Platformer Level Design for Player Believability

Elizabeth Camilleri
Institute of Digital Games
University of Malta
Msida, Malta
elizabeth.camilleri.12@um.edu.mt

Georgios N. Yannakakis
Institute of Digital Games
University of Malta
Msida, Malta
georgios.yannakakis@um.edu.mt

Alexiei Dingli
Dept. of Intelligent Computer Systems
University of Malta
Msida, Malta
alexiei.dingli@um.edu.mt

***Abstract*—Player believability is often defined as the ability of a game playing character to convince an observer that it is being controlled by a human. The agent's behavior is often assumed to be the main contributor to the character's believability. In this paper we *reframe* this core assumption and instead focus on the impact of the game environment and aspects of game design (such as level design) on the believability of the game character. To investigate the relationship between game content and believability we crowdsource rank-based annotations from subjects that view playthrough videos of various AI and human controlled agents in platformer levels of dissimilar characteristics. For this initial study we use a variant of the well-known Super Mario Bros game. We build support vector machine models of reported believability based on gameplay and level features which are extracted from the videos. The highest performing model predicts perceived player believability of a character with an accuracy of 73.31%, on average, and implies a direct relationship between level features and player believability.**

I. Introduction

Player believability is a highly subjective notion commonly viewed as the ability of a game playing character to convince observers that it is being controlled by a human player [1]–[4]. The problem of creating human-like[1] agents and measuring believability in agents is among the most challenging and popular research areas in the field of game artificial intelligence (AI) [5], [6]. While research in believable game bots has seen recent advances in games such as Unreal Tournament [7] and Super Mario Bros [8], generic methods for creating such bots are far from being available.

It is commonly assumed that believable game characters make games more immersive and entertaining for players [9] and that believability is solely dependent on the algorithm controlling the character's behavior. The evident relationship between believability and character control has driven the majority of studies in the area of game AI [5], [6]. However, the extent to which believable behavior in an algorithm-controlled game agent comes about from the controller has recently been questioned [1]. Inspired by speculations in [1] we instead focus on the degree to which believability is a product of the agent's (player's) environment. We thus introduce a *game content-centered* view on player believability instead of the traditional *controller-centered* perspective.

Taking a crowdsourcing approach to model perceived believability of player characters in a platformer game variant of the well-known Super Mario Bros, we asked over 350 subjects to annotate believability in playthrough videos of the game. In these videos four different players (two AI-controlled and two humans) play dissimilar level configurations of the game. The extracted gameplay characteristics and level features, on one hand, and the obtained annotations of believability, on the other hand, were used as the input and output, respectively, to construct a believability model. The model was built using rank support vector machines (RankSVMs) [10] as the crowdsourced annotations have an ordinal (rank-based) format [11]. A correlation analysis revealed that the average width of gaps in the level has a linear relationship with reported believability. Further, the best SVM model reaches 73.31% accuracy, on average, and maps level and gameplay features to believability, revealing non-linear relationships between the enemy placement and number of gaps in the level and believability.

This paper is novel in that it introduces an approach for modeling player believability using machine learned representations of crowdsourced annotations of believability. Most importantly, it offers the first empirical assessment of the extent to which level design influences the believability of play and sheds light into the association between game content and player believability.

II. Related Work

In this section we outline studies which are relevant to this paper including work in believability and its assessment, studies in player modeling as well as earlier work on the impact of content on believability.

A. Believability, Believable Agents and their Assessment

The notion of believability is highly subjective and cannot be objectively defined trivially. However, in virtual worlds it is generally understood as a form of suspension of the observer's disbelief or the *ability of a fictional or virtual world or character to give the illusion of life* [1], [12]. With respect to game agents or characters there are two main dimensions of believability in literature: *character believability* (i.e., the character *itself* is perceived to be real through its behaviour, emotive expression or graphical appearance) [1] and *player believability* (i.e., by exhibiting human-like *gameplay* behaviour, the character gives the impression that it is being controlled by

[1]For the purposes of this study *human-likeness* and *believability* are two terms used interchangeably.

a human *player*) [1]–[4]. This paper focuses on the assessment and modeling of *player believability* in platformer games.

Arguably, gameplay believability increases player engagement since it makes the game interaction more realistic [9]. There is also evidence suggesting that human players tend to prefer playing with or against other human players rather than AI agents due to the unpredictability in human gameplay behavior [1], [9]. Moreover, non-believable behavior such as repetitiveness and predictability (e.g., constantly falling into a gap in a platformer game or getting stuck in areas where human players could easily get through) seems to make games less challenging, thus deterring players [3]. 'God-like' behavior (e.g., going through an entire level without taking any damage) may also be considered non-believable [1]. Therefore, believability — as objectively as one can define it — likely lies somewhere in the middle of a gameplay spectrum, where the lower end includes poor, predictable behavior and the upper end includes optimal, god-like behavior; both of which are naturally perceived as non-believable.

Attempts at measuring the believability of playing behaviour include a criteria-based approach [2], [3] where believability is based on how many predefined criteria a playing character is observed to meet. However, a more common approach to measuring believability is through subjective assessment [1], [5], [13]–[17] where, similarly to the traditional Turing Test [18], human subjects are asked to observe a character in a game and indicate whether they believe it is being controlled by a human or by a computer. In this paper we follow the subjective assessment approach and we crowdsource rank annotations of believability given to video recorded player characters of a Super Mario Bros variant.

B. Player Modeling

Player modeling has been defined as the study of computational models of players in games which includes the detection, modeling, prediction and expression of human player characteristics which are manifested through cognitive, affective and behavioral patterns [19]. Player believability can be viewed as a core component of playing behavior and player experience [1]. One could thus assess and model believability using an approach similar to how other components of player experience have been modeled [20]–[22]. There are two main approaches to modeling players and aspects of the playing experience: the model-based (top-down) and the model-free (bottom-up) approach [19], [23]. In this study we adopt a bottom-up approach for modeling player believability and we derive the computational model from data. The input of the model contains gameplay and game content data [19] of players playing through varied levels of a platformer game whereas the output contains ordinal annotations of believability. To the best of our knowledge, this is the first crowdsourcing-based study for modeling believability in games.

C. Game Content and Believability

Game content is progressively increasing in demand and volume and, thus, becoming more expensive and time-consuming to create manually [24]. Procedural Content Generation (PCG), the "algorithmic creation of game content with limited or indirect user input" [25], is a natural and direct response to this challenge. Experience-driven PCG is a framework which views game content as the building block of player experience and involves the procedural generation of content based on a model of player experience [23]. The framework first consists of constructing a model of player experience which takes information about game content and the player and outputs some approximation for the current player experience state. The quality of the generated content is evaluated based on the model and a representation for the content is established. New content is then generated by searching for content that optimizes the player's experience with respect to the player experience model [23]. This core PCG approach has been applied in several studies to generate content for various player experience states including engagement, frustration, and challenge [20]–[22].

While game content and experience have an apparent direct relationship, no study has ever attempted to quantify the impact of game content on game playing believability; instead, studies in player believability focus on developing the agent's controller [5]. That said, interactions with game content in BioShock Infinite (Irrational Games, 2013) were used to enhance *character believability* of the player's NPC companion, Elizabeth [26]. Game mechanics have also been designed to hide non-believable characteristics of a simple algorithm in [27]. However, extending the argument that believable virtual agents must act according to the *context* they are situated within [28], we argue that the *modification* of the environment that characters act (play) within is of utmost importance for the observer's suspension of disbelief. Inspired by the core suggestions of [1] this study attempts, for the first time, to investigate the relationship between game level architecture and game playing believability and construct models of player believability based on level design features, gameplay characteristics and crowdsourced annotations of believability.

III. Testbed Game

Our testbed game is a variant of Infinite Mario Bros (a platformer game [29] and a popular benchmark in player modeling and content generation studies [20]–[22], [30]–[34]) as modified for the 2011 Mario AI Competition with sprites from an open-source platformer game called SuperTux.

The game therefore consists of a player character called Tux; a number of platforms (separated by gaps) which Tux can run on and jump between; coin collectibles for Tux to collect; and enemies which Tux must avoid or kill (by stomping on them — thus turning them into shells — or by kicking a shell to hit them). The main goal of the player in the game is to get Tux through levels containing these elements. The player is given three 'lives', or attempts, to complete the level. Each time Tux gets killed by falling into a gap or touching an enemy (or shell), one life is lost. If no lives are left, the game is over. Collecting coins increases the player's score. A screenshot of

Fig. 1. Screenshot of the testbed game used in this study.

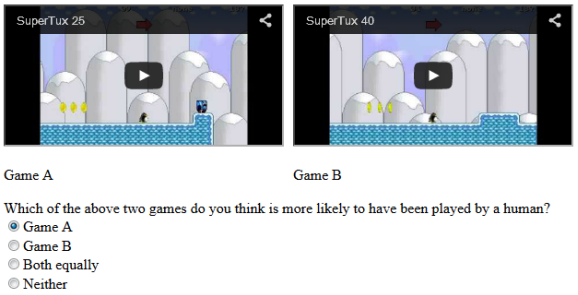

Fig. 2. Screenshot of the 4-AFC part of the online questionnaire.

the game displaying some of the above-mentioned elements is provided in Fig. 1.

IV. Feature Extraction and Data Collection

This section outlines the features extracted from levels and gameplay behaviour and also presents the crowdsourcing process through which believability annotations were collected.

A. Level design features

Following the approach of [21], [22], levels of the test-bed game were represented using the following four features: the number of gaps in the level (G), the average width of the gaps (G_w), the number of enemies (E), and the placement of enemies (E_p). For a particular game level, each parameter could be in one of two states: high or low. For G, the low state was set to 2 gaps and the high state was set to 10 gaps, while for G_w, the low state was set to 3 blocks and the high state was set to 8 blocks. Further, the low value of E was chosen to be 5 enemies while the high value for this feature was chosen to be 15. The E_p feature is represented by three values which define the probabilities of an enemy being placed near a gap, near a box, and on a flat surface. For the low state of E_p, the value of these three probabilities was chosen to be 10%, 10% and 80%, respectively, whereas those for the high state were chosen to be 80%, 10% and 10%. The choice of these features and their corresponding values which were set empirically was inspired by earlier studies in Super Mario level generation [20]–[22], [30], [31], [34]. Given the two states for each of the four level features, the resulting number of all their combinations, and thereby possible levels, is 16. The 16 levels were generated following the approach presented in [21].

B. Gameplay features

The following list of fourteen player metrics used for this study is based on earlier feature extraction attempts for level generation in Super Mario Bros [21], [22]: completion time (t_C), duration of last life (t_{LL}), percentage of time spent running right (t_R), percentage of time spent running left (t_L), percentage of time spent running (t_{Run}), number of times the 'run' button was pressed (P_{Run}), number of jumps (J), number of aimless jumps (J_a), number of coins collected (C), number of enemy shells kicked (S), number of deaths by falling into a gap (D_g), number of deaths by an enemy (D_e), number of enemies killed (K), number of enemies killed by kicking a shell (K_s).

C. Crowdsourced believability annotations

We recorded video clips of four players (two human players and two AI agents) playing through each of the sixteen generated levels. The two humans play the game differently with one clearly performing better than the other. The first AI agent is based on the A* pathfinding algorithm of Baumgarten [35] which won the Mario AI competition in 2009. The second AI agent is a hard-coded rule-based agent inspired by the REALM agent [36] which won the Turing Track of the Mario AI championship in 2010. Collectively, the four players purposely demonstrate varied behavior across all levels played and are therefore assumed to exhibit different levels of believability. The resulting 64 videos of all 16 levels played by all 4 players were stored for the believability annotation experiment described herein.

Human annotators were asked to view a number of gameplay videos which were randomly selected from the 64 videos without replacement. To establish the ground truth of a highly subjective notion such as believability we followed the rank-based approach for reliable annotation as proposed in [11]. The rank-based approach requires that annotators are provided with instances of the investigated variable in pairs (or more) and are asked to rank those instances according to a particular notion. For eliminating any short-term memory biases we chose to present the 64 videos in pairs, amounting to 2016 unique combinations in total. For each pair viewed, the observer was requested to provide a pairwise preference for believability using the 4-alternative forced choice (4-AFC) protocol [20]–[22]. That is, the subject was asked to specify which of the games in the two videos they believed *was more likely to have been played by a human*. Subjects could pick one of four possible responses; the game in video A, the game in video B, both or neither (see Fig. 2). The two videos in the pair were presented in a random order next to each other so that any primacy or recency effects are eliminated.

Prior to proceeding with video annotation the subjects were required to fill in a brief demographics questionnaire. The questions asked were: age; gender; how often do you play games (possible answers: never; few times a month; few times

a week; few hours a day; many hours a day); how would you rate your skill level in playing video games (possible answers: novice; average; good; excellent); have you ever played platformer games (possible answers: never; a few times, I am a novice player; many times, I am a good player; I am an expert player); have you ever played Super Mario Bros (possible answers: never; a few times, I am a novice player; many times, I am a good player; I am an expert player).

The crowdsourcing experiment was advertised widely and run for a whole month during which a total of $1,605$ randomly selected video pairs were ranked by approximately 391 subjects. This amounts to 79.6% of all possible combinations of pairs (i.e., 2016). Subjects reported clearly about their preference (i.e., they selected one of the two videos) in 984 out of the $1,605$ available video pairs; this is the *clear* preference dataset used for the analysis in the remaining of this paper.

V. Statistical Analysis

As a first phase of our analysis, this section presents various descriptive statistics we derived from the obtained data. This includes a brief analysis on the demographical data of the subjects, an analysis of the quality of the believability preferences and a correlation analysis between the game features considered and the reported believability annotations.

A. Demographical Data

A high-level descriptive statistical analysis on the demographical data of the respondents reveals that the subjects (average age is 36.9 years) are relatively balanced in terms of gender (females: 197; males: 168). However, the distribution of demographical data is slightly biased in terms of gaming skills (e.g., only 15 out of 365 subjects play many hours a day), gaming experience (e.g., only 30 subjects identify themselves as expert gamers), and previous exposure to platformer games (e.g., only 15 subjects identify themselves as expert platformer gamers) and Super Mario Bros (e.g., only 13 subjects identify themselves as expert Super Mario Bros gamers).

B. Believability Preferences

Assessing the validity of annotations of highly subjective notions such as believability is a challenging task. For instance, a believability preference for one of two videos depicting different AI players (instead of annotating these videos with the *neither* label) is not objectively invalid since AI players of the game may have the capacity of being perceived as believable. The analysis presented here is not intended to test the validity of the reported believability preferences but, rather, to provide an insight into the nature of these annotations following the evaluation approach presented in [1]. Figure 3 offers a first visualization of the crowdsourced believability preferences with respect to the four characters (the two AIs and the two human players) and three types of subjects: *all* subjects, *expert* subjects who consider themselves a good or an expert Super Mario Bros player, and *novice* subjects who either have never played Super Mario Bros before or they consider themselves novice Super Mario Bros players. For each of the four players and the three annotator types we depict the percentage of the four possible responses.

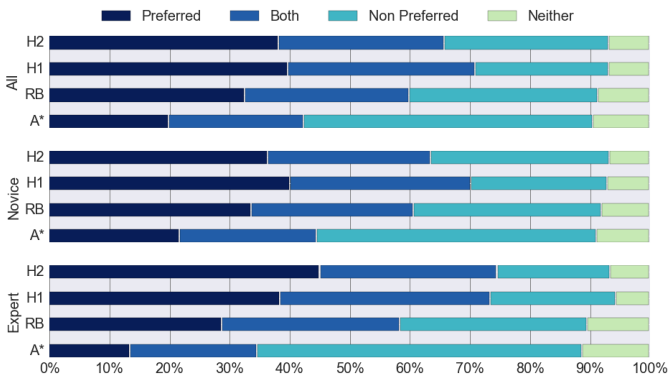

Fig. 3. Percentage of believability annotations across the different player characters and annotator sets. A* and RB (rule-based) are the two AI players; H1 and H2 are the two human players. "Expert", "Novice" and "All" indicate the expert, novice and all subjects, respectively.

What is apparent from Fig. 3 is that expert players, compared to novice players, are capable of better distinguishing the two AIs from the two human players. Further, it is clear that the preferences obtained generally match with the true nature of the players being assessed. In particular, the human players were correctly identified as humans by the majority of subjects. The RB player was able to convince a great number of subjects (especially the novices) that it was likely to be human in most of the video combinations that it was featured. On the other hand, the A* player, which was much more optimal in its actions, was not perceived as believable (especially by the experts). This observation reinforces the findings of Togelius et al. [1] in that there seems to be some lower and upper boundaries for the relation between the playing skill of the players and their believability. To some extent, this may serve as an indication that the quality of the results obtained in this study are at least on par in terms of validity with those obtained in [1].

Although through crowdsourcing we risked obtaining noisy or incorrect data (e.g., preference pairs provided by subjects who made their choices arbitrarily or who did not understand the questionnaire task), these results show that there is a certain degree of agreement between the responses of the majority of subjects, implying that there might indeed be some common reference point (or ground truth) with regards to human perception of player believability. This finding further reinforces the assumption that aspects of believability can be approximated and there are certain factors that affect believability which are common to many.

C. Correlation analysis

In this section we examine the relationship between each of the game features and the reported believability preferences through a correlation analysis. The rank correlation coefficient between each of the eighteen gameplay and level features and the crowdsourced preference pairs was calculated using the test

statistic $c(z) = 1/N \sum_{i=1}^{N} z_i$, where N is the total number of *clear* preference pairs (i.e., where one of the video clips in the pair is preferred over the other) and $z_i = 1$ if the value of the examined feature in the preferred video of the pair is greater than that of the other video (i.e., there is a match) and $z_i = -1$ otherwise (i.e., the feature value is lower in the preferred video and, thus, there is a mismatch). Statistical significance was calculated via a one-tailed binomial test.

The correlation coefficients of all 18 features are provided in Table I across the three different subject datasets: *all* subjects (984 clear preferences), *expert* subjects (206 clear preferences), and *novice* subjects (774 clear preferences). It can clearly be observed that — independently of subject class — 7 out of the 18 features are found to be highly and positively correlated ($p < 0.01$) with reported preferences. These include one level feature, and six gameplay features. When all subjects are considered, findings suggest that believability is perceived higher in levels where gaps are wider (G_w), and when players spend more time running left (t_L), are killed more often by enemies (D_e), kill more enemies by kicking shells (K_s), kick more shells (S), press the 'run' button more often (P_{RUN}) and take longer to die in their last life (t_{LL}). Indeed, running left and making use of shells are considered human playing characteristics and the results are in line with the findings in [1]. An interesting observation is that the above gameplay features have also been found to be associated with reported player *fun* in Super Mario Bros [21].

The difference between the expert and novice subjects is only limited to the degree of effect as some substantial differences are observed in the $c(z)$ values. Further, it seems that for the expert annotators, as opposed to novice annotators, the number of times the 'run' button was pressed and the duration of the player's last life were not factors that contribute to a character's believability. On the other hand, the average gap width was not an indicator of believability for the novice annotators.

From these correlations, it is clear that there exist some linear relationships between aspects of level design and perceived player believability. In particular, it seems that the wider the gaps are in a level, the more the behaviour of the players is showcased; as a result, this level feature seems to be a good predictor of player believability across all four players — particularly for the expert annotators. Interestingly enough, the gap width has already been found to be a good predictor of player *challenge*, *predictability*, *anxiety* and *boredom* in an earlier study in the Super Mario Bros game [21]. That said, an additional, preliminary correlation analysis was also conducted to measure the effect of the features on believability for each of the four players separately. While no effects were found for the rule-based agent, G_w and D_g are positively and highly correlated ($p < 0.05$) with perceived believability in both the A* agent and one of the human players (H1). Further, K_s seems to be a good predictor for the believability of both human players while P_{RUN}, t_L and D_e are highly correlated with the perceived believability of H1.

The correlation analysis presented above limits our findings to *linear* relationships between individual features and annotated believability. However, the effect that features have on believability is clearly dependent on the player (represented by gameplay features). This means that there are relationships in between the features themselves (both level and gameplay features) which may affect how well they can predict believability. Therefore, in the next section, we investigate the creation of *nonlinear* mappings between combinations of features and reported believability via preference learning.

TABLE I
RANK CORRELATIONS $c(z)$ BETWEEN ALL EXAMINED FEATURES AND REPORTED BELIEVABILITY PREFERENCES. VALUES IN BOLD INDICATE SIGNIFICANCE ($p < 0.01$).

Feature	All	Expert	Novice
G_w	**0.1291**	**0.2821**	0.0846
G	0.0825	0.0891	0.0840
E_p	−0.0503	0.0874	−0.0867
E	0.0303	0.1778	−0.0050
t_L	**0.2483**	**0.3548**	**0.2170**
D_e	**0.2134**	**0.3121**	**0.1900**
K_s	**0.2911**	**0.4805**	**0.2342**
S	**0.3043**	**0.3898**	**0.2744**
P_{RUN}	**0.1082**	0.0632	**0.1179**
t_{LL}	**0.0966**	0.0891	**0.0964**
J	0.0700	0.0792	0.0652
J_a	0.0636	0.0707	0.0593
C	−0.0698	−0.1097	−0.0617
t_R	0.0526	0.0051	0.0658
D_g	0.0593	0.1389	0.0373
K	−0.0309	−0.0726	−0.0232
t_C	−0.0260	−0.1179	−0.0041
t_{RUN}	−0.0010	0.0647	−0.0213

VI. PREFERENCE LEARNING FOR MODELING BELIEVABILITY

Preference learning is the task of learning global rankings. Assuming that there is an underlying global order that characterizes the provided rank annotations, the data can be machine learned via preference learning and a global ranking of believability can be derived [37]. This section provides a brief description of the preference learning methodology followed to construct models of player believability: in particular we discuss the feature selection method and the preference learning algorithm adopted. For all experiments reported in this paper we used the Preference Learning Toolbox (PLT) [38]. PLT is an open-source software package which integrates various data normalization, feature selection and machine learning algorithms for the task of learning from preferences.

A. Feature Selection

To select the most meaningful features that maximize the predictive capacity of our computational model of believability we used the Sequential Backward Selection (SBS) algorithm for all experiments presented in this paper. The SBS algorithm starts by feeding the preference learning algorithm with all available features and obtaining a performance value; performance is measured through cross-validation accuracy in this paper. At each iteration, SBS removes from the feature set

the feature which, upon its removal, improves the model's performance (accuracy) the most. This process continues until the feature set contains a single feature or until a certain accuracy threshold is reached.

B. RankSVM

This study's core aim is to infer a computational mapping between level and gameplay features (input) and the believability preferences (output). While the PLT offers a number of preference learning algorithms to choose from, we choose the RankSVM [10] algorithm for its comparatively low computational effort and well-evidenced performance. RankSVM is the preference learning variant of Support Vector Machines (SVMs) introduced by Joachims [10]. SVMs are models which map training instances to data points in a typically high-dimensional space and attempt to divide the data points into two categories via a hyperplane. The algorithm tries to find the dimension (or hyperplane) which best separates the training instances. Once the model is built, unseen instances are mapped to the space represented by the model and an output is produced based on which half of the space they are mapped to [39]. The mapping of training instances to data points in space is performed using a kernel function. In this study, we use the radial basis function (RBF) kernel as it yields superior performance to linear or polynomial kernels compared via trial experiments.

VII. Believability Model Construction

In this section, we describe the core preference learning experiments carried out in order to investigate both the impact of the annotators' experience with respect to Super Mario Bros and the impact of level content on the accuracy of a believability model. In all experiments reported in this section, features are min-max normalized to $[0, 1]$ and RankSVM (see Section VI-B) uses the RBF kernel with the default γ parameter value of 1. Finally, three-fold cross validation is used as the performance measure of all derived models.

A. The Impact of Super Mario Bros Experience

As mentioned earlier in Section IV-C, prior to the video annotation process, subjects were asked if they had ever played Super Mario Bros (i.e., the game that the testbed game is largely based on) before and could respond with one of four different levels of experience. We assume that a level of experience with a particular game may have an impact on a person's perception of believability in that game. That assumption is, in part, validated in the correlation analysis of Section V-C. Thus, in this first round of experiments, we vary the data used for training based on reported experience with Super Mario Bros and investigate their impact on the accuracy of the computational model of perceived believability. For that purpose, as in Section V-C, the full dataset (984 clear preferences) was split into two subsets: the *novice* (774 clear preferences) and the *expert* (206 clear preferences) annotators. We run SBS for each of the three sets and depict progression of the average 3-fold cross-validation accuracy in Fig. 4.

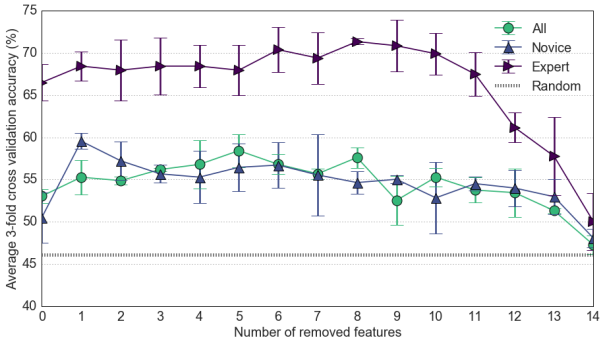

Fig. 4. Impact of Super Mario Bros Experience: 3-fold cross-validation accuracy as features are removed over iterations from the feature set using SBS for all, expert and novice subjects. The x-axis represents the iteration of the SBS process — or the number of removed features. Level features are not considered by SBS in this experiment. The random baseline represents the average accuracy of 10 random multi-layer perceptrons. Error bars depict standard error values.

It is important to note that in this set of experiments, the level features are not considered during the SBS process (i.e., they are purposely forced in the input of the SVM). The purpose of doing so is to allow the model sufficient capacity to capture aspects of reported believability through the content of the level. Furthermore, the performance of such a model can be used as a baseline against any model that does not put an emphasis on content features. In the next section all considered features (gameplay and content) are treated equally, thereby testing the degree to which level features are important for modeling gameplay believability.

The results illustrated in Fig. 4 reveal that the believability annotations provided by subjects who considered themselves to be 'good' or 'expert' players of Super Mario Bros ('Expert') managed to yield significantly higher accuracies compared to the other two datasets (all data and data annotations from novice Super Mario Bros players) and also achieve the highest accuracy improvement over the iterations of SBS. The highest accuracy of a model built solely on expert annotations is that of 71.36% with a corresponding feature set containing all four level features (since they were forced in this set of experiments) and ten gameplay features: t_C (completion time), t_{LL} (duration of last life), P_{RUN} (number of times the run button was pressed), S (number of kicked shells), J (number of jumps) and J_a (aimless jumps). The novice annotators did not seem to yield any significant improvement over the full set of preferences from all annotators. The best accuracy obtained when training RankSVMs on the full dataset is only 58.43%.

B. The Impact of Level Features

In the second round of experiments we chose to treat all features equally and consider them all during the SBS process so as to examine which would be picked as appropriate for capturing reported believability. Given the successful results of the expert subset of annotators in Section VII-A, in this round of experiments we focus on the expert subset and we compare the accuracy obtained between models that are

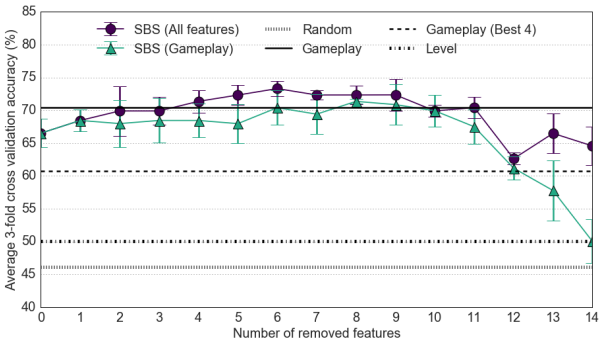

Fig. 5. Impact of level features: 3-fold cross-validation accuracy as features are removed over iterations from the feature set using SBS for the expert annotators. The x-axis represents the iteration of the SBS process — or the number of removed features. Level features are either considered — SBS (All features) — or not — SBS (Gameplay) — by SBS. The random baseline represents the average accuracy of 10 random multi-layer perceptrons. 'Gameplay', 'Gameplay (Best 4)', and 'Level' represent, respectively, the accuracy of SVMs featuring all gameplay, the best 4 gameplay, and all level features. Error bars depict standard error values.

trained when we consider and when we do not consider level features during the feature selection process. The progression of the average model accuracy for the two approaches for the expert annotators is shown in Fig. 5. In addition to the random baseline, the figure contains three more baseline performances: the average 3-fold cross-validation accuracy of an SVM containing solely the gameplay features, another one containing the best four gameplay features, and finally, one containing only the four game level features.

As shown in Fig. 5, the model featuring all (14) gameplay features clearly outperformed the model featuring all (4) level features. The latter is unlikely to be disadvantaged due to the lower number of features since the model featuring the best 4 gameplay features also outperforms it. Nevertheless, the selected features for the best performing model (73.31% accuracy — an improvement over the highest accuracy obtained in Section VII-A) still contains the level features G (number of gaps), G_w (average gap width), and E_p (enemy placement) in addition to 9 gameplay features: t_C, J, P_{Run}, t_{Run}, t_R, K_s, S, J_a and D_e. Although G_w was the only level feature found to be correlated with annotated believability, RankSVMs were able to capture non-linear relationships between more level features (i.e., G and E_p) and gameplay features, and reported believability. Even though the model performance improvement is not large, it is however clear that the best predictors of player believability consist of a combination of gameplay and level features, as also speculated in [1]. Further, it should be noted that the accuracy of 73.31% is considered very satisfactory given the highly subjective notion of believability.

VIII. Discussion

The core findings of this paper suggest that there is a strong and direct relationship between level design and the perceived believability of a character within a level. This validates, to an extent, the hypothesis that level design influences the perception of believability (at least in the platformer genre). Nevertheless, a number of limitations of this study are very likely to have hindered even more promising findings from emerging. First, the performance of all four players may have impacted the believability preferences as some subjects might have based their reasoning on the players' ability to complete levels. More varied and expressive agents would have equipped us with a much wider range of behaviors in the spectrum of believability. Second, the level and gameplay features chosen for this study are a subset of all possible game features that could be encapsulated. We can only envisage that a larger set of features can potentially reveal more information about the relationship we are studying; however, the small level feature set allowed us to both study the relationship and make the crowdsourcing experiment feasible to run (by preventing a combinatorial explosion of the total number of required videos). Third, while RankSVM is a reliable algorithm for learning preferences, other algorithms including backpropagation and neuroevolutionary preference learning must be tested on the data available. Finally, this study is only based on data derived from one particular game; experiments for testing the generality of the methods and the results on other platformer games and other game genres is a future goal of this work.

Despite current limitations, the initial findings of this study suggest that it is possible to optimize player believability merely by modifying the game level architecture, without necessarily adjusting the character behavior per se. Moreover, these findings hint towards the possible use of other forms of game content, beyond levels, to optimize player believability. Finally, other domains such as those of embodied conversational agents and robotics could also benefit from the outcomes of this work by putting an emphasis on modifying the environment within which agents act rather than focusing solely on their controllers to optimize their believable behaviour.

IX. Conclusions

In this study, we have introduced a data-driven modeling approach for constructing a mapping between gameplay, level design and believability. To crowdsource that mapping, we recorded a number of videos showing the gameplay of two human and two AI-controlled players — varying in playing style — playing through platformer levels of dissimilar design. Through the use of an online questionnaire, more than 350 annotators provided pairwise preferences of perceived believability for the characters appearing in the videos.

A first statistical analysis of the data revealed that the annotators' experience with the game impacts their preferences; further, the average gap width of a level and a number of gameplay features were found to be highly correlated to perceived believability. Then, the $1,605$ believability preferences, on one hand, and the gameplay and level features of each video, on the other, defined the output and input, respectively, of a preference learning process via RankSVMs that constructed the desired mapping. Through several experiments, we examined the impact of annotators' experience with Super Mario Bros, and the features considered by the model, on the

model's accuracy. The best accuracy of 73.31% was obtained when both level and gameplay features were considered by the model and the feature selection mechanism. The core findings of this study reveal both a linear and a non-linear relationship between level design and player believability that needs to be further explored in other game genres and domains.

ACKNOWLEDGMENTS

We would like to thank all participants of the crowdsourcing experiment. This work has been supported in part by the FP7 Marie Curie CIG project AutoGameDesign (630665).

REFERENCES

[1] J. Togelius, G. N. Yannakakis, S. Karakovskiy, and N. Shaker, *Believable Bots: Can Computers Play Like People?* Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, ch. Assessing Believability, pp. 215–230. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-32323-2_9

[2] T. Hinkkanen, J. Kurhila, and T. A. Pasanen, "Framework for evaluating believability of non-player characters in games," in *Workshop on Artificial Intelligence in Games*, 2008, p. 40.

[3] D. Livingstone, "Turing's test and believable ai in games," *Comput. Entertain.*, vol. 4, no. 1, Jan. 2006. [Online]. Available: http://doi.acm.org/10.1145/1111293.1111303

[4] F. Tencé, C. Buche, P. D. Loor, and O. Marc, "The challenge of believability in video games: Definitions, agents models and imitation learning," *CoRR*, vol. abs/1009.0451, 2010. [Online]. Available: http://arxiv.org/abs/1009.0451

[5] P. Hingston, *Believable Bots*. Springer, 2012.

[6] G. N. Yannakakis and J. Togelius, "A panorama of artificial and computational intelligence in games," *IEEE Transactions on Computational Intelligence and AI in Games*, 2014.

[7] J. Schrum, I. V. Karpov, and R. Miikkulainen, "Human-like combat behaviour via multiobjective neuroevolution," in *Believable bots*. Springer, 2013, pp. 119–150.

[8] N. Shaker, J. Togelius, G. N. Yannakakis, L. Poovanna, V. S. Ethiraj, S. J. Johansson, R. G. Reynolds, L. K. Heether, T. Schumann, and M. Gallagher, "The turing test track of the 2012 mario ai championship: entries and evaluation," in *Computational Intelligence in Games (CIG), 2013 IEEE Conference on*. IEEE, 2013, pp. 1–8.

[9] J. D. Miles and R. Tashakkori, "Improving the Believability of Non-Player Characters in Simulations," in *Proceedings of the 2nd Conference on Artificiel General Intelligence (2009)*. Atlantis Press, 2009.

[10] T. Joachims, "Optimizing search engines using clickthrough data," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 133–142.

[11] G. N. Yannakakis and H. P. Martínez, "Ratings are Overrated!" *Frontiers in ICT*, vol. 2, p. 13, 2015.

[12] J. Bates, *The nature of characters in interactive worlds and the Oz project*, 1992.

[13] I. Umarov and M. Mozgovoy, "Believable and effective ai agents in virtual worlds: Current state and future perspectives," *International Journal of Gaming and Computer-Mediated Simulations (IJGCMS)*, vol. 4, no. 2, pp. 37–59, 2012.

[14] S. McGlinchey and D. Livingstone, "What believability testing can tell us," in *Proceedings of the International Conference on Computer Games: Artificial Intelligence, Design, and Education*, 2004, pp. 273–277.

[15] B. Gorman, C. Thurau, C. Bauckhage, and M. Humphrys, "Believability testing and Bayesian imitation in interactive computer games," in *From Animals to Animats 9 : Proceedings. 9th International Conference on Simulation of Adaptive Behavior*, S. Nolfi, Ed. Springer-Verlag, 2006.

[16] B. Mac Namee, "Proactive persistent agents-using situational intelligence to create support characters in character-centric computer games," Ph.D. dissertation, 2004.

[17] J. E. Laird and J. C. Duchi, "Creating human-like synthetic characters with multiple skill levels: A case study using the soar quakebot," *Proceedings of the AAAI 2000 Fall Symposium: Simulating Human Agents*, pp. 54–58, November 2000.

[18] P. Hingston, "A turing test for computer game bots," *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 1, no. 3, pp. 169–186, 2009.

[19] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. André, "Player modeling," *Dagstuhl Follow-Ups*, vol. 6, 2013.

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

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

[22] N. Shaker, G. N. Yannakakis, and J. Togelius, "Towards Automatic Personalized Content Generation for Platform Games," 2010.

[23] G. N. Yannakakis and J. Togelius, "Experience-Driven Procedural Content Generation," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 147–161, July 2011.

[24] M. Hendrikx, S. Meijer, J. Van Der Velden, and A. Iosup, "Procedural content generation for games: A survey," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 9, no. 1, pp. 1:1–1:22, Feb. 2013. [Online]. Available: http://doi.acm.org/10.1145/2422956.2422957

[25] J. Togelius, N. Shaker, and M. J. Nelson, "Introduction," in *Procedural Content Generation in Games: A Textbook and an Overview of Current Research*, N. Shaker, J. Togelius, and M. J. Nelson, Eds. Springer, 2015.

[26] IGN, "BioShock Infinite - The Revolutionary AI Behind Elizabeth," https://www.youtube.com/watch?v=2viudg2jsE8, Mar 2013.

[27] M. Cerny, "Sarah and sally: Creating a likeable and competent ai sidekick for a videogame," in *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*, 2015.

[28] F. de Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. De Carolis, "From Greta's Mind to Her Face: Modelling the Dynamics of Affective States in a Conversational Embodied Agent," *Int. J. Hum.-Comput. Stud.*, vol. 59, no. 1-2, pp. 81–118, Jul. 2003. [Online]. Available: http://dx.doi.org/10.1016/S1071-5819(03)00020-X

[29] G. Smith, M. Cha, and J. Whitehead, "A framework for analysis of 2d platformer levels," in *Proceedings of the 2008 ACM SIGGRAPH Symposium on Video Games*, ser. Sandbox '08. New York, NY, USA: ACM, 2008, pp. 75–80. [Online]. Available: http://doi.acm.org/10.1145/1401843.1401858

[30] N. Shaker, G. N. Yannakakis, and J. Togelius, "Digging deeper into platform game level design: session size and sequential features," in *Applications of Evolutionary Computation*. Springer, 2012, pp. 275–284.

[31] ——, "Crowdsourcing the Aesthetics of Platform Games," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 5, no. 3, pp. 276–290, Sept 2013.

[32] J. Ortega, N. Shaker, J. Togelius, and G. Yannakakis, "Imitating human playing styles in super mario bros," *Entertainment Computing*, vol. 4, no. 2, pp. 93–104, 4 2013.

[33] N. Shaker, M. Nicolau, G. N. Yannakakis, J. Togelius, and M. O'Neill, "Evolving levels for super mario bros using grammatical evolution," in *2012 IEEE Conference on Computational Intelligence and Games (CIG)*, Sept 2012, pp. 304–311.

[34] N. Shaker, G. N. Yannakakis, and J. Togelius, "Feature analysis for modeling game content quality," in *2011 IEEE Conference on Computational Intelligence and Games (CIG'11)*, Aug 2011, pp. 126–133.

[35] J. Togelius, S. Karakovskiy, and R. Baumgarten, "The 2009 Mario AI competition," in *Evolutionary Computation (CEC), 2010 IEEE Congress on*. IEEE, 2010, pp. 1–8.

[36] S. Bojarski and C. B. Congdon, "REALM: A rule-based evolutionary computation agent that learns to play Mario," in *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*. IEEE, 2010, pp. 83–90.

[37] L. Rigutini, T. Papini, M. Maggini, and M. Bianchini, *Artificial Neural Networks - ICANN 2008: 18th International Conference, Prague, Czech Republic, September 3-6, 2008, Proceedings, Part II*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, ch. A Neural Network Approach for Learning Object Ranking, pp. 899–908. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-87559-8_93

[38] V. E. Farrugia, H. P. Martínez, and G. N. Yannakakis, "The preference learning toolbox," *arXiv preprint arXiv:1506.01709*, 2015.

[39] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," *Data mining techniques for the life sciences*, pp. 223–239, 2010.