Human Computation for Procedural Content Generation in Platform Games

Willian M. P. Reis
Departamento de Informática
Universidade Federal de Viçosa
Viçosa, Minas Gerais, Brazil

Levi H. S. Lelis
Departamento de Informática
Universidade Federal de Viçosa
Viçosa, Minas Gerais, Brazil

Ya'akov (Kobi) Gal
Department of Information Systems
Ben-Gurion University of the Negev
Beersheva, Israel

***Abstract*—One of the major challenges in procedural content generation in computer games is to automatically evaluate whether the generated content has good quality. In this paper we describe a system which uses human computation to evaluate small portions of levels generated by an existing system for the game of Infinite Mario Bros. Several such evaluated portions are then combined into a full level of the game. The composition of the small portions into a full level is done by accounting for the human-annotated information and the mathematical model of tension arcs used in interactive drama and storytelling. We tested our system with human subjects and the results show that our approach is able to generate levels with better visual aesthetics and that are more enjoyable to play than other existing approaches.**

I. Introduction

In procedural content generation (PCG) one is interested in using computer systems to automatically generate content for specific problem domains. For example, when applied to computer games, PCG systems automatically produce levels, rules, textures, and other contents traditionally generated by human professional designers. PCG in computer games has drawn a lot of attention in the recent years—see [1], [2], [3] for surveys. One of the reasons PCG systems have attracted so much attention in the games community is that these systems can be used to reduce the cost and time required for producing computer games. Moreover, PCG can be a way of generating content tailored to specific players [4], [5] and also to increase the replayability of the games [6].

One of the major challenges in PCG is to automatically evaluate whether the generated content has good quality. In the context of computer games it is important to be able to evaluate whether the generated content is rated highly by users with respect to different measures such as visual aesthetics, enjoyment, and satisfaction. Researchers have aimed at understanding the concept of enjoyment to develop methods to automatically evaluate content in computer games. For example, Togelius et al. [7] use a player's behavior to generate race tracks which are more fun to the player; Liapis et al., [8] introduce general evaluation functions which are applicable to different games; Sorenson et al. [9] learn a model of enjoyment based on levels generated by professional designers. Such works are usually motivated by the fact that it is not possible to have humans evaluating content produced by machines. For example, Shaker et al. [10] stated that "because of the large amount of content that can be generated, it is not feasible to humanly judge the results, and automatic evaluation becomes a necessity".

A. Our Contributions

In this paper we show that human computation [11] is a valid alternative to evaluate content generated by PCG systems for the game of *Infinite Mario Bros* (IMB) [12]. That is, we rely on human workers to measure whether particular content is of good or bad quality. Human evaluations can be quickly obtained in environments such as the Amazon Mechanical Turk (AMT) for a modest price. When using environments such as the AMT we do not need to ask (and perhaps bother) the player about their preferences as we can have human workers evaluating the content generated.

Our system uses human computation to quickly evaluate small portions of the game generated by an existing PCG system. In this paper we refer to these small portions of the game as *small levels*. Human workers provide annotations about the visual aesthetics, enjoyment, and difficulty of the small levels. Several annotated levels are then combined into a full level of IMB (we also refer to the full levels as *larger levels*). The composition of the small levels into larger levels is done by accounting for the human-annotated information and the mathematical model of tension arcs used in interactive drama and storytelling [13].

In this paper we perform two experiments. In the first experiment human workers annotated a collection of almost 2,000 small levels of the game of IMB. In the second experiment human subjects evaluated our proposed approach and other approaches encountered in the literature. Our quantitative and qualitative results on the second experiment show that the levels generated by our method can be more enjoyable to play and can have better visual aesthetics than the levels generated by other schemes. Our results show that the human-computation scheme is practical and thus a good alternative approach to generate good-quality content for the game of IMB.

II. Related Work

Here we describe how our work differentiates from a few other PCG works in platform games. For a thorough literature review we refer the reader to [1], [2], [3].

Smith et al. [14] presented Tanagra, a system for developing levels for 2D platform games such as IMB. Tanagra allows the game designer to specify parts of the level and the system completes the level while respecting the designer's decisions. Sorenson et al. [9] presented a system which uses the idea of rhythm groups introduced by Smith et al. [15] to define a computational model of player enjoyment to evolve levels of IMB. Our idea of applying tension arcs to generate full levels of IMB is somewhat similar to the rhythm groups. The main difference between the tension arcs (as we define in this paper) and the rhythm groups is that the former are based on human-annotated content while the latter are based on a mathematical model for approximating the player's anxiety. Moreover, rhythm groups capture the low-level challenges of the game. By contrast, as we explain later in this paper, tension arcs aim at controlling the player's tension by controlling the high-level challenges of the game.

Shaker et al. [4] describe a system for generating player-specific content for IMB which directly asks questions to the players about their preferences. By contrast, as our system is not designed to generate player-specific content, we do not ask questions to players directly, we ask questions to human workers as a pre-processing step instead. In another work, Shaker et al. [16] showed how to extract features to learn predictive models of the player's experience in IMB. By contrast, we use the annotations provided by humans to directly generate levels of IMB—we assume that human annotations can be quickly obtained, therefore we do not learn a model to generalize the annotated data. Another difference is that we introduce a novel method for connecting the small levels to generate a large level of the game of IMB.

The systems mentioned above are orthogonal to the ideas we introduce in this paper. That is, one could use any PCG system for IMB in conjunction with our system by having them generating small levels which are then evaluated by human workers and combined by our tension arc-based system into a full level of the game.

The mathematical model of tension arcs we use for composing the IMB levels have been successfully used in interactive drama [13] and storytelling [17]. We show that such model can also be effective in PCG for platform games. Recently, algorithms using human computation have become a good alternative for solving tasks which are hard for computers to solve. For example, reCAPTCHA [18] uses human computation to digitalize words that computer programs are not able to accurately recognize. In this paper we show that human computation is a viable approach for evaluating the content generated by computational intelligence methods.

III. The Problem Domain of Mario Bros

In this paper we are interested in the problem of automatically generating levels of the game of IMB, a game which has been used by other researchers to evaluate PCG systems—for more details on the use of IMB in research please refer to the work of Togelius et al. [19]. The advantage of using IMB in our experiments is that we are able to compare the quality of the content generated by our system with that of other systems found in the literature.

A screenshot of IMB is shown in Figure 1. The player controls Mario (on the center of screen). Mario's goal is to reach the rightmost spot of the level. In order to succeed, Mario has to avoid enemies and other challenges. The IMB levels are grid spaces containing a set of objects such as mountains and enemies. Figure 1 depicts part of a level which contains four mountains of different widths and heights, several enemies, and a few boxes which Mario can break to collect power-up items. Every object is associated with a location on the grid (x and y coordinates) and some of the objects such as mountains and pits can have different heights and widths—boxes, a few enemies, and the small version of Mario himself occupy a single cell on the grid. In this paper all full levels are represented as a grid of size 160×15.

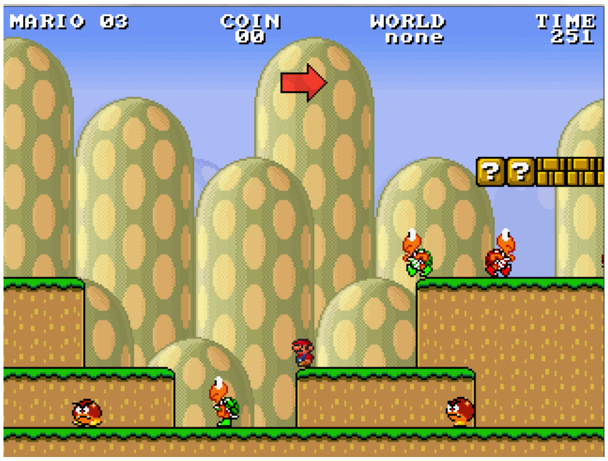


Fig. 1: Screen shot of the game of Infinite Mario Bros.

Let $\mathbf{L} = \{o_1, o_2, \cdots, o_n\}$ be a level of IMB where $o_1, o_2, \cdots, o_n$ are the n objects composing $\mathbf{L}$. The PCG problem for IMB is to choose the set of objects in $\mathbf{L}$ as well as the objects' x and y coordinates. For some of the objects such as pits and mountains we also need to define their height and width values. Our goal is to generate a level $\mathbf{L}$ which is both visually appealing and enjoyable to play.

IV. The Human Computation Approach to PCG

In this section we describe our approach for automatically generating levels of IMB. We call our system Human-Computation Tension Arc-Based (HCTA) level generator. A high-level description of HCTA is as follows.

1) A PCG system generates a collection Γ of levels of IMB with grid size of 20×15.
2) Workers annotate each $l \in \Gamma$ with respect to three measures: enjoyment, visual aesthetics, and difficulty.
3) Different levels of Γ are concatenated to form a full level of larger size.

The central idea behind HCTA is to have human workers quickly annotating a large number of small levels (of size 20×15) of IMB. Then, these small levels are combined in different ways to generate an even larger number of different IMB levels of larger size. It is important to note that HCTA does not assume that the human workers are professional game designers. HCTA uses the annotation provided by anyone able to play IMB.

A large number of annotated small levels can be quickly obtained for a modest cost in environments such as AMT. Another option is to make our system available online and ask for volunteers to annotate the levels. In this paper we use the latter. HCTA assumes that the workers' perceived visual aesthetics and enjoyment on the small levels will be similar to those of people playing the resulting larger level. Also, while we recognize that different players might enjoy different styles of gameplay, we expect HCTA to produce levels with good aesthetics and which are enjoyable to play on average. Another observation is that HCTA does not disrupt the gameplay as it asks questions to human workers as a preprocessing step and not to the players. We believe that asking questions to the players could break their gameplay immersion.

In the next sections we describe each of the steps of HCTA mentioned above. First, we describe the basic PCG system used to generate the library Γ, then the experiment in which volunteers annotated the small levels in Γ, and finally, we explain the method we introduce to concatenate small levels into full-sized levels of IMB.

V. Basic System for Level Generation

The system we use for generating Γ is referred as the Notch Level Generator (which we abbreviate as NLG) after the game designer Markus "Notch" Persson. NLG receives as input a difficulty value d for stochastically determining the number of enemies and other challenges to be placed in the level. The levels NLG generates will tend to be harder for larger values of d. NLG starts with an empty level—in our case an empty grid of size 20×15—and it iteratively adds objects to the grid according to the value of d. NLG follows simple heuristics for adding objects to the levels. For example, when adding a hill, NLG limits the hill's height to a value that Mario is able to reach by jumping from the ground or from some other hill.

We use NLG to generate more than 2,000 levels of size 20×15 with values of d selected uniformly at random to ensure a collection of levels Γ with different difficulty.

VI. Human-Annotated Levels

The NLG system follows a set of hardcoded rules to stochastically create levels of IMB. Although simple, the levels NLG generates can have good visual aesthetics and be enjoyable to play. Unfortunately, however, due to the stochasticity of the generator and the lack of a systematic evaluation, NLG also produces levels which are not visually appealing and are not necessarily enjoyable to play.

In HCTA human workers play all levels in Γ and annotate each of the levels with respect to three measures: enjoyment, visual aesthetics, and difficulty. We note that the value of d provided by the NLG system offers a good indication of the player's perceived difficulty of the level. However, we noticed in preliminary experiments that sometimes a level l has a large number of enemies and challenges (determined by a large value of d) but l is not necessarily a difficult level to play. This happens because there could be alternative paths the player can choose in order to avoid facing all challenges posed by the level. Thus, workers also evaluated the difficulty of the levels in Γ.

We made our system available for download and invited undergraduate and graduate students in the Departamento de Informática, at Universidade Federal de Viçosa, in Brazil to play the levels in Γ. The students voluntarily and anonymously played the levels. Before playing the levels the volunteers were instructed that the levels played would be much smaller than regular levels of the game of IMB. We assumed that most of the volunteers had played the game of Mario before and we did not want them to be disappointed by the reduced size of the levels. After playing each level the volunteers provided a score from 1 to 9 to each of the following criteria: enjoyment, visual aesthetics, and difficulty.

A score of 1 for enjoyment, visual aesthetics, and difficulty means that the level is enjoyable, is visually appealing, and is difficult, respectively. Similarly, a score of 9 for enjoyment, visual aesthetics, and difficulty, means that the level is not enjoyable, is not visually pleasing, and offers no challenge to the player, respectively.

The order in which the volunteers annotated the levels in Γ was random, and every level l in Γ could be annotated only once by each volunteer. By using such an approach the volunteers annotated most of the levels in Γ, and some of the levels were annotated more than once by different volunteers. We use the average score given by different volunteers in case a level was annotated more than once. We removed from Γ the levels that were not annotated by any volunteer.

Since the levels could be played in a few seconds, a few minutes of work was worth several annotated levels. We counted 1,928 annotated levels in our server 30 days after we advertised our system in our department's mailing list.

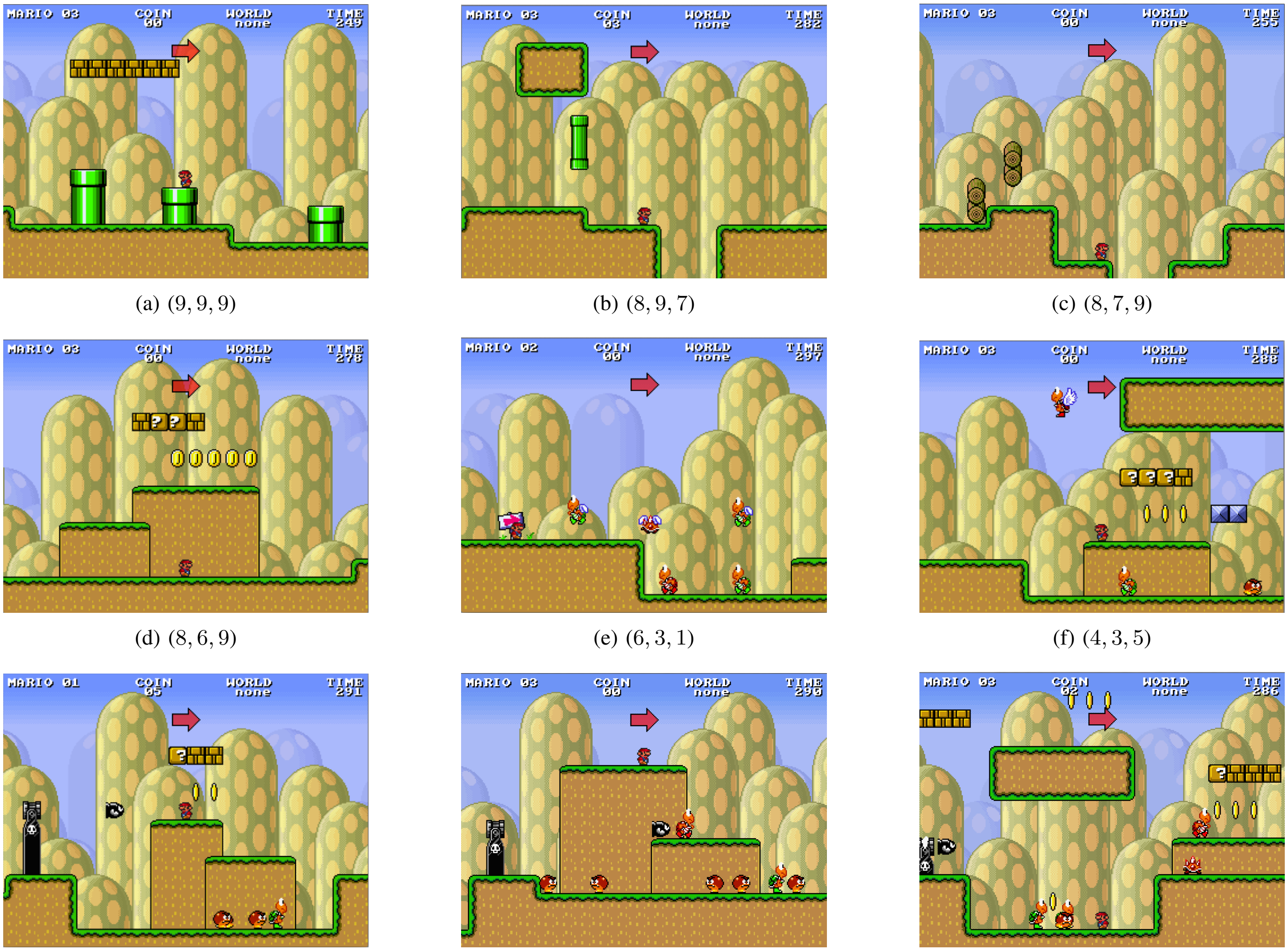
A. Representative Annotated Levels

We now present a few annotated levels from our experiment with human workers. First we show the scores of a few representative levels (quantitative results), then we report a few comments provided by the workers (qualitative results).

1) Quantitative Results: Figure 2 shows a few representative human annotated levels from our experiment. The numbers in parenthesis show the value in a scale from 1 to 9 of enjoyment, visual aesthetics, and difficulty, respectively. While the evaluation is subjective and the volunteers do not explain the scores provided, we now try to interpret the evaluation scores of the levels shown in Figure 2.

The level shown in Figure 2 (a) shows a level with no challenge—there are no enemies or pits. Moreover, the objects are oddly placed on the screen. That is, the platform of blocks on the top of the screen has no purpose since it is not reachable by Mario. These reasons support the score of $(9, 9, 9)$ provided by the worker to the level.

The levels shown in (b) and (c) are similar to each other: both have a pit and thus offer some challenge to the player. However, the volunteer who annotated (c) found the level very easy and marked its difficulty as a 9, while the volunteer who annotated (b) decided to give a 7 to the level. It is expected to see some variance on the scores provided by the workers. However, we note that the scores for difficulty for both (b) and (c) are somewhat similar. We also notice in these two levels that the score of aesthetics is 9 for (b) and 7 for (c). This is justified by the unreachable mountain above the green pipe in (b) which deteriorates the visual aesthetics of the level. The level shown in (d) has a better aesthetics score than (a), (b), and (c), which is reasonable since all objects are well distributed on the screen and are all reachable by Mario.


MARIO 03 COIN 00 WORLD none TIME 249
MARIO 03 COIN 03 WORLD none TIME 282
MARIO 03 COIN 00 WORLD none TIME 255
MARIO 03 COIN 00 WORLD none TIME 278
MARIO 02 COIN 00 WORLD none TIME 297
MARIO 03 COIN 00 WORLD none TIME 288
MARIO 01 COIN 05 WORLD none TIME 291
MARIO 03 COIN 00 WORLD none TIME 290
MARIO 03 COIN 02 WORLD none TIME 286

(a) $(9, 9, 9)$

(b) $(8, 9, 7)$

(c) $(8, 7, 9)$

(d) $(8, 6, 9)$

(e) $(6, 3, 1)$

(f) $(4, 3, 5)$

(g) $(1, 1, 4)$

(h) $(6, 5, 7)$

(i) $(3, 3, 4)$

Fig. 2: A few representative human evaluations for levels in Γ. The numbers in parenthesis show the value of enjoyment, visual aesthetics, and difficulty, respectively in a scale from 1 to 9. For example, a level with values $(9, 9, 9)$ mean that the level is not enjoyable, it has bad visual aesthetics, and it is not challenging.

The level shown in (e) is considered more difficult (difficulty of 1) than level (h) (difficulty of 7), despite the fact that (h) has more enemies than (e). This is because in (h) the player has an alternative path going through the top of the mountain while in (e) the player must act quickly before being caught by three flying enemies. Similarly, (f) is considered more difficult than (h) because the flying turtle comes from the top of the mountain after Mario while in (h) the player can safely stay on the top of the mountain. However, (f) is considered easier than (e) probably because of the reduced number of enemies and because in (f) Mario can hide from the flying turtle below the blocks in the center of the level. These examples clearly illustrate the fact that simply counting the number of enemies on the screen can be ineffective in determining the actual difficulty of the level.

Finally, levels (g) and (i) pose interesting challenges which could make the levels more enjoyable to play. In (g) Mario has to dodge bombs while crossing the level and collecting the power-up item available in the question-mark block on the center of the level; in (i) Mario has to face several enemies to reach the other side of the screen.

Interestingly, we noticed a high positive correlation between enjoyment and visual aesthetics in the levels annotated by the volunteers—coefficient of 0.72 amongst all evaluations. This result suggests an interesting relation of visual aesthetics and enjoyment: levels with well-placed objects tend to be more enjoyable to play. Enjoyment and difficulty are also correlated (coefficient of 0.67). The relation between difficulty and enjoyment is well known. Namely, Piselli et al. [20] showed that the Yerkes-Dodson law [21] applies to computer games in the sense that pleasure will be maximum somewhere in between the largest and the smallest challenge. That is, enjoyment increases with difficulty up to some point, where the level gets too difficult to be enjoyable.

The smallest correlation coefficient is between aesthetics and difficulty (coefficient of 0.47). Clearly these two metrics can be very uncorrelated. For example, there can be levels which are too difficult due to a large number of enemies and also have bad aesthetics due to the poor placement of objects on the screen.

2) Qualitative Results: We also allowed the volunteers to optionally enter comments on the levels played. One of the volunteers stated that the levels were too small to be enjoyable, specifically they wrote: *“After playing 5 levels I noticed that they were too short to be fun”*. Such a comment is an indication that the enjoyment on the small level could differ from the enjoyment on the same level on a larger context. Nonetheless, other volunteers apparently enjoyed playing some of the small levels as they wrote comments such as *“Perfect!”* and *“This level is very good!”*.

B. Relation of Level Size and Enjoyment

The qualitative results shown above can be worrisome as it shows that people’s enjoyment while playing the small levels (*Fun-Small*) could differ from people’s enjoyment while playing the larger levels (*Fun-Large*). One solution for making *Fun-Small* being closer to *Fun-Large* is to increase the size of the small levels. For example, instead of using levels of size 20×15, one could use levels of size 40×15, which are closer to the actual full level size. However, using larger levels in the human computation process will imply in an increase in the time required to evaluate each individual level and also in an increase on the number of levels evaluated in order to have a collection of levels with good quality, as we now explain.

Let p be the probability of the NLG system generating a small level of size 20×15 with good visual aesthetics and which is enjoyable to play. In order to have one level of this size with good visual aesthetics and enjoyment in our collection of levels, one has to generate and humanly evaluate $\frac{1}{p}$ levels on average. If the evaluated levels are twice as large (size of 40×15), then NLG has to generate two small levels of size 20×15 in a row with good aesthetics and enjoyment. It is reasonable to assume that NLG generates such a level with probability of p^2. Thus, one has to generate approximately $\frac{1}{p^2}$ levels of size 40×15 on average to have a single small level with good visual aesthetics and enjoyment in our collection.

To illustrate the reasoning above, let $p = 0.01$. On average we would need to generate 100 levels of size 20×15 to have one good level in our collection. By contrast, we would need approximately 10,000 levels on average to have one good level of size 40×15 in our collection—we would have to evaluate two orders of magnitude more levels when doubling the size of the levels. We believe there is a compromise between the relation of *Fun-Small* and *Fun-Large* and the number of levels the workers have to evaluate to create a collection of levels containing good-quality levels. We believe that the size of 20×15 offers a good tradeoff between the human computation effort and relation of *Fun-Small* with *Fun-Large*. Previous works also used levels of size similar to our small levels. For example, the rhythm groups in Sorenson et al.’s system [9] and Tanagara’s beats [14] have sizes similar to our small levels.

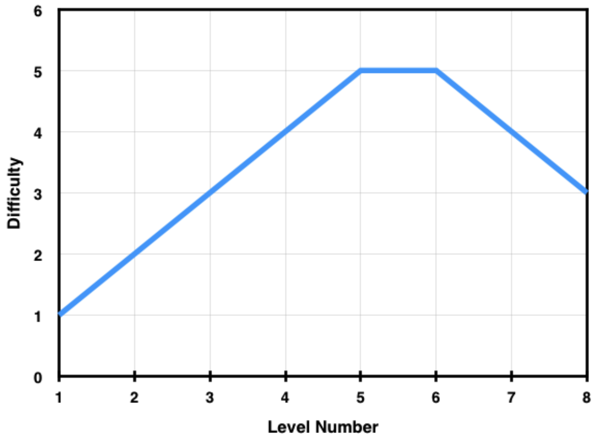


Fig. 3: Tension arc used in our experiments. The difficulty values are obtained with human computation, and the values on the x-axis denote the tension arc’s ordering.

VII. Tension-Arc Based Small Level Combination

The human computation procedure described above outputs a collection of small annotated levels. In this section we describe how HCTA combines such levels into a full IMB level. For doing so we borrow the idea of tension arcs used in interactive drama [13] and story writing [17]. Vogler [17] argues that in storytelling, either in books, movies or TV shows, the story usually follows a pattern in which the tension builds up to a climax and it drops before concluding. In HCTA we use similar idea to try to control the player’s tension throughout a full level of IMB as we now explain.

Algorithm 1 Tension-Arc Concatenation

```
Require: Collection of annotated small levels $\Gamma$, tension arc
    $T = \{d_1, d_2, \cdots, d_M\}$, quality parameter $k$.
Ensure: $\nabla = \{l_1, l_2, \cdots, l_M\}$
for $i = 1$ to $M$ do
    choose at random a small level $l$ in $\Gamma$ with difficulty $d_i$
    which is at same time in the set of $k$ levels with highest
    visual aesthetics score and in the set of $k$ levels with
    highest enjoyment score.
    append $l$ to $\nabla$
end for
```

Let a $\nabla = \{l_1, l_2, \cdots, l_M\}$ be a totally ordered set with $\nabla \subseteq \Gamma$ and each level $l \in \nabla$ with size $x \times y$. ∇ represents a level of size $x \cdot M \times y$ formed by the concatenation of the levels $l \in \nabla$ according to some ordering $\mathbf{O}$. ∇ is the final output of HCTA.

In HCTA a *tension arc* defines the ordering $\mathbf{O}$. A tension arc T is a sequence of difficulty values $\{d_1, d_2, \cdots, d_M\}$ where d_1 is the difficulty of the first small level composing ∇, d_2 is the difficulty of the second small level, and so on. Figure 3 shows a tension arc as a function of the levels in ∇ and their difficulty value. The numbers on the x-axis from left to right define the ordering $\mathbf{O}$ of the small levels composing ∇. Such a tension arc follows the idea of building up the tension by increasing the difficulty of the level as the player evolves into the game, until reaching the climax (small levels 5 and 6 in ∇) and dropping on small levels 7 and 8.

For a given collection of annotated levels Γ, tension arc T, and integer k, HCTA builds ∇ as shown in Algorithm 1. HCTA selects a level l in Γ for each difficulty value d in T and appends l to ∇. The parameter k controls ∇'s quality by allowing HCTA to select only the levels amongst the k with highest scores for visual aesthetics and enjoyment. For small values of k ∇ will be composed only by small levels from Γ with the highest human perceived visual aesthetics and enjoyment, but the levels produced by HCTA might be repetitive—larger values of k allow for less repetition but for a likely loss in quality. In our experiments we use $k = 50$.

HCTA can guarantee level playability by generating playable small levels and connecting them with "safe walking areas", as explained by Smith et al. [22]. However, we did not implement such areas in our system as we did not observe the generation of non-playable levels in our preliminary experiments with NLG as the generator of the small levels.

One can easily notice that our proposed approach can be implemented in ways different than the one we describe in this paper. For example, instead of using a tension arc with growing tension, one could try to use a constant difficult value throughout, or even random values. Moreover, instead of using NLG to generate the collection Γ one could use any other PCG system. Also, instead of using human computation, one could also use other measures (e.g., the metrics introduced in [23]).

VIII. Empirical Evaluation

In this section we present our empirical evaluation of the HCTA system with human subjects.

A. Methodology

1) Systems Tested: We evaluate four different systems: HCTA with the tension arc shown in Figure 3 (HCTA+P, where the P stands for "parabolic", the shape of the tension arc), HCTA with a random tension arc (HCTA+R), the Occupancy-Regulated Extension generator (ORE) [24] which was the winner of the 2011 Mario AI Competition (MAIC), and NLG. We wish we could evaluate more systems in our experiment (e.g., other entries of the MAIC), but that would substantially increase the time required to run the experiment and would also require more subjects. We chose to use ORE because it was the 2011 MAIC winner and its code was available online.[1]

2) Evaluated Metrics: The systems were evaluated according to the following criteria: enjoyment, visual aesthetics, difficulty, and Turing. The Turing criterion was meant to measure whether the participant thought that the level played was designed by a human or machine. In the beginning of the experiment we stated that the participants could be playing levels generated either by humans or by machines.

Each participant was asked to answer how much they agreed or disagreed, in a 7-likert scale, with the following:

1) This level is enjoyable to play.
2) This level has good visual aesthetics.
3) This level is difficult.
4) This level was developed by a machine.

A score of 1 for enjoyment and visual aesthetics mean that the participant strongly agrees that the level played is enjoyable and has excellent visual aesthetics; a score of 1 for the Turing criterion means that the participant strongly agrees that the level was designed by a machine; finally, a score of 1 for difficulty means that the participant strongly agrees that the level is difficult.

3) Participants: Our within-subject experiment had 34 participants: 30 males and 4 females with an average age of 23.73 and standard deviation of 4.31. The experiment was carried out online. Namely, we made our system available in the Internet and advertised our experiment in different mailing lists. Participation was anonymous and volunteered.

4) Experimental Design: In the beginning of the experiment the subjects filled a questionnaire informing their age, and their skills in the game of Mario (i.e., how much Mario they played before). Subjects were instructed about the controls of the game before playing a practice level. The practice level is important for participants to get acquainted with the keyboard control. The practice level was generated by the NLG system. Only after playing the practice level that the participants evaluated the levels generated by the PCG systems. After playing each level the participants gave scores according to the criteria described above in a 7-likert scale. In addition to the scores, the participants had the option to enter comments justifying their scores, informing us of technical issues they might have had, or making general suggestions on the experiment. Since all participants played levels generated by the four tested systems, we used a balanced Latin square design to counteract ordering effects.

In order to have a fair comparison of the levels generated by different systems we had all systems generating levels of the same size: 160 × 15. We chose such size because we did not want the experiment to be too long. In total each participant played 5 levels (1 practice level and 4 other levels for evaluation), and we could not afford creating larger levels as it could be tiring for the participants.

Moreover, we controlled the systems HCTA+P, HCTA+R, and NLG to generate levels with difficulty similar to the levels generated by ORE. This was done in the HCTA approaches by bounding the difficulty value used in the tension arcs, and in the NLG approach by bounding the value of d. We evaluate difficulty in this experiment only to make sure we were able to control such variable on the levels tested.

5) Data Cleaning: The data provided by the participants who were not able to play all levels in our experiment is not included in our results. We also removed the data of a single participant who had never played the game of Mario before. By examining the logs of the experiment we noticed that this participant was not able to get too far into the game and thus not able to properly evaluate the levels. The number of 34 participants was obtained after removing such data.

B. Hypotheses

We are interested in testing if HTCA is able to generate levels with good visual aesthetics and that are enjoyable to

[1] see http://www.marioai.org/LevelGeneration for details.

TABLE I: Empirical evaluation of PCG systems. Lower values of enjoyment and visual aesthetics indicate levels that are more enjoyable to play and have better visual aesthetics; larger values of Turing indicate levels which participants were more prone to believe that were generated by humans.

	HCTA+P	HCTA+R	ORE	NLG
Enjoyment	2.45 +/- 1.87	2.91 +/- 2.04	3.54 +/- 1.93	2.82 +/- 1.98
Aesthetics	2.48 +/- 1.77	2.42 +/- 1.93	3.60 +/- 1.65	3.02 +/- 2.19
Turing	3.42 +/- 2.03	3.22 +/- 2.13	2.71 +/- 2.09	3.17 +/- 2.32
Difficulty	3.45 +/- 1.69	3.22 +/- 2.28	3.08 +/- 1.57	3.71 +/- 1.72

play. We also want to test if the participants are tricked into thinking that the levels the HTCA approach generates are produced by human designers. Finally, we are also interested in testing the effect of different tension arcs in HCTA.

Specifically, we test the following hypotheses:

H1 On average, the HCTA+P and HCTA+R approaches generate levels which are more enjoyable to play than the levels generated by the other approaches tested.

H2 On average, the HCTA+P and HCTA+R approaches generate levels with better visual aesthetics than the levels generated by the other approaches tested.

H3 On average, the HCTA+P approach is better at tricking players into thinking that the levels were produced by humans.

H4 Different tension arcs can influence how much enjoyment the player has while playing levels generated by HCTA.

C. Quantitative Results

The mean results +/- the standard deviations of our experiment are shown in Table I. Shapiro-Wilk tests show that our data is not normally distributed ($p<.0001$ for all criteria). Thus, we use the non-parametric Friedman test which shows a significant difference on enjoyment ($\chi^2(3)=8.18$, $p<.05$) and on visual aesthetics ($\chi^2(3)=9.18$, $p<.05$) across different systems; there was no statistical significance for Turing and difficulty. The small difference in the difficulty scores is an evidence that difficulty was indeed controlled in our experiment, allowing a fair comparison of the different approaches. We now turn to post-hoc tests (Wilcoxon signed-rank) of the systems with respect to enjoyment and visual aesthetics.

1) Pairwise Comparison on Enjoyment: HCTA+P generates levels which are significantly more enjoyable to play than the levels HCTA+R generates ($p<.05$) and the levels that ORE generates ($p<.005$). The levels HCTA+R generates are significantly more enjoyable to play than the ones ORE generates ($p<.05$). Finally, the levels NLG generates are significantly more enjoyable to play than the ones ORE generates ($p<.05$).

2) Pairwise Comparison on Visual Aesthetics: HCTA+P generates levels with significantly better visual aesthetics than the levels ORE generates ($p<.05$) and than the levels NLG generates ($p<.05$). Also, HCTA+R generates levels with significantly better visual aesthetics than the levels ORE generates ($p<.01$) and than the levels NLG generates ($p<.05$).

D. Discussion of the Quantitative Results

1) Testing H1: In general we observe that the participants enjoyed playing the levels generated by all systems. For example, a score of two and three for enjoyment means that the participant agrees and somewhat agrees, respectively, that the level is enjoyable to play. HCTA+P was the system which generated levels which the participants enjoyed playing the most, with a score of 2.45. ORE was the system which generated levels which participants enjoyed playing the least. The average score for enjoyment for ORE was of 3.54, which is the closest to the score of 4 which means that the participant neither agrees nor disagrees with the statement that the level is enjoyable to play.

The levels HCTA+P generates are significantly more enjoyable to play than the levels generated by HCTA+R and ORE. The difference between HCTA+P and NLG was not significant in our experiment. These results partially support **H1** in the sense that both HCTA+P and HCTA+R generated levels which were more enjoyable to play than those generated by ORE, but no statistical difference was observed with respect to the levels NLG generates. We conjecture that the issue regarding the level size and enjoyment discussed in Section VI-B played a role in the results.

2) Testing H2: Similar to the results on enjoyment, in general the participants liked the visual aesthetics of the levels generated by all systems. HCTA+R was the system with best score (2.42) and HCTA+P was the second best (2.48). Both ORE and NLG had an average score above 3. Both HCTA systems generated levels with significantly better visual aesthetics than ORE and NLG. These results support **H2**.

3) Testing H3: The system which performed best in tricking people into thinking that the level was developed by a human designer was HCTA+P, followed by HCTA+R, NLG, and ORE. The average score of 3.42 for HCTA+P means that people somewhat agree that the level was developed by a machine. Also, the differences on the Turing score values were not statistically significant. Our results do not support **H3**.

4) Testing H4: Enjoyment was the only criterion in which there was a significant difference between the HCTA approaches. HCTA+P was slightly better than HCTA+R in the Turing criterion and it produced levels which were slightly easier than HCTA+R. However, these results were not statistically significant. The average score of visual aesthetics for both systems is nearly the same. We conjecture that the perceived visual aesthetics depends only on what the player sees on the screen at a given time. This is in contrast with enjoyment, where there could be a relation with the level's structure as discussed in Section VI-B.

It is interesting to observe that the tension arc shape can influence the enjoyment the players experience. While it is standard to have tension arcs depicting growing tension in storytelling [17], it was not immediately clear to us that such a tension arc could also have a positive impact in IMB. This result fully supports **H4**. While the tension arc shown in Figure 3 performed better than the random tension arc, we do not know whether there are better tension arcs to be used to generate levels in platform games or even if the tension arc shape should be player-dependent. Investigating these issues are interesting directions of future work.

E. Qualitative Results

Written evaluation was optional in our experiment and only a few participants entered comments about the levels played. Here we highlight a few of such comments. A participant noticed the carry-on effect of playing multiple levels: *"this level seems to be easier than the first I played, maybe it is because I am getting used to the game"*. Although the ordering effect exists, as mentioned by the participant, an indication that our balanced Latin square design counteracted such effects is the small difference in the difficulty scores (see Table I).

Another participant wrote about a level generated by ORE: *"The randomness of how the objects are placed on the screen make me believe this level was generated by a machine."*. and *"The aesthetics isn't good because there are objects in places where they aren't needed."* A level generated by ORE also received a compliment: *"If this level was developed by a machine, then this is really good!"*. Another participant wrote about a level generated by HCTA+P: *"Despite the large number of enemies, I believe that the structure of the level was built by a human designer"*.

IX. Conclusions

In this paper we introduced HCTA, a PCG system for IMB which uses human computation to evaluate the content generated. HCTA uses an existing PCG system to generate a large number of small levels which are evaluated by human workers. Then, HCTA uses the mathematical model of tension arcs to combine a number of small annotated levels into a larger level of the game. We also performed a systematic experiment with human subjects to evaluate levels generated by the proposed approach and also by other systems. Our results showed that (i) the levels generated by the HCTA approaches had better visual aesthetics than the levels generated by all other schemes tested on the average case; (ii) the levels generated by the HCTA approaches were more enjoyable to play than one of the other schemes tested on the average case.

Our results suggest that the human-in-the-loop approach is feasible for the task of PCG in platform games and that such an approach can produce content of good quality. Finally, we believe that the HCTA approach is general to be applied to other Mario-like platform games such as Sonic the Hedgehog.

X. Acknowledgements

This research was supported by CAPES, CNPq, FAPEMIG, Science Without Borders, and Gapso. We also thank Rob Holte, Noam Tractinsky, and Nathan Sturtevant for fruitful discussions on earlier drafts of this paper.

References

[1] J. Togelius, G. N. Yannakakis, K. O. Stanley, and C. Browne, "Search-based procedural content generation: A taxonomy and survey," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 3, no. 3, pp. 172–186, 2011.

[2] M. Hendrikx, S. Meijer, J. Van Der Velden, and A. Iosup, "Procedural content generation for games: A survey," *ACM Transactions Multimedia Computing, Communications. Applications*, vol. 9, no. 1, pp. 1:1–1:22, 2013.

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

[4] N. Shaker, G. N. Yannakakis, and J. Togelius, "Towards automatic personalized content generation for platform games," in *Conference on Artificial Intelligence and Interactive Digital Entertainment*. AAAI Press, 2010, pp. 63–68.

[5] A. M. Smith, C. Lewis, K. Hullet, and A. Sullivan, "An inclusive view of player modeling," in *Proceedings of the International Conference on Foundations of Digital Games*. New York, NY, USA: ACM, 2011, pp. 301–303.

[6] K. Compton and M. Mateas, "Procedural level design for platform games." in *Conference on Artificial Intelligence and Interactive Digital Entertainment*. AAAI Press, 2006, pp. 109–111.

[7] J. Togelius, R. De Nardi, and S. M. Lucas, "Making racing fun through player modeling and track evolution," in *Proceedings of the SAB Workshop on Adaptive Approaches for Optimizing Player Satisfaction in Computer and Physical Games*, 2006.

[8] A. Liapis, G. N. Yannakakis, and J. Togelius, "Towards a generic method of evaluating game levels." in *Conference on Artificial Intelligence and Interactive Digital Entertainment*. AAAI Press, 2013.

[9] N. Sorenson, P. Pasquier, and S. DiPaola, "A generic approach to challenge modeling for the procedural creation of video game levels," *IEEE Transactions on Computing Intelligence and AI in Games*, vol. 3, no. 3, pp. 229–244, 2011.

[10] N. Shaker, M. Nicolau, G. N. Yannakakis, J. Togelius, and M. O'Neill, "Evolving levels for super mario bros using grammatical evolution," in *Proceedings of the Conference on Computational Intelligence and Games*. IEEE Press, 2012, pp. 304–311.

[11] A. J. Quinn and B. B. Bederson, "Human computation: A survey and taxonomy of a growing field," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1403–1412.

[12] J. Togelius, N. Shaker, S. Karakovskiy, and G. N. Yannakakis, "The mario AI championship 2009–2012," *AI Magazine*, vol. 3, no. 34, pp. 89–92, 2013.

[13] M. Mateas and A. Stern, "Façade: An experiment in building a fully-realized interactive drama," in *Game Developers Conference*, 2003.

[14] G. Smith, J. Whitehead, and M. Mateas, "Tanagra: a mixed-initiative level design tool," in *Proceedings of the International Conference on Foundations of Digital Games*. ACM, 2010, pp. 209–216.

[15] G. Smith, M. Cha, and J. Whitehead, "A framework for analysis of 2D platformer levels," in *ACM SIGGRAPH Symposium on Video Games*. ACM, 2008, pp. 75–80.

[16] N. Shaker, G. Yannakakis, and J. Togelius, "Crowdsourcing the aesthetics of platform games," *IEEE Transactions on Computational Intelligence and AI in Games*, vol. 5, no. 3, pp. 276–290, 2013.

[17] C. Vogler, *The Writer's Journey: Mythic Structure For Writers*. United States: Michael Wiese Productions, 2007.

[18] L. von Ahn, B. Maurer, C. Mcmillen, D. Abraham, and M. Blum, "recaptcha: Human-based character recognition via web security measures," *Science*, vol. 321, no. 5895, pp. 1465–1468, 2008.

[19] J. Togelius, S. Karakovskiy, J. Koutnik, and J. Schmidhuber, "Super mario evolution." in *Proceedings of the Conference on Computational Intelligence and Games*. IEEE Press, 2009, pp. 156–161.

[20] P. Piselli, M. Claypool, and J. Doyle, "Relating cognitive models of computer games to user evaluations of entertainment." in *Proceedings of the International Conference on Foundations of Digital Games*, J. Whitehead and R. M. Young, Eds. ACM, 2009, pp. 153–160.

[21] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit formation," *Journal of Comparative Neurology and Psychology*, vol. 18, pp. 459–482, 1908.

[22] G. Smith, M. Treanor, J. Whitehead, M. Mateas, M. Treanor, J. March, and M. Cha, "Launchpad: A rhythm-based level generation for 2d platformers," *IEEE Transactions on Computing Intelligence and AI in Games*, vol. 3, no. 1, pp. 1–16, 2011.

[23] B. Horn, S. Dahlskog, N. Shaker, G. Smith, and J. Togelius, "A comparative evaluation of level generators in the mario ai framework." in *Proceedings of the International Conference on Foundations of Digital Games*. ACM, 2014.

[24] P. A. Mawhorter and M. Mateas, "Procedural level generation using occupancy-regulated extension." in *Proceedings of the Conference on Computational Intelligence and Games*. IEEE Press, 2010, pp. 351–358.