An Experiment on Game Facet Combination

Raphael Patrick Prager
Department of Information Systems
University of Muenster
Münster, Germany
raphael.prager@wi.uni-muenster.de

Laura Troost
Department of Information Systems
University of Muenster
Münster, Germany
l_troo01@uni-muenster.de

Simeon Brüggenjürgen
Department of Information Systems
University of Muenster
Münster, Germany
s_brue33@uni-muenster.de

David Melhart
Institute of Digital Games
University of Malta
Msida, Malta
david.melhart@um.edu.mt

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

Mike Preuss
LIACS
Universiteit Leiden
Leiden, Netherlands
m.preuss@liacs.leidenuniv.nl

***Abstract*—Procedural Content Generation of game content has been vastly improved over the last years and is more and more adopted also in the game industry. It relies mostly on evolutionary and related optimization methods but usually only treats a single of the many available facets as visuals, levels, audio, etc. The problem of how to combine several facets of generation is largely unsolved, but nevertheless very important. One of its subproblems is that we currently do not know in advance how users will react to machine-generated combinations. Based on a simple maze game with exchangeable visuals and audio styles we test how users receive 'usual' and 'unusual' facet compositions by means of rank trace based annotations of their own play-throughs. By means of machine learning techniques, we establish a model in order to learn and predict user reactions. Understanding the effects of facet composition on the user is fundamental if we want to rise evolutionary generation of content to the next level.**

***Index Terms*—game facets, user study, procedural content generation**

I. Introduction

Driven mostly by evolutionary optimization algorithms, Procedural Content Generation (PCG) is nowadays applied for generating many different aspects of games [1]. More precisely, so-called game facets describe high-level components of a game such as visuals, audio, narrative, game design, level design or gameplay [2]. While it is often feasible to generate one facet of a game – for example the selected mazes of this experiment were at first procedurally generated – it is of major challenge to combine multiple (by independent algorithms) generated facets in a way that the resulting game is perceived as harmonic by a human player [3].

The process of harmonizing the facets of a game has been coined as facet orchestration [4]. The definition is inspired by a musical orchestra where each instrument needs to be played carefully to create a comprehensive and harmonic interaction. If one instrument is out of tune, this can be easily noticeable and may create an interfering experience for the listener.

Facet orchestration would be a necessity for a general game generator. Probably due to the complexity of the task, the amount of approaches in this area is currently limited [5]. A recent survey [4] introduces the problem of multiple facet generation and provides an overview of attempts to combine generation for different facet combinations. One of the difficulties here is that it is hard to foresee how human users will react to content combinations. Will they really only accept harmonic matches, e.g. horror sounds with horror visuals? Or will they also willingly accept seemingly non-matching combinations? Is this shift rather gradual or pivotal? With this study, we attempt to shed more light on how users perceive different compositions of multi-faceted game content in order to obtain directions for multi-faceted content generation. From our point of view, this problem is relevant not only to research but to practitioners as well. One could easily imagine that comforting visuals paired with tensed sound might elicit an even more distressing game environment for the player as he expects something to happen. This is only one of the plethora of possible combinations of facets and facet manifestations

We have therefore set up a simple maze runner game with two exchangeable sets of visuals and audio content, respectively. One set of each facet resembles a *horror* scenario, and the other set a *happy* scenario. In our study, the content is manually designed, but it may also have been generated separately for each scenario. The question is, how humans will react if, for a single facet, the content is swapped to resemble the other scenario.

Based on our expectations concerning user reaction towards the different combinations of happy and horror facets, we are interested in testing the following hypotheses:

- Users generally prefer homogeneous combinations (happy with happy, horror with horror) over heterogeneous ones.
- If at all, the video facet is more important than the audio facet concerning user reactions.
- The facet combinations are so different that a model predicting the effect of one combination on the player cannot generally be used for another combination.

978-1-7281-1884-0/19/$31.00

The main contribution of this work is to provide more insight into the interaction between facet combination and user reaction. With respect to facet combination, a better understanding of this interaction would clearly be beneficial for better combining generated game content from multiple generators in a way that is embraced by human players.

In Section II and III, we describe the maze runner game used for the study in more detail and elaborate on the experimental setup. Thereafter, we present the analysis of the obtained information. This comprises an exploratory data analysis of the study results. In Section V, a machine learning approach is introduced which is able to scrutinize derived insights of the former section. The outcomes are presented in Section VI. This is followed by a discussion where we relate our results to the field of facet orchestration. Finally, we summarize our findings with respect to the hypotheses.

II. The Maze Runner Game

The game we use for collecting data falls into the category of maze runner games. Inspired by this genre, the general objective is to escape a maze in a specific time frame. A single maze consists of exactly one entrance and one exit. While dead ends are incorporated, loops are excluded. Hence, the mazes to be solved are called *simply connected* or *perfect* mazes.

At first, a player can try to escape the labyrinth unbothered. After a given amount of time, an enemy spawns at the entrance point. Using the A*-algorithm to construct the shortest path [6], the enemy aims to reach the player. If that occurs, the player will die, and thus loses the game. In addition, so-called *action chests* are located within the maze. A chest can simply be triggered by running against it. Thereby, an effect is activated that is not known beforehand. Exemplary chest events are teleportation to a predefined location within the maze, jump scares and so forth. A full account of chest types and their different effects is provided in Table I.

The game contains a visual representation for the previously mentioned chests and the enemy but also for basic elements as the walls of the maze, the floor, and the players avatar. These visual representations are accompanied by a soundscape. This soundscape is constituted of background music, and sound effects for certain chests and other game events, for instance, the spawn of an enemy. For both facets, visual representation and soundscape, two different settings are available, i.e., a selected visual representation can either convey a dark and tensed, or a soothing and joyful game environment. Figure 1 provides two screenshots of these representations in the game. On the left-hand side, the visual representation is set to *horror*. The right-hand side depicts a *happy* visual representation. The same applies to the soundscape. For the sake of simplicity, we label the former as the *horror setting* and the latter as the *happy setting*. The game allows combining the different settings of visuals and soundscape in an arbitrary manner. In other words, visual representation and soundscape can either complement (e.g. both are set to *horror*) or contradict each other (e.g. soundscape is set to *horror* while visuals are set to *happy*).

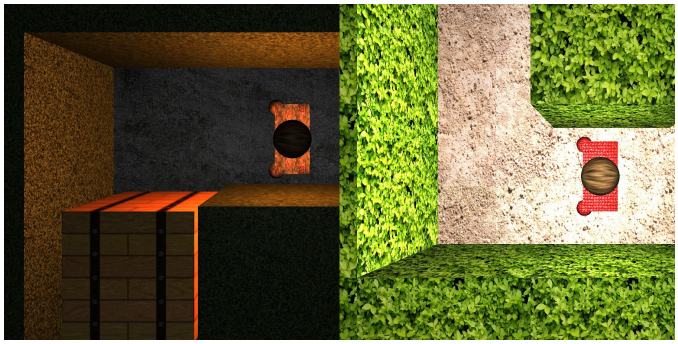

Fig. 1. Different visual representations. Left: player in a *horror* environment, next to an action chest. Right: a *happy* setting.

To provide a consistent language, we will use the following notation in the remainder of this paper when referring to the setting. The set of possible visual representations is denoted as $V = \{\text{horror}, \text{happy}\}$ and the soundscape respectively as the following $S = \{\text{horror}, \text{happy}\}$. The setting configuration sc of a game instance is captured in the tuple $sc = (v, s)$ with $v \in V$ and $s \in S$. Hence, a game configuration which uses *happy* visual representations and a *horror* soundscape is denoted as (happy, horror).

To summarize the aforementioned game elements and concepts, a player's objective is to escape the maze within the given time frame and without being caught by the enemy, by either utilizing chests to his advantage or not, while being subjected to the different configurations of visual representation and soundscape[1].

III. Experimental Setup

In order to measure how human test players perceive the different combinations of facets, we perform an experiment that provides us with different types of information. Next to some demographic data, we obtain annotated play-throughs of different facet combinations together with preference rankings also provided by the players.

Whereas the annotated play-throughs are modeled via rank based SVM later on, we first separately analyze the rankings done by the players concerning the different facet combinations. However, we now start with describing how the experiment was performed.

The survey was conducted with 20 participants. The rules and game elements were explained beforehand. A tutorial allowed each participant to get acquainted with the game for five minutes. This enables players to familiarize themselves with the controls. In this tutorial, visual elements, soundscape, and game elements were reduced to a minimum, i.e., no enemy, chests and sound effects were present. Visuals consisted only of the player, the ground and walls. The reasoning for this minimal audiovisual representation was to avoid introducing a biased perception as much as possible. In other words, when

[1]The game can be found here: https://github.com/MS-Lolstars/PCGMaze. Note that depending on your location the first level might lag in the beginning until all resources are loaded

TABLE I
CHEST TYPES AND THEIR EFFECTS

Chest Type	Effect
Rotation	Rotates the camera by either 90, 180 or 270 degrees. This effect holds on until the current maze is finished.
Teleportation	Instantly moves a player to a predefined location within the maze.
Reduce Time	The amount of time to escape the maze is reduced by six seconds while a typical game takes about one minute to finish.
Increase Time	The amount of time to escape the maze is increased by six seconds while a typical game takes about one minute to finish.
Zoom out	The camera moves further away from the player. Hence, the player is able to see roughly twice as much of the maze as before. This effect is active for five seconds.
Horror Jump Scare	One out of four different animated jump scares is played. The jump scare covers the entire screen for less than a second and is accompanied by a respective sound effect. This chest type only occurs in combination with *horror* visuals.
Happy Video	This is the antagonist of the horror jump scare. A short animated video is played. The content comprises what we define as possibly funny or 'sweet', e.g., animal videos. During that, the countdown timer is paused. This chest type only occurs in combination with *happy* visuals.

a participant would play a specific setting configuration in the non-monitored tutorial, he might rate this setting configuration differently in the monitored survey. After playing the tutorial, each participant began the survey.

A single survey consists of four iterations of a specific procedure. This procedure is constituted of three distinct phases: a *playing phase*, an *annotation phase*, and a *rating phase*. In the playing phase, the survey participant tries to escape a given maze. During this play-through, valuable information of the game state is written to a corresponding log file. The log file comprises information regarding the player and enemy position in a 250 milliseconds interval, which sound effects occurred at a specific point of time, which chests were opened, and whether the game ended successfully or not. In addition, the entire play-through is recorded. That means the screen was captured using the Windows Gaming Bar. This video is used in the annotation phase. During this second phase, participants were presented their recording and instructed to annotate their video in terms of arousal. Arousal is a well established emotional dimension which was already the focus of different research [7]–[9]. We define arousal as the product of one or multiple endogenous stimuli. This can be a positive emotion induced by certain game elements, for instance, joy and laughter. But also emotions which typically have a negative connotation, like stress, pressure, panic, and fear. In either case, when the degree of overall stimuli rises, participants are obliged to increase the arousal value. Respectively, when the game cannot uphold the strength of stimuli, participants are presumed to decrease the arousal value. The gradient of change can be freely chosen by a participant. The initial arousal value is zero, and there are no lower or upper bounds. In addition, the produced values are discrete. This results in a time series of arousal values, meaning for a given participant and configuration c we gather the arousal value in the beginning ($t = 0$), the arousal value in $t = 1$, and so forth. When a participant does not explicitly assign an arousal value in t, the last arousal value in $t \quad 1$ is usedThese arousal values are collected using an external hardware component, *Griffin Power Mate*[2], in combination with the software called *Ranktrace*[3]. According to [10], the unbounded annotation via Ranktrace complemented by a wheel-like external hardware component, such as the Griffin Power Mate, yield good results when measuring arousal or similar emotions. The results of the annotation phase are matched with the log files of the playing phase to deduce meaningful insights. In the rating phase, participants compare a given maze with a previous maze played. It is of interest here which setting configuration is perceived as more 'fun' and more 'difficult' over others. Hence, we capture these preference decisions using a three alternative forced choice questionnaire schema. In other words, in case of 'fun', participants have to choose whether a given setting configuration was 'more', 'less' or 'same' fun as the previous one. This also applies to the question concerning difficulty.

This procedure is repeated for the different setting combinations. The available setting configurations $sc \quad 2 \quad SC$ with set $SC = f(\text{horror}, \text{horror}), (\text{happy}, \text{horror}), (\text{horror}, \text{happy}), (\text{happy}, \text{happy})g$ theoretically allows for $\sum_{i=1}^{|SC|-1} i = 6$ different comparisons between two setting configurations. To improve statistical relevance, only four setting comparisons are considered within this paper. These are namely:

- (horror, horror) versus (happy, horror)
- (happy, horror) versus (horror, happy)
- (horror, happy) versus (happy, happy)
- (happy, happy) versus (horror, horror)

On the one hand, this weakens the independence of the measured results because there are combinations not tested. On the other hand, this prevents the measured results to degrade due to fatigue and keeps the amount of data large enough to apply statistical procedures. With a larger test set (one participant needs about half an hour, not including the training level), a more complete comparison would have been possible.

As already mentioned, the described procedure, comprising the three phases, is conducted four times for a single survey participant. To ensure comparability between participants, four different mazes were designed beforehand. In each of these mazes, several chests with predefined effects were integrated. In the first iteration, the **initial** setting configuration as well as one of the predefined mazes are **chosen at random**. The setting of the second iteration is determined by the setting of the first iteration, i.e., if the first iteration used the (horror,

[2]https://support.griffintechnology.com/product/powermate/

[3]http://pagan.davidmelhart.com/upload.php

horror)-configuration, the second iteration will always attain the (happy, horror)-configuration (c.f. the listing above). The order of mazes on the other hand is not fixed, rather they are sampled uniformly. The only constraint enforced is that a maze cannot be played again within one survey (so sampling without replacement). For the rating phase, this leads to three different setting comparisons per survey, that is, (1) comparison between the first and second setting combination played, (2) comparison between the second and third setting combination played, (3) and finally the comparison between the third and fourth setting combination. Note that the first iteration of the survey procedure only consists of the playing phase and the annotation phase. The rating phase is omitted simply because there is no previous game played to compare it to.

IV. Data Analysis

First of all, the dataset was split into several parts. Splitting criteria were the four different mazes as well as demographics, like gender and age. Thereafter, we checked if these factors had an impact on the effect we want to measure, and whether we should divide the dataset into several smaller ones for further analysis. That was not the case. In addition, each setting combination was approximately equally often the first configuration of a survey. Thereby, we can assume that an external bias is minimized to the best of our knowledge. Regardless, the dataset is divided into two subsets due to their distinct nature. The first dataset consists of the information of the rating phase extracted from the questionnaire. The second dataset comprises the matched game logs and the arousal annotations which have a time-series character.

A. Questionnaire Data

As a reminder, each survey comprises three rating phases where a participant compares the current level with the previous one. One of the questions is directed at the enjoyability, i.e., is the current level more enjoyable (more 'fun') than the previous one? The results of these comparisons between the different setting combinations are depicted in Figure 2. Players substantially prefer the heterogeneous (horror, happy)-setting over the heterogeneous (happy, horror)-setting. This also applies to the comparison of the (horror, happy)-setting against the (happy, happy)-setting. The (happy, happy)-setting is favored in that case. In the remaining two comparisons, that is, (horror, horror) vs. (happy, horror), and (happy, happy) vs. (horror, horror), none setting configuration clearly dominates the other. The (horror, horror)-setting is only slightly preferred over the (happy, happy)-setting and the (happy, horror)-setting.

However, rankings can still be extracted but have to be regarded with caution. Assuming a scenario where we have two objects a and b. For arbitrary reasons, if a has a better rank than b, then this is denoted as $rank(a) < rank(b)$. Consequently, the aforementioned preferences of setting combinations can be written as follows:

- $rank$((horror, horror)) $<$ $rank$((happy, horror))
- $rank$((happy, horror)) $>$ $rank$((horror, happy))
- $rank$((horror, happy)) $>$ $rank$((happy, happy))
- $rank$((happy, happy)) $>$ $rank$((horror, horror))

This ranking is transitive. That means that the homogeneous settings are preferred over the heterogeneous settings in any of the four considered comparisons.

At the same time, participants perceived the heterogeneous settings as more difficult compared to the homogeneous setting combinations. An external bias induced by the different mazes is most likely not the case. All four mazes were roughly perceived as similarly difficult. This trend is illustrated in Figure 3. It is likely that the heterogeneous combination of visual elements and soundscape acts as a catalyst for confusion, and is therefore a possible reason for an increase in perceived difficulty. However, these rankings are of relative nature. While heterogeneous setting combinations might be perceived as more difficult, that does not necessarily mean that the (horror, horror) and (happy, happy)-configuration are too simple.

B. Arousal Data

Besides the dataset of the questionnaire, the arousal dataset is also utilized for analysis purposes. Of special interest is here the density of the arousal values distinguished for each setting combination. The hypothesis (derived from the questionnaire data) that homogeneous setting combinations are preferred over heterogeneous ones seems to be present in the arousal data as well. Figure 4 shows the mirrored density of each audiovisual configuration. The (happy, happy)-setting and the (horror, horror)-setting exhibit both a relative stable density with fatter tails in the direction of increasing values. The heterogeneous settings on the other hand cover a larger range of values on the y-axis. While these violin plots do not illustrate the time series of arousal values, it can still be inferred that the increase and decrease of arousal in the homogeneous settings is more smooth. This holds especially true for the (horror, horror)-setting. In contrast, the (happy, horror)-setting and the (horror, happy)-setting exhibit more rugged properties. To confirm the intuition that each density is different, the Kolmogorov-Smirnov test is used. H_0 assumes an equality of distribution. In every pairwise comparison (e.g. (horror, horror) and (happy, happy)), H_0 can be rejected with a significance level $\alpha = 0.01$

Surprisingly, the analysis of the chest types in context of different audiovisual aesthetics does not yield any insights. Especially in the case of the *Horror Jump Scare* and the *Happy Video* chest, we expect differences in the distinct setting combinations which are not present. While the absolute arousal

TABLE II
Arithmetic mean and median of arousal for the different setting configurations.

Setting configuration	Mean	Median
(horror, horror)	37:27	9
(happy, horror)	17:31	2
(horror, happy)	25:41	2
(happy, happy)	24:50	4

Fig. 2. The four considered comparisons between two different setting configurations. The title of each bar plot indicates the order of settings played. For instance, (horror, horror) vs. (happy, horror) means that the previous maze had the (horror, horror)-setting whereas the current maze consists of the audiovisual aesthetics (happy, horror). The x-axis depicts the three different possible answers to the questions 'How much did you enjoy the current maze compared to the previous?'. The y-axis represents the number of answers by participants. Note that due to the randomized experimental setup the number of total answers to each question varies slightly.

Fig. 3. This figure is closely connected to Figure 2. The same description of that figure applies here. The only difference is the asked question: Whereas the previous figure illustrates the enjoyability of the setting combination, the information of this figure targets the perceived difficulty. The question was 'How difficult is the current maze compared to the previous?'.

values are divergent and cannot be explicitly connected to the chests, the magnitude of change, after a chest is activated, is approximately the same.

Besides the information depicted in the violin plots, the (horror, horror)-setting has the highest median value in terms of arousal. This is followed by the (happy, happy)-setting. The arithmetic mean on the other hand does not bare any meaningful interpretation due to several extreme outliers in some setting configurations (c.f. Figure 4). However, for full disclosure the arithmetic means and medians of the arousal distinguished for each setting combinations are summarized in Table II.

Nevertheless, the derived insights of the arousal data cannot fully support the hypothesis that homogeneous setting combinations are preferred over heterogeneous ones. One of the measured features of the questionnaire is 'fun' which assumes a positive valence. In contrast, arousal itself can be both, positive and negative, as it only expresses an emotional response in one dimension. Consequently, an increasing amount of arousal does not necessarily mean that it will automatically be preferred by players. However, there exists a correlation between the features fun and arousal. A more gradual and smooth density of arousal seems to imply that it will be more fun for players. This is the case for the homogeneous setting configurations, that is, (horror, horror) and (happy, happy).

V. Ranked SVM based Arousal Model

This section introduces a machine learning approach to model arousal based on the different audiovisual configurations of the maze runner game. Comparative predictive modeling is applied to examine the consistency of user reactions to stimuli within and between audiovisual configurations. Our results provide further insights into the capacity of different facet combinations to influence the emotional outcome of a gaming experience. This approach focuses on the predictive strengths of different models based on given configurations between subjects and across different setups.

A. Datasets and Features

For the machine learning task, the data is preprocessed and divided into different sets based on the configuration of the audiovisual aesthetics of the given session of a player. Both the input and output features are preprocessed and aggregated on a low level. Each element of the gameplay logs and the corresponding annotations are grouped and assigned to three second time windows. This procedure helps to smooth

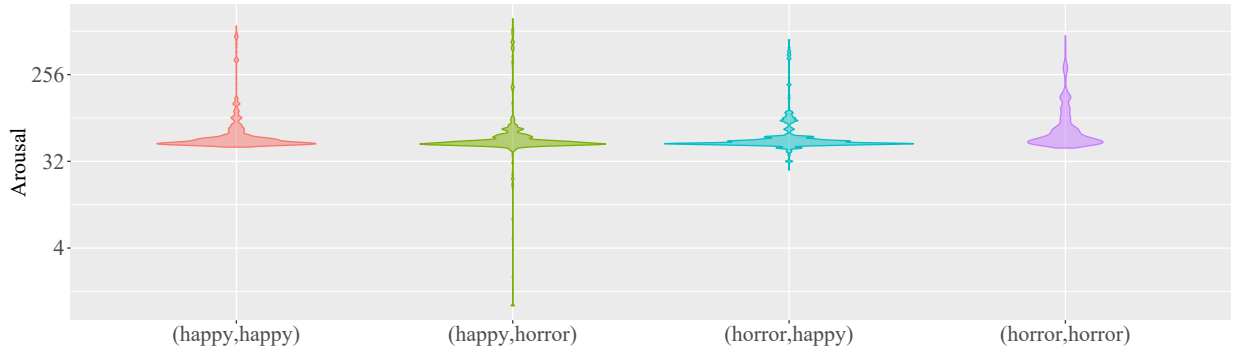

Fig. 4. Each violin plots depicts the density as well as the skewness and kurtosis of a given setting combination (x-axis). The y-axis indicates the arousal value on a logarithmic scale. Arousal values have been linearly transformed beforehand to ensure positive values only. The arousal values are not aggregated, i.e., for each participant and a given configuration the values of the entire time span are taken.

out the data and aligns the input and output features while preserving the moment-to-moment dynamic of gameplay. The whole dataset of 1320 feature vectors is divided into 4 sets of 330 samples for each audiovisual configuration.

Input features are translated from gameplay logs into both general gameplay metrics (*Time Passed*, *Overall Distance Travelled*, *Delta Distance Moved*, *Euclidean Distance to the Starting Location*) and gameplay events related to audiovisual aesthetics (*Horror Jump Scare*, *Happy Video*, *Sound 1: Distant scream*, *Sound 2: Crow call*, *Sound 3: Alien noise*, *Sound 4: Crying baby*), enemies (*Enemy Spawned*, *Enemy Present*, *Enemy Catches Player*), navigation (*Rotation* 180°, *Rotation* 270°, *Teleportation*, *Zoom Out*, *Player Finds Exit*) and time constraints (*Decrease Time*, *Time Runs Out*, *Increase Time*).

The target output for all models is the annotated arousal. This ground truth is processed through different means, which are modeled independently. Six different signals are generated through three processing methods and the same three-second window method which is applied to the input features is employed here as well. For each window, the amplitude ($\hat{A}$), the gradient (A), and mean value (μA) is calculated. For each of these combinations, a second feature is generated as well with the application of a one second lag (l), which shifts the annotation back, potentially aligning them with the gameplay logs to account for the reaction time of the player.

B. Pairwise Preference Learning

Preference Learning (PL) [11] is a supervised machine learning paradigm, in which an algorithm learns to infer the relative association of datapoints. In contrast to classification and regression—which treat data as nominal and interval values respectively—PL regards datapoints as ordinal variables [12], hence it is more flexible [13], [14]. To infer relative relationships PL applies a *pairwise transformation* on the given dataset [15], translating it into a new representation, which can be solved by any binary classifier. During the *pairwise transformation* for each pair of (x_i, x_j) 2 X the preference relation of the variables is observed based on their associated ground truth value (y_i, y_j) 2 Y. If $(y_i > y_j)$, then x_i is preferred over x_j (x_i x_j). For each of these observations, two new datapoints and associated labels are created, signifying the difference between the two datapoints and the direction of the preference relation. In case of x_i x_j, the new datapoints are $x'_1 = (x_i \quad x_j)$ with label $\lambda_1 = 1$ and $x'_2 = (x_j \quad x_i)$ with label $\lambda_2 = \quad 1$. During the transformation a preference threshold parameter (P_t) is applied to control the minimum significant difference between datapoints for the inference of the preference relationship. As a result of the transformation, the training set is inflated to 386 on average.

C. Ranking Support Vector Machines

The present study applies pairwise preference learning through Ranking Support Vector Machines (rankSVM) [16] as they are implemented in the *Preference Learning Toolbox* [17]. This application of the algorithm is based on LIBSVM library [18]. SVMs operate by maximizing the margins of a separating hyperplane, often in higher dimensional feature space [19]. The present study uses both linear kernels to infer the distance between the transformed points. During training, our models rely on the C regularization term, which controls the trade-off between the correctly classified training examples and the width of the margins of the separating hyperplane. RankSVMs are chosen for this study because they have been shown to produce robust results in affect modeling tasks [20], [21].

D. Model Validation

To find the best hyperparameters and get a better picture of the internal consistency of the data, cross-participant validation (CPV) is applied in combination with grid search. In the CPV process, we define folds as separate participants based on their id, and run traditional (leave one out) cross-validation over these folds. We have 20 folds in total, with each fold combining data of 4 play sessions. We find the best C parameter in the 1–100 range (with steps of 10); and P_t in 0.01–0.09 range (with steps of 0.01) range.

For the tests comparing the predictive power of different datasets, we fit the models with the best parameters found for each output to the whole dataset and test them on all of the other three datasets, first separately then combined. Due

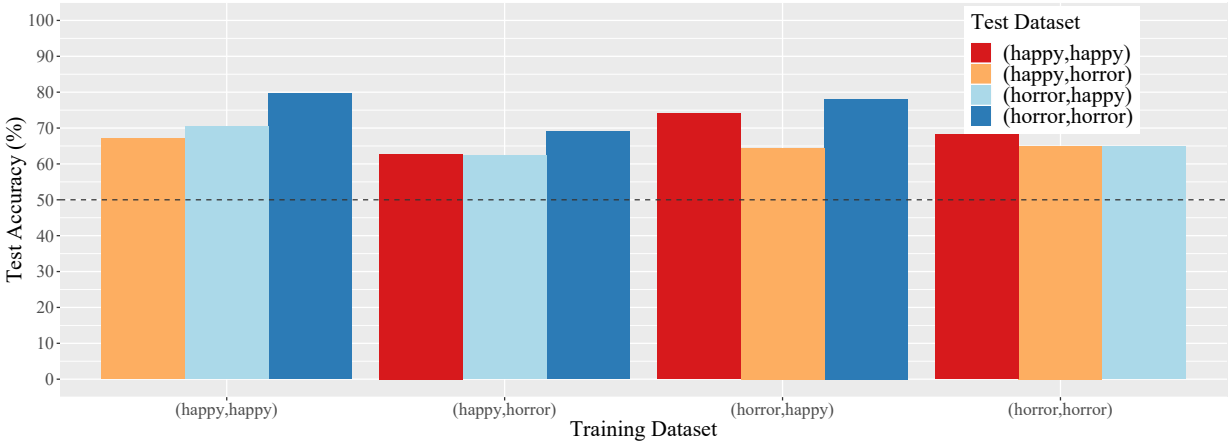

Fig. 5. Test accuracy (y-axis) of the SVM with linear kernel when predicting $A(l = 1)$ arousal. The x-axis indicates which audiovisual constellation was used as a training data set. The different colors represent the results for a given test set. The (happy, happy)-configuration reaches 79.56% accuracy when predicting arousal on a (horror, horror)-setting.

to space constraints, the presented models are always using their best parameters. Since the aim of the predictive modeling is to compare how well models can be trained to perform on different datasets, we compare their best performance and disregard the differences in their hyperparameters.

VI. Predictive Modeling Results

This section displays the results of the rank based SVM modeling and is structured as follows: Section VI-A describes the CPV results briefly, Section VI-B presents the predictive power from one dataset to another, and shows the results of models predicting results from a pooled dataset of three other configurations. Due to the *pairwise transformation* of the dataset, the baseline for all models is 50%.

A. Cross Validation Results

An initial CVP result shows that the internal consistency of homogeneous audiovisual configurations are higher, with the average accuracy of linear SVMs are at 80.73%, 72.02%, 68.45%, 61.37% for the (happy, happy), (horror, horror), (horror, happy), and (happy, horror) setups respectively.

B. Comperative Tests of Single Configurations

To test the robustness of models trained on specific setups, a set of comparative tests is carried out. Although previous tests already show the effectiveness of homogeneous audiovisual configurations, these tests reveal more about the effects between different aesthetic setups. As the results show on Figure 5, models trained with *happy* audio generalize with a higher average accuracy (by approximately 6.8%), while the (horror, horror) setup is the easiest to predict with approximately 8.5% increase in accuracy on these test sets on average.

A final test examines which models generalize better over not just one but all other setups. These results underline and validate prior observations about the robustness of models trained on *happy* audio. Results of (happy, happy) and (horror, happy) configurations show 72.61% and 72.26% accuracy, while (horror, horror) and (happy, horror) only reaching 65.01% and 64.34%.

VII. Discussion

In this study, we performed different types of analysis in order to investigate how users perceive different compositions of multi-faceted game content and how they can be predicted. The results of the exploratory data analysis claim that homogeneous settings are perceived as more fun and less difficult than the heterogeneous ones. This also correlates with the derived arousal values which are higher and more stable for the homogeneous settings. In contrast, arousal values for the heterogeneous settings cover a large range of arousal values which indicates that it might be hard to foresee how players will react to these setting configurations.

The predictive modeling results support this observation as the homogeneous audiovisual facets can be predicted with greater accuracy, especially in case of more intense and transgressive configurations (i.e. (horror, horror)). As for the predictive power of a facet combination, there seems to be a divergence between the combinations. The (happy, happy)-model is able to predict the arousal value for (horror, horror) with an accuracy of nearly 80%. The *happy* audio setting generally performed better than the *horror* audio setting, regardless of the visual configuration. This may indicate that the audio facet has a bigger impact on the player's perception of the game content than the visual facet. The results of predictive modeling extend beyond what the statistical analysis has revealed in two main ways: A) using a pleasant soundscape yields data with a more consistent inter-rater agreement (see configuration-dependent but user-independent model accuracies in Section VI-A) and thus subsequent models generalize better; and B) homogeneous, especially transgressive (horror, horror) audiovisual configurations elicit emotions which are easier to model but not necessarily easier to transfer to less intense input.

Nonetheless, these results must be interpreted with caution and some limitations should be borne in mind. This study only considered a subset of the possible comparisons of two audiovisual configurations, namely four out of six. Additionally, only two facets, audio and visual, were included which only allows for restricted statements about multi-faceted game content. These limitations were endured in order to keep the amount of data large enough for statistical procedures and so that at least some findings can be presented.

VIII. Conclusions

By means of this study, we want to provide more insight into the interaction between facet combination and user reaction. Therefore, we conducted an experiment with a maze runner game and four different audiovisual setting combinations, namely (happy, happy), (happy, horror), (horror, happy) and (horror, horror). In doing so, we derived information about the perceived *arousal* for each game and rankings of the different facet combinations. Based on this, we first performed an exploratory data analysis. The results of this analysis are in line with our first hypothesis, users generally prefer homogeneous combinations over heterogeneous ones, as these were overall perceived as more 'fun'. Additionally, the densities of arousal values were smoother for those setting combinations and exhibit overall larger arousal (in terms of median).

Next, we applied pairwise preference learning with ranking support vector machines to investigate if facet combinations can be used to predict the effect of one combination on the player. As a first result, we observed that the homogeneous settings can be modeled with greater accuracy than heterogeneous ones which supports our first hypothesis. Moreover, there is a high predictive power between some of the facet combinations. This partly contradicts our third hypothesis, facet combinations are generally too different to predict the effect of one combination on the player. Additionally, the audio facet seems to exhibit more predictive power than the visual one which counteracts our second hypothesis that the video facet is more important concerning user reactions. Investigating the role of audio facets in games more detailed might be a promising venture for future research.

Collectively, this study examines the effects of facet compositions on the user. The corresponding results are of valuable insight for PCG with respect to evolutionary algorithms in order to generate games with multiple facets which are perceived as enjoyable by human players. Targeting a specific arousal value, a well performing PL model (consisting of more than two facets and setting configuration) can be used as black-box function for which an evolutionary algorithm aims to find a suitable combination of facets. When we move from the automated generation of individual and independent facets of a game to the harmonized generation of multiple facets using facet orchestration, we are able to reach astonishing advances in the area of PCG, maybe even including a general game generator. This study is only a first step. To move further in that direction, more comprehensive studies with a larger amount of different facets and participants are necessary.

IX. Acknowledgements

This project has received funding from the European Union's Horizon 2020 programme under grant agreement No 787476.

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, Sep. 2011.

[2] A. Hoover, W. Cachia, A. Liapis, and G. Yannakakis, "Audioinspace: Exploring the creative fusion of generative audio, visuals and gameplay," vol. 9027, 04 2015, pp. 101–112.

[3] J. Togelius, A. Champandard, P. L. Lanzi, M. Mateas, A. Paiva, M. Preuss, and K. Stanley, "Procedural content generation: Goals, challenges and actionable steps," *Dagstuhl Follow-Ups*, vol. 6, pp. 61–75, 01 2013.

[4] A. Liapis, G. N. Yannakakis, M. J. Nelson, M. Preuss, and R. Bidarra, "Orchestrating game generation," *IEEE Transactions on Games*, 2019, https://doi.org/10.1109/TG.2018.2870876. [Online]. Available: http://graphics.tudelft.nl/Publications-new/2019/LYNPB19

[5] G. Yannakakis and J. Togelius, *Artificial Intelligence and Games*. Springer, 2018.

[6] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.

[7] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 12 1996.

[8] J. A. Russel, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[9] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions: An emerging approach," *IEEE Transactions on Affective Computing*, 2018.

[10] P. Lopes, G. N. Yannakakis, and A. Liapis, "Ranktrace: Relative and unbounded affect annotation," in *Affective Computing and Intelligent Interaction (ACII), 2017 Seventh International Conference on*. IEEE, 2017, pp. 158–163.

[11] J. Fürnkranz and E. Hüllermeier, "Preference learning," in *Encyclopedia of Machine Learning*. Springer, 2011, pp. 789–795.

[12] G. N. Yannakakis, R. Cowie, and C. Busso, "The ordinal nature of emotions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 248–255.

[13] H. Martinez, G. Yannakakis, and J. Hallam, "Dont classify ratings of affect; rank them!" *IEEE transactions on affective computing*, no. 1, pp. 1–1, 2014.

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

[15] J. Fürnkranz and E. Hüllermeier, "Pairwise preference learning and ranking," in *European conference on machine learning*. Springer, 2003, pp. 145–156.

[16] 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.

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

[18] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *Transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[19] V. Vapnik, "Chapter 5 constructing learning algorithms," *The Nature of Statistical Learning Theory*, pp. 119–157, 1995.

[20] E. Camilleri, G. N. Yannakakis, and A. Liapis, "Towards general models of player affect," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2017, pp. 333–339.

[21] D. Melhart, K. Sfikas, G. Giannakakis, G. N. Yannakakis, and A. Liapis, "A motivational model of video game engagement." *Proceedings of Machine Learning Research, 2018 IJCAI workshop on AI and Affective Computing*, vol. 86, pp. 26–33, in print.