$) | 0.390 |

TABLE III
Statistically Significant CGFs and Top ten Statistically significant (p-value $< 1\%$) PBFs Correlation Coefficients for Challenge.

| **Feature** | $c(z)$ |
|---|---|
| Controllable Level Features | |
| Area of level ($G_A$) | 0.686 |
| Player Behaviour Features | |
| Time of Max Acceleration ($A_{tMax}$) | 0.467 |
| Time of Max Velocity ($V_{tMax}$) | 0.476 |
| Time of Max Ang. Acceleration ($RA_{tMax}$) | 0.504 |
| Time of Max Ang. Jerk ($RJ_{tMax}$) | 0.523 |
| Maximum Ang. Velocity ($RV_{max}$) | 0.467 |
| Max-Min Ang. Velocity ($RV_D$) | 0.467 |
| Time of Min Ang. Velocity ($RV_{tMin}$) | 0.467 |
| Time of Max Ang. Velocity ($RV_{tMax}$) | 0.504 |
| Score Rate ($SR$) | -0.714 |
| Time ($T$) | 0.771 |

*a) Boredom:* Participants had a preference of *Boredom* $50.43\%$ of the time. Order testing showed a significant ($p = 0.001$) effect: subjects tended to find the second game more boring. Table II shows the significant PBFs and CGFs.

*b) Challenge:* Participants had a preference $89.74\%$ of the time. Order testing was not significant. Statistical testing of the features showed that area of the level was the only CGF that had a significant effect (details provided in table III) while 39 PBFs showed a significant effect. We report only the top ten correlation coefficients in table III.

*c) Excitement:* Participants had a preference 76.06% of the time. Order testing was not significant. Among all the features, only *Treasure in Level* $G_T$ (a CGF) had a significant effect with $c(z) = 0.339$.

*d) Frustration:* Participants had a preference 69.23% of the time. Order testing was not significant. Feature tests showed that area of the level was the only CGF that had a significant effect (details in table IV) while 38 PBFs showed a significant effect. We report only the top ten correlation coefficients for *Frustration* in table IV.

*e) Fun:* Participants had a preference 90.60% of the time. Order testing showed a significant ($p = 0.002$) effect,

TABLE IV
STATISTICALLY SIGNIFICANT CGFS AND TOP TEN STATISTICALLY SIGNIFICANT (P-VALUE $< 1\%$) PBFS CORRELATION COEFFICIENTS FOR FRUSTRATION.

| **Feature** | $c(z)$ |
|---|---|
| Controllable Level Features | |
| Area of level ($G_A$) | 0.691 |
| Player Behaviour Features | |
| Maximum Acceleration ($A_{max}$) | 0.481 |
| Max-Min Acceleration ($A_D$) | 0.481 |
| Time of Max Jerk ($J_{tMax}$) | 0.530 |
| Time of Max Velocity ($V_{tMax}$) | 0.444 |
| Time of Max Ang. Acceleration ($RA_{tMax}$) | 0.481 |
| Time of Max Ang. Jerk ($RJ_{tMax}$) | 0.555 |
| Minimum Ang. Velocity ($RV_{min}$) | -0.481 |
| Time of Max Ang. Velocity ($RV_{tMax}$) | 0.456 |
| Score Rate ($SR$) | -0.679 |
| Time ($T$) | 0.802 |

subjects tended to find the first game more fun. Among the CGFs, *Area of level ($G_A$)* had a significant effect with $c(z) = -0.198$. Among the PBFs *Completion ($C$)* had a significant effect with $c(z) = 0.245$.

### B. Machine Learning

In this section, we present results from preference learning techniques. The extracted features are used to predict preferences. A summary of findings is shown in Table V which provides the accuracy and standard deviation from 10-fold CV of the Logistic Regression, LDA and SVM classifiers, to predict the various emotional dimensions of preferences in both binary and ternary classification scenarios. Corresponding accuracies of the best performing feature subset from the feature selection techniques (SFS, SFFS) along with the number of features in the subset have also been provided in Table V. It has been a common observation across all dimensions and types of linear classifiers that the base (all 65 features) performance without feature selection performs poorly with very high standard deviation (as high as $\pm 28.55\%$ for the ternary *Fun* LDA classifier). However, all base classifiers perform higher than random chance (50% for binary and 33.34% for ternary).

*a) Boredom:* The best binary classifier was SVM with a subset of 14 features found with SFFS - the performance was $86.33 \pm 10.2\%$. The best ternary classifier was SVM with a subset of 21 features found with SFS - the performance was $60.29 \pm 13.8\%$.

*b) Challenge:* The best binary classifier was SVM with a subset of 5 features. Both SFS and SFFS found the same feature set - the performance was $93.36 \pm 4.4\%$. The best ternary classifier was SVM as well with the same subset of 5 features (found with both SFS and SFFS) - the performance was $84.85 \pm 5.5\%$.

*c) Excitement:* The best binary classifier was LDA with a subset of 15 features found with SFFS - the performance was $75.61 \pm 12.5\%$. The best ternary classifier was SVM with a subset of 14 features found with SFFS - the performance was $56.92 \pm 10.7\%$.

*d) Frustration:* The best binary classifier was SVM with a subset of 24 features found with SFFS - the performance was $93.89 \pm 6.1\%$. The best ternary classifier was SVM as well with a subset of 11 features found with SFFS - the performance was $75.43 \pm 8.3\%$.

*e) Fun:* The best binary classifier was SVM with a subset of 16 features found with SFFS - the performance was $79.21 \pm 9.7\%$. The best ternary classifier was SVM with a subset of 10 features found with SFFS - the performance was $69.42 \pm 7.0\%$.

### C. Feature Recommendations

In this subsection, we present a list of feature combinations that can be used as a starting point in similar work to predict player preferences. The recommendations are based on a grounded analysis of features in terms of statistical effects and likelihood of each being selected in the best performing feature subset. We have observed 2 common features in predicting all the emotion dimensions, given by: $\Theta \in \{A_{min}, RV_{min}\}$. A number of feature sets were also found that could predict 4(/5) dimensions. A set of 4 features can be used to predict *Boredom, Challenge, Excitement and Frustration* (not *Fun*), given by: $\Lambda \in \{A_m, A_{std}, A_{max}, V_m\}$. A single feature can be used to predict *Boredom, Challenge, Frustration, Fun* (not *Excitement*), given by: $\Pi \in \{G_A\}$. Similarly, feature sets emerge that can predict 3(/5) dimensions. A single feature was found to be able to predict *Challenge, Frustration, Fun*, given by $\Phi \in \{C\}$. A single feature can be used to predict *Boredom, Challenge, Frustration*, give by: $\Psi \in \{J_m\}$. Common feature sets also emerge that can predict pairs of emotions. A large set of features could predict *Challenge and Frustration*, given by: $\Omega \in \{T, A_D, J_{std}, RA_{tMin}, RJ_{tMin}, RV_{max}\}$. Another pair of features could predict *Boredom, Excitement*, given by: $\Delta \in \{G_T, RJ_{tD}\}$. Feature recommendations for each dimension of preference are given by the composition of the sets of features presented above, illustrated in fig 2. For instance, *Challenge* is a set of 15 features, given by: $\mathbb{CH} \in \{\Theta, \Lambda, \Pi, \Psi, \Phi, \Omega\}$. Showing the recommendations as compositions of other feature sets allows us to appreciate important relationships across emotions. For instance, a total overlap in the features that predict *Challenge* and *Frustration* is observed (implications are discussed in the next section).

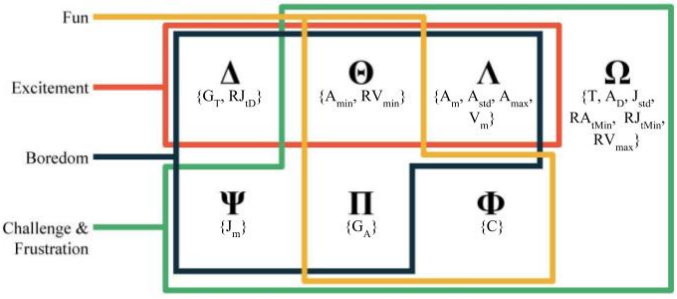


Fig. 2. The figure shows how recommended feature sets for the various dimensions of emotional preference can be expressed as compositions of important feature sets.

TABLE V

SUMMARY OF RESULTS FROM THE PREFERENCE LEARNING TECHNIQUES INCLUDING FEATURE SELECTION FOR BOTH BINARY AND TERNARY SCENARIOS. FOR EACH CLASSIFIER, THE NUMBER OF FEATURES USED, THE SAMPLE ACCURACY AND STANDARD DEVIATION FROM 10-FOLD CV ARE SHOWN. THE BEST PERFORMING BINARY AND TERNARY CLASSIFIER FOR EACH EMOTION HAS BEEN HIGHLIGHTED.

| | | | **Log. Reg.** | | | **LDA** | | | **SVM** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | $F_\#$ | Acc | $\pm SD$ | $F_\#$ | Acc | $\pm SD$ | $F_\#$ | Acc | $\pm SD$ |
| Boredom | Binary | All | 65 | 64.33% | 15.8% | 65 | 50.67% | 28.6% | 65 | 78.00% | 12.9% |
| | $N_s = 59$ | SFS | 28 | 79.67% | 9.9% | 27 | 80.00% | 12.5% | 15 | 84.67% | 9.1% |
| | | SFFS | 29 | 79.33% | 15.3% | 40 | 84.67% | 11.8% | 14 | 86.33% | 10.2% |
| | Ternary | All | 65 | 45.71% | 17.0% | 65 | 37.58% | 12.9% | 65 | 47.64% | 8.0% |
| | $N_s = 117$ | SFS | 3 | 57.37% | 14.7% | 6 | 58.21% | 13.0% | 21 | 60.29% | 13.8% |
| | | SFFS | 36 | 57.45% | 16.3% | 6 | 58.21% | 13.0% | 32 | 58.69% | 10.9% |
| Challenge | Binary | All | 65 | 81.00% | 10.3% | 65 | 74.54% | 13.1% | 65 | 81.10% | 10.0% |
| | $N_s = 105$ | SFS | 15 | 91.54% | 5.0% | 4 | 91.45% | 5.0% | 5 | 93.36% | 4.4% |
| | | SFFS | 3 | 91.45% | 5.0% | 4 | 91.45% | 5.0% | 5 | 93.36% | 4.4% |
| | Ternary | All | 65 | 67.70% | 12.7% | 65 | 62.80% | 14.1% | 65 | 72.86% | 9.1% |
| | $N_s = 117$ | SFS | 2 | 82.20% | 5.2% | 12 | 82.26% | 6.0% | 5 | 84.85% | 5.5% |
| | | SFFS | 2 | 82.20% | 5.2% | 12 | 82.26% | 6.0% | 5 | 84.85% | 5.5% |
| Excitement | Binary | All | 65 | 55.44% | 15.8% | 65 | 57.81% | 20.5% | 65 | 56.39% | 8.2% |
| | $N_s = 89$ | SFS | 12 | 73.64% | 13.6% | 9 | 72.39% | 13.7% | 8 | 68.53% | 9.6% |
| | | SFFS | 9 | 73.64% | 14.5% | 15 | 75.61% | 12.5% | 12 | 71.89% | 10.5% |
| | Ternary | All | 65 | 44.05% | 12.9% | 65 | 36.41% | 11.7% | 65 | 43.01% | 6.9% |
| | $N_s = 117$ | SFS | 31 | 54.99% | 14.1% | 20 | 54.88% | 11.4% | 2 | 55.17% | 14.7% |
| | | SFFS | 15 | 53.05% | 13.3% | 11 | 55.78% | 11.2% | 14 | 56.92% | 10.7% |
| Frustration | Binary | All | 65 | 85.38% | 8.8% | 65 | 68.29% | 14.1% | 65 | 82.60% | 8.4% |
| | $N_s = 81$ | SFS | 1 | 90.41% | 8.9% | 1 | 90.41% | 8.9% | 4 | 92.78% | 5.9% |
| | | SFFS | 1 | 90.41% | 8.9% | 34 | 92.92% | 9.5% | 24 | 93.89% | 6.1% |
| | Ternary | All | 65 | 50.57% | 14.4% | 65 | 45.57% | 20.4% | 65 | 59.64% | 14.6% |
| | $N_s = 117$ | SFS | 11 | 71.83% | 10.6% | 16 | 72.65% | 11.6% | 16 | 72.92% | 8.3% |
| | | SFFS | 13 | 73.57% | 9.8% | 20 | 74.05% | 10.0% | 11 | 75.43% | 8.3% |
| Fun | Binary | All | 65 | 65.62% | 13.9% | 65 | 58.47% | 11.9% | 65 | 60.37% | 5.2% |
| | $N_s = 106$ | SFS | 42 | 73.68% | 8.9% | 35 | 74.33% | 6.5% | 16 | 74.68% | 6.8% |
| | | SFFS | 32 | 76.26% | 8.6% | 34 | 77.59% | 11.4% | 16 | 79.21% | 9.7% |
| | Ternary | All | 65 | 59.56% | 9.2% | 65 | 47.53% | 11.7% | 65 | 54.77% | 5.0% |
| | $N_s = 117$ | SFS | 4 | 65.75% | 7.8% | 1 | 65.67% | 8.1% | 9 | 67.68% | 6.8% |
| | | SFFS | 32 | 65.94% | 8.4% | 18 | 68.99% | 9.0% | 10 | 69.42% | 7.0% |

## VII. DISCUSSION

The results and recommendations made indicate that combinations of CGFs and PBFs can be used to accurately predict dimensions of emotional preferences. We have chosen to compare several linear classifiers and have found SVM classifiers (unexplored in previous research for this problem) were the best performing classifiers for both accuracy and stability, indicated by higher accuracy and lower standard deviation of 10-fold CV accuracies. Although all classifiers perform better than random chance, we find that binary classifiers outperform ternary classifiers in both accuracy and stability.

*Boredom, Excitement* and *Fun* are difficult to model, which is observed in statistical analysis as well as the low accuracy and stability. Predicting *Challenge* and *Frustration* shows higher accuracy and stability. Due to the small data-set, we believe that classifiers for *Boredom, Excitement* and *Fun* show over-fitting due to a high standard deviation of 10-fold CV accuracies. The performance of *Challenge* and *Frustration* shows more acceptable stability with both binary and ternary classifiers showing a variance of $\approx 5\%$. We believe that this variance will be further reduced if a larger data-set is used. Our results are similar to results from other studies that model players' preferences [2]. Although the authors investigate Super Mario in their work, they find that *Fun* and *Boredom* were the most difficult to predict, while *Challenge* and *Frustration* were the best performing classifiers.

We propose that our observations are due to the underlying relationship between game activities and the specific emotional dimension. Some theories [15] consider emotions as being constructed from more fundamental properties called Valence and Arousal: Valence is the amount of goodness or badness in experiences, while Arousal is the psychological state of being awake. *Fun, Excitement,* and *Boredom* are more resonant with the emotional dimensions of valence. *Frustration* and *Challenge* are resonant with the emotional dimension of arousal. *Frustration* is a construct of negative valence and high arousal, while *Challenge* is strongly linked to player performance and high arousal states. As our approach uses movement data, this information medium could be more useful to detect variable arousal rather than variable valence based emotional states. Although it is useful to use valence and arousal to interpret these results, the current approach of asking players about preferences across easily understandable emotions has obvious advantages as it is more intuitive for people to compare 2 experiences based on *Fun* or *Frustration* rather than valence and arousal. Studies about emotions [15] also tell us that people are different in their ability 'to represent their experiences as categorically distinct events' and this

ability is influenced by context and language abilities. This is observed in our study by the variable proportions of clear preference across the emotions tested. For this reason, we also caution that the nature of the game task can bias players' interpretation of the preferential comparison being made.

Our game requires considerably high amounts of walking; we believe that our results are applicable to similar AR games that require movement in local space. The techniques proposed in this paper have considerable potential to create content that is optimized for an ideal balance of *Challenge* and *Frustration* (i.e. the best balance of most challenging and least frustrating). Currently, we explore a simple game design space and it is possible that *Frustration* and *Challenge* are being predicted by the same underlying feature correlations. In this game, most participants appear to prefer a large amount of Treasure (more fun, exciting and less boring) and do not prefer walking a large amount (more challenging and frustrating). In this case, it would be impossible to find a game experience that is both challenging but not frustrating. It would be interesting to see how this approach scales in more complex game design spaces. This can easily be achieved by using a CGF set of higher complexity, for example, $G_A \in \{XS, S, M, L, XL\}$. This set contains possible values for the area of the game parameter that allows for a more diverse range of levels moving from small to large.

This study serves as a starting point in a better understanding of how player behaviour in AR environments can be used to model their preferences. An aspect of this problem that is unresolved is guidelines of following a binary or ternary approach to classification. Binary classification performs better and is more stable. However, this is unsurprising since it models a simpler problem. This advantage over ternary classification seems preferable and could be accomplished by forcing a binary choice onto participants referred to as the 2-AFC [3]. A critique of this approach is that it seems a naive way of achieving better performance and could prove detrimental to optimization for true player experience.

Our work is built on studies of detecting emotions from movement and we apply it to predict emotional preferences, a complex problem that we have begun probing. Future research in addressing the discussed gap in establishing ground truth, improved feature extraction, and different models to address this problem (clustering players, non-linear classifiers or models for time-series data) would increase our understanding. Current trends of casual play on mobile devices show the need to explore other sensor information available to model players' emotional preferences. For instance, location and time of day are important features to consider for personalization in mobile gaming.

## VIII. Conclusion

To the best of our knowledge, this work is the first known study that exploits mobile sensor information to explore to what extent this can be used to personalize game experiences. Investigation in this area will allow AR games to become more engaging. Niantic (makers of *PoGO*) has announced the launch of *Harry Potter: Wizards Unite*, an AR game based on the HP universe. Game designers have begun to design immersive digital experiences based on popular fiction situated in the real world. Borrowing an example from the same universe, it would be difficult to imagine the ideal AR Quidditch[5] without the support of AI agents driven by a player experience model built from similar approaches we describe in this paper.


## Acknowledgment

The authors would like to thank all subjects who participated in this study, and our anonymous reviewers. This work is supported by the EPSRC and AHRC Centre for Doctoral Training in Media and Arts Technology (EP/L01632X/1).


## References


[1] Linowes J, Babilinski K, "Augmented Reality for Developers: Build Practical Augmented Reality Applications with Unity, ARCore, ARKit, and Vuforia." Packt Publishing Ltd, pp. 7–8, Oct 2017.

[2] Pedersen, Christopher, Julian Togelius, and Georgios N. Yannakakis. "Modeling player experience for content creation." IEEE Transactions on Computational Intelligence and AI in Games 2.1, pp. 54–67, 2010.

[3] S. Tognetti, M. Garbarino, A. Bonarini and M. Matteucci, "Modeling enjoyment preference from physiological responses in a car racing game," Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games, Dublin, 2010, pp. 321–328.

[4] Azad S, Saldanha C, Gan CH, Riedl MO. "Mixed Reality Meets Procedural Content Generation in Video Games." In Twelfth Artificial Intelligence and Interactive Digital Entertainment Conference, Sep 2016.

[5] Yannakakis, Georgios N., and John Hallam. "Entertainment modeling through physiology in physical play." International Journal of Human-Computer Studies 66.10, pp. 741–755, 2008.

[6] Rogers K, Colley M, Lehr D, Frommel J, Walch M, Nacke LE, Weber M. "KickAR: Exploring Game Balancing Through Boosts and Handicaps in Augmented Reality Table Football." InProceedings of the 2018 CHI Conference on Human Factors in Computing Systems, pp. 166, 2018

[7] Yannakakis, Georgios N., and Julian Togelius. "Experience-driven procedural content generation." IEEE Transactions on Affective Computing 2.3, pp. 147–161, 2011.

[8] Garber-Barron, Michael, and Mei Si. "Using body movement and posture for emotion detection in non-acted scenarios." 2012 IEEE International Conference on Fuzzy Systems. IEEE, 2012.

[9] Kleinsmith, Andrea, Nadia Bianchi-Berthouze, and Anthony Steed. "Automatic recognition of non-acted affective postures." IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 41.4, pp. 1027–1038, 2011

[10] Mandryk, Regan L., and M. Stella Atkins. "A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies." International journal of human-computer studies 65.4,pp. 329–347, 2007.

[11] Preece SJ, Goulermas JY, Kenney LP, Howard D. A "Comparison of feature extraction methods for the classification of dynamic activities from accelerometer data." IEEE Transactions on Biomedical Engineering, 2009

[12] Li B, Zhu C, Li S, Zhu T. "Identifying Emotions from Non-Contact Gaits Information Based on Microsoft Kinects." IEEE Transactions on Affective Computing, Dec 2016.

[13] Bernhardt, Daniel, and Peter Robinson. "Detecting affect from non-stylised body motions." International conference on affective computing and intelligent interaction. Springer, Berlin, Heidelberg, 2007.

[14] Fiechter, Claude-Nicolas, and Seth Rogers. "Learning subjective functions with large margins." ICML, 2000.

[15] Barrett, Lisa Feldman. "Solving the emotion paradox: Categorization and the experience of emotion." Personality and social psychology review 10.1, pp. 20–46, 2006.


[5]Quidditch is a fictional sport invented by author J. K. Rowling for her fantasy book series Harry Potter. For this example, it is important to note that each ball in this game has a distinct behavioural identity (because of 'magic').