Affect in Spatial Navigation: A Study of Rooms

Emmanouil Xylakis Antonios Liapis, *Member, IEEE,* and Georgios N. Yannakakis, *Fellow, IEEE*

Abstract—How do spaces make us feel? What is the perceived emotional impact of built form? This study proposes a framework to identify and model the effects that our perceived environment can have by taking into consideration illumination and structural form while acknowledging its temporal dimension. To study this, we recruited 100 participants via a crowd-sourcing platform in order to annotate their perceived arousal or pleasure shifts while watching videos depicting spatial navigation in first person view. Participants' annotations were recorded as time-continuous unbounded traces, allowing us to extract ordinal labels about how their arousal or pleasure fluctuated as the camera moved between different rooms. Given the subjective nature of the task and the noisy signals from real-time annotation, a number of processing steps are applied in order to convert the data into ordinal relationships between affect metrics in different rooms. Experiments with random forests and other classifiers show that, with the right treatment and data cleanup, simple interior design features can be adequate predictors of human arousal and pleasure changes over time. The dataset is made available in order to prompt exploration of additional modalities as input and ground truth extraction.

Index Terms—Affective Computing, Virtual Spaces, Lighting, Affect Annotation, Ordinal Modeling, Random Forests

1 Introduction

SPACES and their surrounding structures can have a strong impact on occupants residing within them or just passing through them. Practitioners and researchers of architectural design have long strived to define the psychological impact of the built space [1], [2], [3], [4]. These endeavors have resulted in a plethora of models that map the emotions involved in the process of decoding our surroundings. These models have been revisited and reworked extensively in the post-modern era, when our built tendencies shifted from the functional approaches to more humanistic and environmental design methodologies [5], [6]. These design approaches emphasize human comfort in interior environments, and have shifted our interest to the emotional and well-being capabilities of the built space [7], [8], [9].

The quantitative study of the effect that architectural design has on humans and its potential to deliver the intended experiences is not trivial. Visual perception is inherently a subjective matter and the affective responses to our environments are subject to personal taste, values, cultural influences, experiences and interactions. This emphasis on the subject is commonly referred as "the internalist's view" [10], where the context of the experience plays a more important role than objective facts and descriptors regarding the surrounding environment. To mitigate this subjectivity inherent in all research fields associated with human factors and human behavior, large samples of the population are typically required. In the field of architecture, such user studies fall under the umbrella of co-design or participatory design [11], where the occupant or user of the space becomes a crucial entity feeding into the design of the built artifact in question.

The impact our environment has on our emotions is embedded deep within our survival instincts and has been passed on throughout our evolution, as traits that have kept us safe and at an advantage against the harmful environments that surround us. Under this lens, Hildebrand in his book *Origins of Architectural Pleasure* [12] highlights the mechanisms that make up our spatial emotional reactions and links them to our sense of survival and Darwin's theory of evolution. Similarly, Appleton's concept of prospect-refuge [13] suggests that our perceptual aesthetics are linked to our preference towards what we consider as a safe place, a place that offers a good view to the outside world but can easily be retreated to, in cases of danger.

Given how critical the perception of our surroundings is to our emotion reactions, it is important to identify which perceptible "qualities" matter. Human perception is formed mainly unconsciously [11], but partly conscious affective reactions can occur when we make assessments of our environment (e.g., safety of the space as per the prospect-refuge view). To understand how different design features impact these affective responses, bottom-up approaches tend to be a common practice in architectural and design research by isolating and adjusting the parameters that comprise the environment. In this way, relations between design features and emotions can arise, be processed and quantified. How features of architectural form elicit emotions is less studied than environment variables such as lighting and color [3].

The present study aims to explore—quantitatively—the impact of architectural form on emotions by developing computational models based on theory-driven parameters and metrics. In particular, we combine theories from design and psychology with established practices from affective computing to solicit evidence on the spatiotemporal relationship between form and emotion. This work is an extension of our previous work [14] which studied how arousal changes according to three expert annotators viewing a large dataset of recorded spatial navigation sequences. This paper extends significantly our earlier work in several ways. Firstly, this study introduces a new dataset named *AffRooms12*, which comprises a broader collection of shorter spatial navigation videos that incur less fatigue and cognitive load from participants, compared to the one-

- *All authors are with the Institute of Digital Games, University of Malta. E-mail: {emmanouil.xylakis.21, antonios.liapis, georgios.yannakakis}@um.edu.mt˙*

hour annotation task of [14]. Second, we collect unbounded time-continuous annotations of both arousal and pleasure dimensions, allowing us to study their fluctuations when properties of the built environment change; in contrast, the study of [14] only collected arousal annotations. Third, we solicit feedback from a broader sample of the population, through crowdsourcing platforms, and collect (after data cleanup) 39 participants' arousal annotations and 37 participants' pleasure annotations. These participants were not experts in the annotation protocol used in this study, unlike the three expert annotators of [14]. Fourth, inspired by [15], [16], this paper explores different ways of processing the annotations in order to ensure inter-rater annotator agreement and to derive ground truth labels of affect. This is a significant departure from the methodology of [14] which observed changes between the mean affect in each room. Finally, we go beyond the one-to-one mapping between each design feature and affect reported in [14], and instead leverage supervised learning that combines the different spatial features together to train models of arousal and pleasure from spatial transitions.

2 Related Work

In this section we provide a brief overview of theoretical constructs of affect and ways of capturing it (Section 2.1), followed by an extensive review of studies on the impact of features of the built environment on affect (Section 2.2).

2.1 Capturing Affect

In practice, the ways in which affect states can be captured fall under three categories: self-reports, behavioral measures and measurement of physiological reactions [10], [17]. Self-reports deal with user-aware feedback to the stimuli while the other two record data as a bodily or subconscious result of the elicited affective state in question. While self-reports are easier to design by utilizing the plethora of available emotional models, they are the most challenging to interpret: participants are asked to indicate their direct affect state, making them vulnerable to participant interpretations if not explained carefully on the task in question.

The two main theoretical models that we find in the field of affective computing for capturing affect via subjective input are divided into dimensional models and discrete models [18], [19], [20]. Dimensional models describe emotions usually in a two-dimensional (arousal-valence) or three-dimensional space, with the added dimension of control or dominance. They idea behind dimensional models is that all emotions can be placed within bivariate axes and can be described as coordinates along these dimensions [21], [22], [23]. The alternative approach is to describe emotions in a discrete or categorical manner of emotional labels such as happiness, anger, fear, disgust, etc. [24], [25]. In this study, we follow the dimensional model and focus on the two dimensions of arousal (emotion intensity) and pleasure (valence) in order to solicit annotations of one of these two dimensions from participants during spatial navigation footage. The appeal of the dimensional model for this study is, among other reasons, its ability of representing affect as scalar variables and thus enabling annotators to indicate value changes in a time-continuous manner. Even though dimensional models of affect are known to be sensitive to inter-rater variability due to their continuous nature (a) they offer a more intuitive model for time-continuous annotation and (b) variations in the annotated magnitude can be alleviated using second-order data processing methods [26].

2.2 Affect and Features of the Built Environments

While research has explored many different properties of space [27], [28], [29], we focus on three important, general features of architectural form and one feature of lighting.

Curvature—or the organic-like resemblance as compared to the strict angular property—has been studied both as a property of objects and within architecture, for the shape of a building contour [3], [4] or the shape of perforated facades [30], [31]. Research on this type of design parameter points to a positive preference by participants [32] while other work links this to higher activation when shown to users [3]. The resulting reactions are interpreted as an instinct of threat that angular and rectilinear forms may convey. Vartanian et al. [2] instructed participants to point their preferred choice on the approach-avoidance dimension regarding the curved properties of depicted two-dimensional shapes. Other studies have explored the dimensional model regarding impact of curvature on arousal and pleasure dimensions. Banaei et al. [4] categorized generated rooms into clusters according to their pleasure, arousal and dominance ratings annotated by participants via the self-assessment manikin [22]. They noted that rooms with higher curvature properties result in higher arousal and pleasure while higher angularity in interior forms correlates negatively with both dimensions. The connection to curvature with higher valence and lower arousal has been observed in [33], [34] where angularity and sharp objects result in higher arousal ratings by participants.

Color appearance and materials is another core parameter explored in studies that wish to uncover their perceptual affective impact on inhabitants. Garip et al. [35] studied the reactions to concrete materials in a children's school with an experiment that involved 33 children aged 6-7 viewing the proposed classroom designs in Virtual Reality (VR). The study revealed that the proposed material variations and lighting settings did not produce significant effects, as children at these ages seemed to focus more on identifiable objects around them rather than on the background. Coşgun et al. [36] recruited 298 students to assess wall coverings in wood, metal and concrete (in two tones) in the context of café environments along 11 semantic bipolar dimensions. In this context, light-colored wall coverings were perceived more favorable than dark-colored ones, while light-colored wooden wall coverings were considered a warmer material than concrete and metal. Gomez-Tone et al. [37] recruited 22 students to rate their preference for materials in combination with size, enclosure, degree of perforation and style during VR navigation. Responses evidenced relations between student study year and sensation responses from professional architects. Lastly, Lipson-Smith et al. [38] studied the impact of colored walls in different contexts of virtual living rooms, waiting rooms and empty rooms. Results obtained from 100 participants in a VR experience per room showed significant correlations between two annotation models: a dimensional

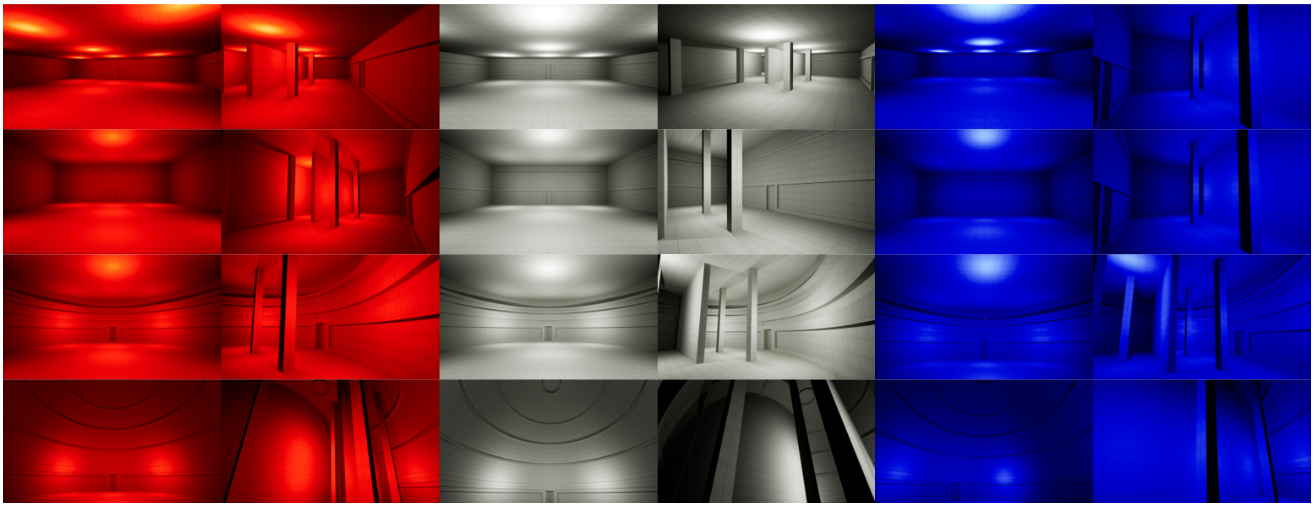


Fig. 1. Views of the 24 rooms examined in the AffRooms12 corpus

annotation for valence and a discrete annotation via the *Pick-A-Mood* scale [24].

Instead of focusing on a specific design feature and adjust it to study participant responses, a number of studies explore the spatial properties as a whole by using the stimulus (e.g. room) in its context. Bianconi et al. [11] sought to measure the impact of virtually constructed prospect spaces in VR on different emotional labels. Spaces such as creativity rooms, study areas, socializing and relaxing rooms were ranked by 40 participants on the emotional labels of careful, pleasant, stimulating, happy, serene, calm, relaxing and satisfying. In that study, researchers collected data regarding a school renovation project in order understand whether the expected reactions for their designs is inline with what the potential users believe. Similar work [2], [39] explored the spatial properties with the use of photographs, where participants must usually choose preferred designs for a given emotion or affect state.

3 AffRooms12 Dataset

Our work hinges on soliciting human annotations of affect during spatial navigation tasks; to achieve this, a diverse set of stimuli representing features of the built environment need to be designed. We build on previous work [14] and identify three dimensions of architectural form and one dimension of lighting in order to produce a diverse set of rooms that one can navigate through. This paper re-uses rooms designed in earlier work for producing the *AffRooms* dataset [14]: a dataset of 20 videos of spatial navigation passing through the same 24 rooms (in different order per video). To lower cognitive load and time requirements for annotating such long videos, in this paper we introduce the *AffRooms12* dataset[1] which contains 55 videos of spatial navigation between 12 different rooms, chosen randomly from the 24 possible room configurations. See Table 1 for dataset properties.

TABLE 1
AffRooms12 Dataset properties

Properties	
Total number of videos	55
Number of different rooms	24
Rooms per video	12
Mean duration per video (secs)	141
Total database duration (hours)	17.2

3.1 Designing Spaces and Recording Navigation

This paper explores four design features in terms of their impact on perceived arousal and pleasure: *contour curvature*, *ceiling height*, *occlusion*, *illumination color*. The first three are dimensions of architectural form, and in our dataset can take two values (feature is present or absent); the last one is a dimension of interior lighting and can take three values (neutral, blue, or red light). Ceiling height can be high (present) or low (absent), while rooms may be empty (occlusion is absent) or populated with two interior walls and 6 columns that block both visibility and the player's path (occlusion is present). When curvature is absent, rooms are rectangular; when curvature is present, the room is cylindrical if the ceiling is low or domed if the ceiling is high. Figure 1 shows all possible room configurations based on these dimensions; more details can be found in [14].

Each video in the *AffRooms12* dataset was produced by randomly selecting 12 room configurations among the 24 possible and building a sequence of rooms procedurally within the Unreal game engine. Every room sequence begins at a small, empty hallway which is connected to the first room through a sliding door (which opens automatically when the player approaches it); each room after this is similarly connected with a sliding door to the room before it. The sliding doors are always placed at opposite sides of the room, and therefore the player can simply walk straight (unless occlusion is present) to access the next room. Once the room sequence is built, the first author of the paper navigated from the starting hallway to the final room and recorded their interaction. The video recording includes pauses to the movement to look around, to give annotators

[1] The dataset is available at https://osf.io/b3ra5

a better view of the surroundings. To allow more reaction time for real-time annotation, raw recorded videos were slowed down by 50%, which did not affect the viewing or navigation experience.

3.2 Design Parameters as Inputs

To assess the impact of space on emotion, we need ways of quantifying properties of the space. For each of the design features explicitly used to build our 24 rooms, it is straightforward to derive features: 0 and 1 for absence and presence of curvature, height, occlusion, and -1 (blue), 0 (neutral), and 1 (red) for illumination color.

When using these metrics to predict affect changes between a room A and a room B (see Section 4), the metrics used as inputs include the four features (curvature, height, occlusion, illumination color) of each room, as well as their difference (Δ). In addition, since navigation times in the recorded videos differ between rooms (especially in the case of rooms with many occlusions), we include in the model's inputs the total duration (in seconds) spent traversing each room.

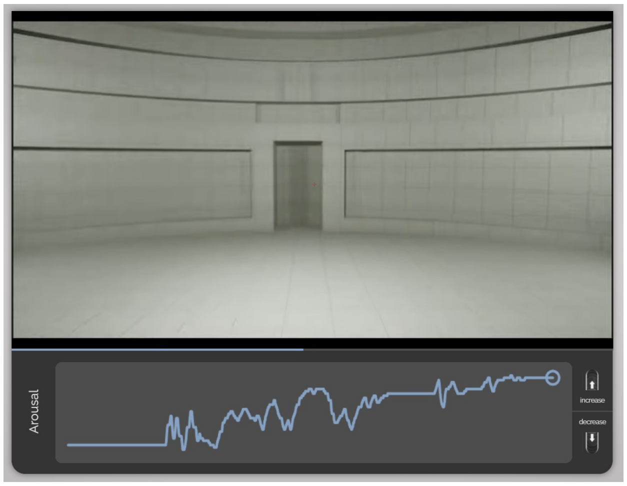


Fig. 2. Screenshot of PAGAN [40] using the RankTrace [19] tool during an arousal annotation task: navigation video (top) and continuous arousal annotation (bottom).

4 Annotation Protocol & Data Collection

Annotations for arousal and pleasure were collected using the PAGAN annotation tool [40]. The tool enables users to report their moment to moment affect changes (up or down) in a single affect dimension. Annotators register their affect in a continuous and unbounded manner using the RankTrace protocol [19], while viewing pre-recorded videos of spatial navigation. RankTrace allows users to define the degree of change of the affect dimension in an unbounded fashion, while showing the users the entirety of this session's annotation so far (see Fig. 2).

We conducted two separate experiments: one for annotating arousal and another for annotating pleasure. Participants of each experiment would annotate a number of pre-recorded navigation videos from the *AffRooms12* dataset (see Section 3) on the same affect dimension each time. At the beginning of the experiment, the user was shown a description of the annotation task. For arousal annotation, the following instructions were provided: *You will be asked to register Arousal changes in the videos that follow by decreasing or increasing the appropriate level. Arousal is the intensity of emotion. Arousal increase means excitement, tension, stimulation, while arousal decrease is connected with boredom, fatigue and/or calmness.* For pleasure annotation instead, we provided the following instructions: *You will be asked to register Pleasure changes in the videos that follow by decreasing or increasing the appropriate level. Pleasure characterizes positive emotions. Pleasure increase is connected with beautiful, exciting, calm, while pleasure decrease describes dull, uncomfortable and/or tense environments.*

4.1 Participants and Data Cleaning

Participants were employed to contribute to the study using the Mechanical Turk crowd sourcing environment by Amazon. The survey was open to individuals regardless of background, ethnicity, gender or any relevance to design and the arts. A total of 100 individuals participated in the annotation task (50 per affect dimension). Each participant annotated 6 videos, selected at random from the 55 in the *AffRooms12* dataset. Ethics approval for the present test protocol was obtained from the University of Malta Research Ethics Committee and participants were informed on the purpose of the study, the use of their contributed data and their option to withdraw their participation at any moment throughout this experiment. A number of steps were taken to clean up the dataset. First, incomplete sessions and duplicates were removed from the dataset. Second, the interactions with the annotation tool are checked: since RankTrace can be used in a continuous fashion, whenever a user interacts with it (e.g. registers a change in affect), we log it. Following the literature on time-continuous affect annotation [41], we remove traces with less than 10 interactions in total (i.e. affect changes), as we consider that the user was idle and not meaningfully interacting with the experiment. Lastly, navigation videos that were not annotated by at least 2 participants are also rejected, as inter-rater agreement is important for the analysis of our data. After this cleanup process, the dataset contains 224 annotated videos of arousal from 39 participants and 215 annotated videos of pleasure from 37 participants. While we specified on Mechanical Turk that participants could only participate on either pleasure or arousal annotation, 5 participants in the clean dataset performed both annotation tasks. However, since the navigation videos are randomly selected every time, it is unlikely that these participants saw the same stimuli in their second annotation session. As per their self-reports, these 71 total participants were in their majority (77%) between 25 and 45 years old, and had completed a university education (82%); 33% identified as female and 67% as male.

5 Data analysis framework

To analyze our data for patterns between affect and room properties, we process the annotations per video. To capture the ground truth of affect reactions to space, we rely on inter-annotator agreement for the same stimulus (navigation

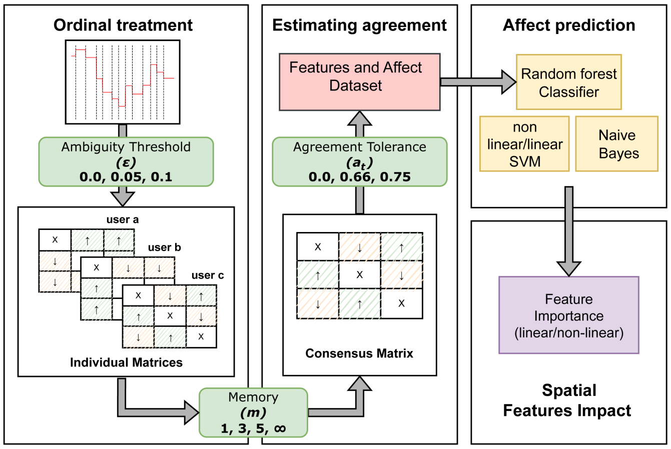


Fig. 3. Framework for processing and analyzing the data, deriving relative labels, estimating inter-annotator agreements, leading up to training the classifier and Preference score ordering. The pipeline follows the structure of Sections 5 and 6.

\	0	1	2	3	4	5	6	7	8	9	10	11
0	X	=	↓	↓	↑	=						
1	=	X	↓	↓	↑	=	↑					
2	↑	↑	X	=	↑	↑	↑	↑				
3	↑	↑	=	X	↑	↑	↑	↑	=			
4	↓	↓	↓	↓	X	↓	↑	↑	↓	↓		
5	=	=	↓	↓	↑	X	↑	↑	↓	↓	↓	
6		↓	↓	↓	↓	↓	X	↓	↓	↓	↓	↓
7			↓	↓	↓	↓	↑	X	↓	↓	↓	↓
8				=	↑	↑	↑	↑	X	=	↓	↑
9					↑	↑	↑	↑	=	X	↓	↑
10						↑	↑	↑	↑	↑	X	↑
11							↑	↑	↓	↓	↓	X

\	0	1	2	3	4	5	6	7	8	9	10	11
0	X	=	↓	↑								
1	=	X	↓	↓	↑							
2	↑	↑	X	=	↑	↑						
3	↑	↑	=	X	↑	↑	↑					
4		↓	↓	↓	X	↓	↑	↑				
5			↓	↓	↑	X	↑	↑	↓			
6				↓	↓	↓	X	↓	↓	↓		
7					↓	↓	↑	X	↓	↓	↓	
8						↑	↑	↑	X	=	↓	↑
9							↑	↑	=	X	↓	↑
10								↑	↑	↑	X	↑
11									↓	↓	↓	X

Fig. 4. Impact of memory m considered in IMs for $m = 5$ (left), $m = 3$ (right), with upwards relation shown as up arrow, downwards relation with down arrow and ambiguous with equal sign. Only shaded cells (upwards or downwards) are retained as valid labels.

video) and affect dimension (arousal or pleasure). Figure 3 summarizes our data processing steps.

5.1 Extracting Affect Measures from Annotation Traces

Following the methodology of [14], we treat the room where the player is in at the time in the recording as the (sole) elicitor of the affect annotation. Therefore, we split the time-continuous annotation signal collected by RankTrace based on which room is currently visited in the recording. Each annotation is thus split into 12 time windows (one per room), with each time window starting from the moment the player entered the room and ending the moment the next room in the queue is entered. We refer to these time windows as *room windows* [14]. Note that we remove any data where the player is not in one of the tested rooms (e.g. while being in the starting empty hallway).

While the raw annotation data from RankTrace is unbounded, before processing it further we normalize it to the value range of $[0, 1]$ through min-max normalization. This is done individually per annotation trace, as we assume that the idiosyncratic annotations of each stimulus can still contain information regarding higher and lower values if we treat these in an ordinal fashion [26], [42]. We then split the normalized annotation trace per room window. This results in 12 different sub-traces, which can be processed in different ways to derive summary metrics (as scalar values) that can be used in an ordinal fashion. We use the following affect metrics in this paper:

- **Affect Amplitude:** calculated as the difference between the maximum and minimum of the affect data points within the room window. A high value means that there were large shifts in emotion annotation while the user was in the room.
- **Affect Gradient:** calculated as the sum of absolute differences between consecutive data points (sampled at 4Hz) within the room window. We hypothesize that when annotators fluctuate their affect states during the window, different stimuli may be triggering them. While amplitude measures the difference between high and low points, a trace may have a high gradient without a corresponding high value in amplitude if the user was changing annotation direction often.

Following [19], [31], we consider amplitude and gradient as *relative* measures that assess changes of emotion, and are ideally suited for the unbounded affect traces produced by RankTrace [19]. We further process these metrics in an ordinal fashion by observing differences between metrics in different rooms within the same video.

5.2 Ordinal Treatment of Affect Measures

To derive relative labels from the signal, we follow the Qualitative Agreement (QA) approach proposed by Cowie and McKeown [15]. According to the QA approach, each resulting user-annotated trace is split into bins of standard size; after that, pairwise comparisons between the bins for each trace are made and subsequently stored into an individual matrix. The bin size in our case is the room's duration during spatial navigation. Our bin sizes differ here from QA's uniform bin sizes: since each room is not being navigated in the same manner and time-span, each bin contains a different number of frames (and annotation data points). Below we describe how the QA method is applied for each of the metrics of Section 5.1.

Since each video contains 12 rooms, the resulting individual matrix (IM) is a 12 by 12 matrix (i.e. per video and per user) containing the results of the pairwise comparison of all bins. A video may have multiple IMs according to the number of participants that annotated that particular video. To compare how the two affect metrics (detailed in Section 5.1) change between two rooms in the IM, we mark the direction as *upwards*, *downwards* or *ambiguous*. If the affect metric of the room in the IM row is lower than that of the room in the IM column, we mark it as an upwards relation for that bin, and if the opposite is true as downwards. If the absolute difference between the affect metric for the two rooms is below an ambiguity threshold (ϵ), we consider this difference as ambiguous for that bin. In our study we explore three different thresholds for ambiguity (0.0, 0.05 and 0.1) based on past work [43].

Lastly, to create a balanced dataset for our classifier we take into account not only the upper part of the individual matrix as proposed in the QA approach [15], but also the lower part of the matrix marking the opposite trend relationship of the pair. For example, when the pairwise comparison trend between rooms A and B is marked as upwards then the comparison trend between B and A is marked as downwards (see Fig. 3; Ordinal treatment step). This ensures a 50% random guess baseline for our predictive models (see Section 6).

Through this process, the IM contains the relationships of each room versus all remaining rooms for a single user. While the IM contains all relevant information for one annotation session and affect metric, we do not necessary wish to compare affect across all rooms in the video. We hypothesize instead that a comparison in terms of affect responses is more relevant between consecutive rooms, i.e. that annotators may have short-term memory patterns in their annotation that may not hold throughout the entire trace. To cater for this, we introduce a *memory* (m) parameter as the number of neighboring rooms considered in terms of ordinal relationships. This memory parameter can be applied directly on the IM, as shown in Fig. 4. In this paper we test four memory thresholds: $m = 1$ (i.e. only considering consequent rooms), $m = 3$, $m = 5$ and $m = \infty$ (i.e. comparing affect across all rooms).

5.3 Leave-One-Subject-Out Cross-Validation Protocol

To evaluate the performance of our models, we employ a leave-one-subject-out cross-validation (LOSOCV) approach during training and testing. If we assume n participants contributing their annotations, we first isolate three participants for hyperparameter tuning (external validation set), while the remaining $n-3$ participants undergo LOSOCV. RF training uses the scikit-learn Python package [44], and the hyperparameters tuned are: the number of trees, maximum tree depth, minimum number of samples per leaf node and minimum number of samples required to split a tree node[2]. All hyperparameters are tuned via exhaustive search targeted at maximizing accuracy on the validation set. The validation set, which consists of annotations from 3 participants on multiple video stimuli, uses the IMs of each participant to test models trained on the remainder of the dataset ($n-3$ participants). The best performing hyperparameters based on test accuracy on this external set of 3 participants' IMs is used in follow-up LOSOCV sets. The hyperparameter configurations for the top performing models, covering both affect labels and metrics, are presented in Table 4. Note that these 3 participants are never included for either training or testing in follow-up LOSOCV steps discussed below.

For training and testing via LOSOCV, each fold uses $(n-3)-1$ participants for training and one participant for testing. We derive different datasets through this process by exploring parameter tuples of our QA framework (memory, agreement tolerance, ambiguity threshold). Test accuracy is measured on the left-out participant's IM, while the training set undergoes a more involved process based on inter-rater agreement. For each video in the corpus, the IMs of annotators in the training set are joined into a singular consensus matrix (CM) as follows: we assign a label in the CM (*upwards* or *downwards* as per Section 5.2) if the ratio of annotators agreeing with this label is above an *agreement tolerance threshold* (a_t). In this paper we explore different a_t values of 50% (simple majority), 66% and 75% as well as a baseline with $a_t = 0\%$ where all users' data is treated as valid. The training set uses the CMs calculated per video (only considering the $(n-3)-1$ participants), while the trained model is then tested on videos that the left-out participant has annotated individually (IM).

6 Results

Through our crowdsourcing experiment, we have collected a corpus of annotations as listed in Table 1. By processing our annotations as discussed in Section 5, and calculating model inputs as discussed in Section 3.2, we can compare how the space impacts the affect dimensions of arousal and pleasure annotated by our participants. The following sections present the analysis of our results.

6.1 Parameter Tuning for Affect Prediction

In order to assess the potential of Random Forest classifiers (RF) for the task of predicting changes in affect amplitude or affect gradient when parameters of the environment change, we performed an extensive experiment with different parameter settings. Since RFs are stochastic by nature, experiments with leave-one-subject-out were repeated 10 times. The results are reported in Tables 2 and 3, with the classification accuracy and the number of data points that the model is trained on. Unsurprisingly, we observe that higher thresholds for what constitutes a valid change according to an extended majority results in much smaller datasets. Especially for $a_t = 75\%$ the dataset size drops significantly, at least when comparing only consecutive rooms ($m = 1$); however, those parameter pairings seem to yield some of the most accurate models. An expected behavior is the sub-par performance for $\epsilon = 0$, since the dataset includes every minor change as valid—which in turn confuses the predictive models. Surprisingly, ignoring inter-rater agreement ($a_t = 0\%$) does not seem to lead to a drop in accuracy despite retaining multiple user perspectives. Only comparing consequent rooms ($m = 1$) seems to lead to better models overall; for affect amplitude larger memory windows seem to perform better, but accuracies are generally low. While none of the models reach particularly impressive accuracies on this challenging leave-one-subject-out task, binomial testing [45] was used to establish statistical significance (at $p < 0.05$), with the hypothesis that the observed prediction is significantly different from chance. Based on the binomial test, we calculated the test statistic using the observed accuracies and compared these to the total number of samples on each dataset, with the expected probability set to 50%, reflecting the balancing of the dataset (see Section 5.2). Tests show that accuracies in Tables 2 and 3 are significantly above the 50% baseline except for pleasure amplitude at $a_t = 75\%, \epsilon = 0.05, m = 5$ (at 50.3% accuracy) and pleasure amplitude at $a_t = 50\%, \epsilon = 0.0, m = 3$ (at 49.4% accuracy). It is evident that changes in the affect gradient are easier to predict from the changes in design features, with test accuracies as high as 69% for arousal gradient and 68.1% for pleasure gradient. In comparison, predicting changes in affect amplitude is more challenging,

[2] Apart from the listed hyperparameters, the default values for Random Forests on scikit-learn are used.

TABLE 2
Test accuracies (%) for arousal modeling. Bold highlights single highest scores per affect treatment. Accuracies (and dataset sizes in parentheses) are averaged from 36 leave-one-participant-out experiments.

		Memory (m)			
a_t	ϵ	1	3	5	∞
Arousal Gradient					
0%	0	63% (2.1K)	62% (7.9K)	61% (16.5K)	62% (29.3K)
	0.05	67% (796)	65% (2.8K)	64% (6.1K)	64% (11.1K)
	0.1	69% (403)	66% (1.3K)	64% (2.8K)	65% (5.2K)
50%	0	64% (390)	63% (1.1K)	61% (1.8K)	62% (2.6K)
	0.05	66% (336)	64% (912)	64% (1.4K)	65% (2.1K)
	0.1	**69%** (264)	65% (641)	64% (1K)	66% (1.5K)
66%	0	65% (225)	62% (670)	62% (1.1K)	63% (1.6K)
	0.05	67% (124)	66% (517)	64% (1K)	64% (1.6K)
	0.1	**69%** (56)	66% (309)	65% (719)	64% (1.1K)
75%	0	65% (130)	63% (371)	62% (694)	63% (1.1K)
	0.05	66% (42)	65% (180)	65% (559)	65% (1.1K)
	0.1	67% (8)	67% (57)	65% (323)	65% (741)
Arousal Amplitude					
0%	0	55% (2K)	53% (7.6K)	55% (16K)	54% (28.4K)
	0.05	60% (1.5K)	60% (5.7K)	60% (12K)	61% (21.5K)
	0.1	60% (1K)	61% (4K)	62% (8.6K)	62% (15.6K)
50%	0	57% (383)	54% (1.1K)	55% (1.7K)	58% (2.5K)
	0.05	60% (360)	60% (1.1K)	60% (1.6K)	61% (2.4K)
	0.1	60% (332)	60% (975)	61% (1.5K)	62% (2.3K)
66%	0	61% (190)	56% (610)	54% (1K)	54% (1.5K)
	0.05	62% (159)	60% (626)	60% (1.1K)	61% (1.6K)
	0.1	62% (132)	60% (558)	61% (1.1K)	62% (1.6K)
75%	0	59% (91)	57% (281)	54% (624)	53% (1K)
	0.05	62% (69)	59% (261)	60% (651)	60% (1.1K)
	0.1	**62%** (42)	59% (207)	61% (591)	60% (1.1K)

TABLE 3
Test accuracies (%) for pleasure modeling. Bold highlights single highest scores per affect treatment. Accuracies (and dataset sizes in parentheses) are averaged from 34 leave-one-participant-out experiments.

		Memory (m)			
a_t	ϵ	1	3	5	∞
Pleasure Gradient					
0%	0	57% (2K)	58% (7.4K)	58% (15.6K)	59% (27.6K)
	0.05	66% (753)	64% (2.5K)	64% (5.5K)	60% (9.9K)
	0.1	66% (388)	65% (1.3K)	63% (2.7K)	63% (4.9K)
50%	0	64% (409)	59% (1.1K)	57% (1.8K)	59% (2.5K)
	0.05	66% (362)	63% (914)	64% (1.4K)	63% (2K)
	0.1	66% (263)	65% (623)	64% (933)	64% (1.4K)
66%	0	64% (216)	62% (692)	55% (1.1K)	57% (1.6K)
	0.05	66% (144)	64% (549)	63% (1K)	63% (1.5K)
	0.1	**68%** (57)	65% (325)	64% (677)	64% (1K)
75%	0	64% (120)	62% (362)	53% (683)	54% (1.1K)
	0.05	64% (34)	65% (178)	64% (572)	63% (1K)
	0.1	60% (9)	65% (63)	64% (334)	64% (692)
Pleasure Amplitude					
0%	0	56% (1.9K)	54% (7.3K)	56% (15.3K)	57% (27.2K)
	0.05	59% (1.3K)	59% (4.8K)	52% (10.2K)	52% (18.4K)
	0.1	58% (840)	60% (3.3K)	60% (7K)	59% (12.8K)
50%	0	60% (363)	49% (1.1K)	52% (1.7K)	54% (2.4K)
	0.05	60% (347)	59% (1K)	52% (1.6K)	55% (2.3K)
	0.1	60% (344)	60% (960)	61% (1.5K)	**62%** (2.1K)
66%	0	58% (163)	56% (575)	52% (996)	52% (1.5K)
	0.05	56% (148)	59% (592)	57% (1K)	53% (1.5K)
	0.1	59% (109)	60% (541)	59% (1K)	60% (1.5K)
75%	0	58% (85)	57% (286)	51% (550)	52% (912)
	0.05	60% (66)	58% (263)	50% (620)	54% (1.1K)
	0.1	59% (26)	58% (171)	57% (551)	56% (1K)

with the highest test accuracies for arousal amplitude at 62.3% and for pleasure amplitude at 61.8%. For arousal gradient, the highest accuracy (69%) is with two datasets of $\epsilon = 0.1$, $m = 1$ and either $a_t = 50\%$ or $a_t = 66\%$; since the number of data points is much higher for $a_t = 50\%$ (264) than for $a_t = 66\%$ (56), we prefer the former and treat it as our best model for arousal gradient. For pleasure gradient, the best model is with $\epsilon = 0.1$, $m = 1$ and $a_t = 66\%$ (68.1% accuracy); the second best model is at $a_t = 50\%$ and the same m and ϵ (65.9% accuracy), but on a far larger dataset (263 data points versus 57 for the best model). Since the QA method is expected to lead to more robust ground truth data, for the sake of a larger dataset we opt to use data with $\epsilon = 0.1$, $m = 1$ and $a_t = 50\%$ for both pleasure and arousal gradient as "best" for additional experiments in Section 6.2.

6.2 Different Predictive Models

Through the extensive validation process of Section 6.1, we confirm that affect gradient as a metric is more easily predicted by the design features of architectural spaces. Therefore, we use the datasets with the highest accuracy and sufficient dataset size from Tables 2 and 3, i.e. with $a_t = 50\%$, $\epsilon = 0.1$ and $m = 1$. For these datasets, the average test accuracy of the RF are 69% for arousal gradient and 65.9% for pleasure gradient.

We compare the RF method with other popular classification algorithms from the literature: linear and non-linear Support Vector Machines (SVM) [46] and Naive Bayes (NB) [47]. For the non-linear SVM we utilize the Radial Basis Function (RBF) kernel while for all classifiers we follow the same leave-one-subject-out approach described in Section 5.3. Hyperparameters that are tuned here are the C hyperparameter for both SVMs, as well as gamma for the non-linear SVM (for the linear SVM gamma is set to 1). For the Naive Bayes classifier we tune the smoothing parameter alpha. All machine learning algorithms use the scikit-learn Python package [44], with default hyperparameters except those noted.

Table 4 compares the performance of the different models on predicting arousal and pleasure gradient and amplitude changes between rooms. Beyond accuracy, we include precision (true positives versus all positives), recall (true positives versus all samples that should have been identified as positive), and F1 score (harmonic mean of the precision and recall) as established measures for classification tasks. We establish statistical significance via Wilcoxon signed-rank tests on the same test folds (for paired data) at $p < 0.05$; when reporting significance in multiple comparisons, the Bonferroni correction is applied [48]. Table 4 indicates that the RBF SVM underperforms, reaching significantly lower F1 scores and recall values compared to all other models on all affect metrics, and significantly lower test accuracies than all other models in all affect metrics except pleasure amplitude. Other models reach similar test accuracies, between 68% and 70% for gradient metrics and lower (between 60% and 63%) for amplitude metrics. There are no significant differences between these models for test accuracies, except the NB model outperforming Linear SVM for pleasure amplitude. On the other hand, RF has a significantly higher F1 score and recall than other models for arousal amplitude, but for other affect metrics there are no significant differences. While NB models tend to have higher test accuracies

TABLE 4
Classification performance of the best affect datasets for different predictive models. Results are averaged from leave-one-subject-out runs and 95% confidence intervals are included.

Model	Accuracy	F1 Score	Precision	Recall	Hyperparams*
Arousal gradient					
RF	69%±2.6%	.68±.028	.67±.026	.70±.039	{100, 3, 6, 6, T}
Lin. SVM	68%±3.0%	.68±.030	.65±.025	.72±.043	{1}
RBF SVM	59%±3.3%	.39±.056	.66±.061	.29±.055	{100, 1}
NB	68%±3.3%	.66±.038	.67±.031	.66±.050	{89}
Pleasure gradient					
RF	66%±5.5%	.65±.057	.67±.059	.63±.060	{50, 5, 10, 10, T}
Lin. SVM	69%±4.5%	.66±.058	.65±.057	.67±.062	{9}
RBF SVM	60%±4.1%	.40±.059	.67±.081	.29±.051	{1, 1}
NB	70%±4.6%	.64±.060	.71±.063	.59±.060	{1}
Arousal amplitude					
RF	61%±2.5%	.55±.032	.61±.042	.52±.039	{50, 3, 2, 2, T}
Lin. SVM	62%±2.5%	.51±.034	.67±.042	.43±.043	{3}
RBF SVM	54%±2.0%	.29±.036	.56±.062	.21±.030	{1, 10}
NB	63%±2.3%	.50±.032	.69±.042	.40±.032	{56}
Pleasure amplitude					
RF	60%±5.1%	.50±.056	.64±.069	.41±.051	{50, 7, 2, 2, T}
Lin. SVM	58%±3.9%	.57±.045	.57±.047	.58±.051	{5}
RBF SVM	55%±3.7%	.29±.050	.53±.094	.21±.036	{1, 0.1}
NB	63%±4.3%	.49±.058	.66±.075	.40±.050	{56}

* The tuned hyperparameters for these models are shown in order as follows. For RF the array is {Number of estimators, max. tree depth, min. number of samples required to split an internal node, min. number of samples required to be at a leaf node, whether bootstrap samples are used when building trees (True or False)}. For Lin. SVM the array is {C} and for RBF SVM the array is {C, gamma}. For NB the array is {smoothing parameter alpha}.

than RF and linear SVM models, those differences are not significant; instead, recall values for RF models are significantly higher than NB models for all affect metrics. The linear SVM performs generally on par with RF models, with no significant differences for test accuracy. Based on this, RF models perform comparably to other models in general; we argue in favor of using RFs in Section 7 and leverage them again for finding which spatial features impact emotion in Section 6.3.

6.3 Impact of Different Spatial Features

Looking at the feature importance of the different inputs, Table 5 lists the five most important features for the RF (based on mean decrease in impurity within each tree) for all affect metrics. We use RFs for this analysis as they perform well (see Section 6.2) and are more interpretable for the task of deriving feature importances. We observe that occlusion plays a major role for both affect labels and metrics. Considering that the gradient captures how often a user changes their arousal or pleasure annotation within the same room and amplitude captures the range of change within that room, this finding is not surprising: rooms with occlusions reveal parts of the room at different times, and there are more surprising moments that may result in annotation changes. Illumination color has some impact as a predictor for both arousal and pleasure, which aligns with existing theories for the impact of illumination color in digital games [49]. Height as a predictor contributes mildly to both arousal models while curvature seems to contribute more to pleasure models. Both findings align with theories regarding the impact of scale and arousal or awe [50] and curvature or non-rectilinearity on pleasure [30], [51]. Finally, we observe that the duration of the navigation in both rooms is an important feature for all models, indicating that factors such as recorded viewing behavior and navigation pace play an important role in the manner the environment is perceived and annotated by the participants.

TABLE 5
Top five features of RF predictors for affect gradient and amplitude changes between rooms A and B (and their difference Δ for this metric), calculated based on mean decrease in impurity.

Stat	Arousal		Pleasure	
	feature	weight	feature	weight
Gradient	Occlusions $_A$	.408	Occlusions $_\Delta$	.518
	Duration	.219	Occlusions $_A$	.217
	Occlusions $_\Delta$	.186	Duration	.118
	Height $_\Delta$	.030	Illumination $_\Delta$	.055
	Illumination $_\Delta$	.030	Curvature $_\Delta$	.022
Amplitude	Occlusions $_\Delta$	.572	Occlusions $_\Delta$	.321
	Duration	.162	Duration	.195
	Illumination $_A$	.083	Occlusions $_A$	.193
	Height $_B$	.049	Illumination $_\Delta$	.121
	Occlusions $_B$	.044	Curvature $_\Delta$	.063

7 Discussion

Our framework introduced a method for capturing and treating unbounded and dimensional data of arousal and pleasure in order to find a mapping between affect and environmental stimuli of the spaces we navigate in. Taking advantage of a large pool of annotators (i.e. 39 and 37 for arousal and pleasure, respectively) we process the collected data in an ordinal manner seeking agreements across annotators. To find the mapping between affect and space, we leverage non-linear affect models that consider four design parameters as their input, in addition to the duration of the walkthroughs. Results indicate that affect gradient is an easier measure to predict, yielding reasonably accurate models with test accuracies around 69% for the best models when predicting arousal gradient and pleasure gradient.

The methodology presented here took advantage of crowdsourcing platforms and inter-rater agreement to produce a robust dataset with over 200 annotated videos for both arousal and pleasure. This significantly extends our earlier work [14] which was limited in scope to a few expert annotators experienced with the RankTrace protocol. It should be noted, however, that while we explored different ways of treating affect and ways to improve the integrity of the original data annotations, many of the models were rather weak in terms of their test accuracy. Alternative feature representations may be necessary to address this limitation. Preliminary tests for predicting the change in mean arousal or pleasure (calculated on a per-room basis) from one room to the next yielded accuracies close to the baseline; thus we reverted to more robust affect metrics that measure relative change within each room instead. Additional experiments were conducted with linear and nonlinear SVM and Naive Bayes (NB) classifiers, demonstrating comparable behavior to our Random Forest model (see Table 4). While NB shows improvements in test accuracy, it is worth noting that the predictor relies on strong assumptions regarding feature independence. For the current dataset, this seems to works well, but can pose challenges in future additions to the input set where features may be correlated (e.g. room volume and height). Furthermore, both Naive Bayes and SVM classifiers are sensitive to mixed data types, such as continuous and interval data; RFs are versatile in handling both data types. These observations motivate us to use RFs for this problem (including the extensive tuning reported in Tables 2 and 3), as we aim to further enhance the *Affrooms12* dataset with more representations and data types in future work. The analysis on the feature importance of the best performing predictors exhibited similar results for both metrics and affect labels. Occlusion within the room itself affects how often and how much annotators register an affect change as new stimuli come into view. Following that, height and illumination had some impact for arousal prediction, while curvature and illumination had some impact for pleasure prediction. In future work it is necessary to expand the analysis of this study via (a) modeling the current affect metrics with more advanced machine learning algorithms (e.g. using the raw footage in a computer vision approach such as [43], [52]); or (b) identifying new ways of processing the annotation so that we better leverage its unbounded and time-continuous nature.

It is worth noting that we build on several assumptions when constructing the ground truth of affect which we attempt to map to characteristics of the built environment. While we explore the impact of different parameters such as ambiguity threshold (what annotation changes would qualify as different), memory (how far back an annotator remembers and subconsciously compare their current emotion levels) and agreement threshold (how many annotators need to have similar annotation behaviors to consider those "universal"), we chose the best parameters based on their performance on a classification task using the characteristics of the built environment as input. However, more assumptions could be tested: indicatively, the use of a majority rule for deriving "universal" annotation behavior (via the agreement threshold) may overlook clusters of annotators that agree with each other but do not achieve a general consensus. The literature has suggested that inter-rater variability may be due to multiple (equally valid) perspectives [53]; however, with the few annotators per video in this study (average 4 annotators per video), further clustering would make consensus-building difficult. Future studies could collect a larger set of annotations on fewer, likely longer, navigation videos and test whether multiple perspectives exist in the affect fluctuations.

Since the *AffRooms12* dataset is now publicly available, we expect that additional methods for processing or modeling the signals, videos and metadata will be forthcoming. Regarding future work beyond the *AffRooms12* dataset, we have identified that the current spatial stimuli may not be sufficient to elicit visceral affect reactions; this may lead to more noise and biases in the ground truth data. While inter-annotator agreement aimed to cater for individual reporting biases, developing a more stimulating video recording with more diverse environments (e.g. with more objects, interplay between light and darkness, and textures) is more likely to yield more reliable affect annotations at the cost of tractability. A more involved environment with e.g. more occluding elements or branching paths would allow us to study more complex interior arrangements, perhaps processed via isovists and space syntax features [54]. Preliminary experiments for this paper with such metrics did not seem to result in better predictive models, perhaps due to the simplicity of the room structures and navigation experience. Future studies could incorporate room context and intended use, as in the dataset curated by Gregorian et. al. [55] which comprises of first-person-view video recordings of traversals within constructed environments. Combining our annotation and processing approach with such stimuli can offer valuable insights into the interplay between spatial contexts and affect. Future work could also explore more task-oriented navigation, e.g. using game elements such as monsters, locked doors and keys, to make the navigation more tense [56] and camera facing more important (e.g. when searching for a key in the rubble). This was indicatively explored using YouTube streamers' "Let's play" videos of a horror game level, mapping architectural and game design elements to derived emotions from streamers' facial expressions and voice [57]. Finally, while in this paper many users annotated the same pre-recorded videos of spatial navigation, annotators could also be in control of the navigation task itself. In a recent study conducted by the authors [58], interactive virtual environments were leveraged to investigate how a user can annotate while they explore the space (partially, as their movement is restricted). The findings revealed notable inter-annotator agreements on the impact of architectural features to pleasure, highlighting the role of action in perception [59].

8 Conclusion

Motivated by research on the impact of architectural form and light on affect, this paper significantly extends our pilot study [14] by introducing a new and more extensive dataset of 3D spatial navigation videos, *Affrooms12*, and a total of 76 annotation traces of arousal or pleasure. In this paper we performed a first analysis of the annotation data, processing them in an ordinal fashion and minimizing noise through inter-rater agreement. An extensive parameter tuning explored both the annotator's memory and the different extracted metrics of affect from the time-continuous unbounded annotation traces. Results showed that some affect metrics are more robust than others, yielding test accuracies as high as 69% for changes in arousal and pleasure in the challenging leave-one-subject-out protocol. The dataset is made available as part of this publication in order to invite further study in affect modeling during spatial navigation tasks. Extensions could include taking advantage of additional modalities such as the recorded footage or the in-game camera perspective, or new ways of pre-processing the signals to improve annotations' validity.

Acknowledgments

The research leading to these results has received funding from the Malta Council for Science and Technology through the SINO-MALTA Fund 2022, Project OPtiMaL.

References

[1] U. Nanda, D. Pati, H. Ghamari, and R. Bajema, “Lessons from neuroscience: Form follows function, emotions follow form,” *Intelligent Buildings Intl.*, vol. 5, no. sup1, pp. 61–78, 2013.

[2] O. Vartanian, G. Navarrete, A. Chatterjee, L. B. Fich, H. Leder, C. Modroño, M. Nadal, N. Rostrup, and M. Skov, “Impact of contour on aesthetic judgments and approach-avoidance decisions in architecture,” *Proc. of the National Academy of Sciences*, vol. 110, no. Supplement 2, pp. 10 446–10 453, 2013.

[3] M. Banaei, A. Ahmadi, K. Gramann, and J. Hatami, “Emotional evaluation of architectural interior forms based on personality differences using virtual reality,” *Frontiers of Architectural Research*, vol. 9, no. 1, pp. 138–147, 2020.

[4] M. Banaei, J. Hatami, A. Yazdanfar, and K. Gramann, “Walking through architectural spaces: The impact of interior forms on human brain dynamics,” *Frontiers in human neuroscience*, vol. 11, p. 477, 2017.

[5] V. Kosorić, H. Huang, A. Tablada, S.-K. Lau, and H. T. Tan, “Survey on the social acceptance of the productive façade concept integrating photovoltaic and farming systems in high-rise public housing blocks in Singapore,” *Renewable and Sustainable Energy Reviews*, vol. 111, pp. 197–214, 2019.

[6] H.-T. Hamdy Mahmoud, “Interior architectural elements that affect human psychology and behavior,” *The Academic Research Community publication*, vol. 1, 2017.

[7] Y. A. de Kort, “Tutorial: Theoretical considerations when planning research on human factors in lighting,” *Leukos*, vol. 15, no. 2-3, pp. 85–96, 2019.

[8] D. Smith, P. Metcalfe, and M. Lommerse, “Interior architecture as an agent for wellbeing,” *Journal of the Home Economics Institute of Australia*, vol. 19, no. 3, pp. 2–9, 2012.

[9] A. Petermans and A. E. Pohlmeyer, “Design for subjective well-being in interior architecture,” in *Proc. of the Annual Architectural Research Symposium in Finland*, 2014, pp. 206–218.

[10] D. Kirsh, “Atmosphere, mood, and scientific explanation,” *Frontiers in Computer Science*, vol. 5, 2023.

[11] F. Bianconi, M. Filippucci, G. Magrini, and M. Seccaroni, “Designing with emotional awareness,” *The Intl. Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 46, pp. 55–62, 2021.

[12] G. Hildebrand, *Origins of architectural pleasure*. University of California Press, 1999.

[13] J. Appleton, *The experience of landscape*. Wiley Chichester, 1996.

[14] E. Xylakis, A. Liapis, and G. N. Yannakakis, “Architectural form and affect: A spatiotemporal study of arousal,” in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2021.

[15] R. Cowie and G. McKeown, “Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme,” SEMAINE Report D6b, 2010.

[16] V. Sethu, E. M. Provost, J. Epps, C. Busso, N. Cummins, and S. Narayanan, “The ambiguous world of emotion representation,” *arXiv preprint arXiv:1909.00360*, 2019.

[17] A. Pinilla, J. Garcia, W. Raffe, J.-N. Voigt-Antons, R. Spang, and S. Möller, “Affective visualization in virtual reality: An integrative review,” *arXiv preprint arXiv:2012.08849*, 2020.

[18] G. N. Yannakakis and H. P. Martinez, “Grounding truth via ordinal annotation,” in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2015.

[19] P. Lopes, G. N. Yannakakis, and A. Liapis, “RankTrace: Relative and unbounded affect annotation,” in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2017.

[20] E. Harmon-Jones, C. Harmon-Jones, and E. Summerell, “On the importance of both dimensional and discrete models of emotion,” *Behavioral sciences*, vol. 7, no. 4, p. 66, 2017.

[21] J. T. Cacioppo, W. L. Gardner, and G. G. Berntson, “Beyond bipolar conceptualizations and measures: The case of attitudes and evaluative space,” *Personality and Social Psychology Review*, vol. 1, no. 1, pp. 3–25, 1997.

[22] M. M. Bradley and P. J. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[23] J. A. Russell, “A circumplex model of affect.” *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[24] P. M. Desmet, M. H. Vastenburg, and N. Romero, “Mood measurement with pick-a-mood: Review of current methods and design of a pictorial self-report scale,” *Journal of Design Research*, vol. 14, no. 3, pp. 241–279, 2016.

[25] P. Ekman, “Emotions revealed,” *BMJ*, vol. 328, 2004. [Online]. Available: https://www.bmj.com/content/328/Suppl_S5/0405184

[26] G. N. Yannakakis, R. Cowie, and C. Busso, “The ordinal nature of emotions: An emerging approach,” *IEEE Trans. on Affective Computing*, vol. 12, no. 1, pp. 16–35, 2018.

[27] A. H. Maghool, M. Homolja, and M. A. Schnabel, “Cybernetics approach to virtual emotional spaces - an electrodermal activity actuated adaptive space,” in *Proc. of Education and research in Computer Aided Architectural Design in Europe Conf.*, 2020.

[28] S. Kalantari, “A new method of human response testing to enhance the design process,” *Proc. of the Design Society: Intl. Conf. on Engineering Design*, vol. 1, no. 1, p. 1883–1892, 2019.

[29] P. Naghibi Rad, A. A. Shahroudi, H. Shabani, S. Ajami, and R. Lashgari, “Encoding pleasant and unpleasant expression of the architectural window shapes: An ERP study,” *Frontiers in Behavioral Neuroscience*, vol. 13, 2019. [Online]. Available: https://www.frontiersin.org/article/10.3389/fnbeh.2019.00186

[30] N. Ruta, S. Mastandrea, O. Penacchio, S. Lamaddalena, and G. Bove, “A comparison between preference judgments of curvature and sharpness in architectural façades,” *Architectural Science Review*, vol. 62, no. 2, pp. 171–181, 2019.

[31] K. Chamilothori, “Effects of façade and daylight pattern geometry on subjective and physiological responses: Findings from experiments in immersive virtual reality,” in *Proc. of the Kongsberg Vision Meeting: Immersive technologies for eye care and lighting design*, 2019.

[32] G. Gómez-Puerto, J. Rosselló, G. Corradi, C. Acedo-Carmona, E. Munar, and M. Nadal, “Preference for curved contours across cultures.” *Psychology of Aesthetics, Creativity, and the Arts*, vol. 12, no. 4, p. 432, 2018.

[33] J. Aronoff, “How we recognize angry and happy emotion in people, places, and things,” *Cross-cultural research*, vol. 40, no. 1, pp. 83–105, 2006.

[34] M. Bar and M. Neta, “Humans prefer curved visual objects,” *Psychological science*, vol. 17, no. 8, pp. 645–648, 2006.

[35] E. Garip and B. Seymen, “Research for evaluating perception of concrete material by using visual research methods in learning environments,” *AZ ITU Journal of the Faculty of Architecture*, vol. 18, no. 1, pp. 17–28, 2021.

[36] B. Coşgun, K. Yldrm, and M. L. Hidayetoglu, “Effect of wall covering materials on the perception of cafe environments,” *Facilities*, vol. 40, no. 3-4, pp. 214–232, 2021.

[37] H. C. Gomez-Tone, J. Martin-Gutierrez, J. Bustamante-Escapa, P. Bustamante-Escapa, and B. K. Valencia-Anci, “Perceived sensations in architectural spaces through immersive virtual reality,” *VITRUVIO-Intl. Journal of Architectural Technology and Sustainability*, vol. 6, no. 2, pp. 70–81, 2021.

[38] R. Lipson-Smith, J. Bernhardt, E. Zamuner, L. Churilov, N. Busietta, and D. Moratti, “Exploring colour in context using virtual reality: Does a room change how you feel?” *Virtual Reality*, vol. 25, no. 3, pp. 631–645, 2021.

[39] O. Vartanian, G. Navarrete, A. Chatterjee, L. B. Fich, H. Leder, C. Modroño, N. Rostrup, M. Skov, G. Corradi, and M. Nadal, “Preference for curvilinear contour in interior architectural spaces: Evidence from experts and nonexperts.” *Psychology of Aesthetics, Creativity, and the Arts*, vol. 13, no. 1, p. 110, 2019.

[40] D. Melhart, A. Liapis, and G. N. Yannakakis, “PAGAN: Video affect annotation made easy,” in *Proc. of the Intl. Conf. on Affective Computing and Intelligent Interaction*, 2019.

[41] ——, “The affect game annotation (AGAIN) dataset,” *arXiv preprint arXiv:2104.02643*, 2021.

[42] G. N. Yannakakis, R. Cowie, and C. Busso, “The ordinal nature of emotions,” in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2017, pp. 248–255.

[43] K. Makantasis, A. Liapis, and G. N. Yannakakis, “From pixels to affect: A study on games and player experience,” in *Proc. of the IEEE Intl. Conf. on Affective Computing and Intelligent Interaction*, 2019.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[45] D. Cramer, *Fundamental statistics for social research: Step-by-step calculations and computer techniques using SPSS for Windows*. Routledge, 2003.

[46] V. Vapnik and C. Cortes, "Support vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[47] I. Rish, "An empirical study of the naive Bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[48] O. Dunn, "Multiple comparisons among means," *Journal of the American Statistical Association*, vol. 56, no. 293, 2012.

[49] S. Niedenthal, "Patterns of obscurity : Gothic setting and light in Resident Evil 4 and Silent Hill 2," in *Horror Video Games: Essays on the Fusion of Fear and Play*, B. Perron, Ed. McFarland, 2009.

[50] T. Heath, S. G. Smith, and B. Lim, "Tall buildings and the urban skyline: The effect of visual complexity on preferences," *Environment and behavior*, vol. 32, no. 4, pp. 541–556, 2000.

[51] N. Ruta, J. Vañó, R. Pepperell, G. B. Corradi, E. G. Chuquichambi, C. Rey, and E. Munar, "Preference for paintings is also affected by curvature." *Psychology of Aesthetics, Creativity, and the Arts*, vol. 17, no. 3, p. 307, 2023.

[52] K. Makantasis, A. Liapis, and G. N. Yannakakis, "The pixels and sounds of emotion: General-purpose representations of arousal in games," *IEEE Trans. on Affective Computing*, vol. 14, no. 1, 2023.

[53] A. Gingerich, C. P. M. van der Vleuten, K. W. Eva, and G. Regehr, "More consensus than idiosyncrasy: Categorizing social judgments to examine variability in Mini-CEX ratings," *Journal of Machine Learning Research*, vol. 89, pp. 1510–1519, 2011.

[54] P. Koutsolampros, K. Sailer, T. Varoudis, and R. Haslem, "Dissecting visibility graph analysis: The metrics and their role in understanding workplace human behaviour," in *Proc. of the 12th Intl. Space Syntax Symposium*, 2019.

[55] L. Gregorians, P. F. Velasco, F. Zisch, and H. J. Spiers, "Architectural experience: Clarifying its central components and their relation to core affect with a set of first-person-view videos," *Journal of Environmental Psychology*, vol. 82, 2022.

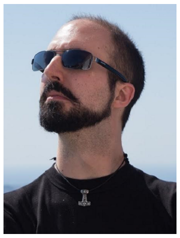
[56] P. Lopes, A. Liapis, and G. N. Yannakakis, "Framing tension for game generation," in *Proc. of the Intl. Conf. on Computational Creativity*, 2016.

[57] E. Xylakis, A. Liapis, and G. N. Yannakakis, "The Scream Stream: Multimodal affect analysis of horror game spaces," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction Workshops and Demos*, 2024.

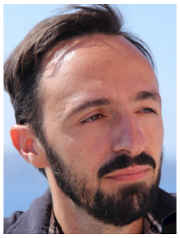
[58] E. Xylakis, A. Najm, D. Michael-Grigoriou, A. Liapis, and G. N. Yannakakis, "Eliciting and annotating emotion in virtual spaces," in *Proceedings of the 41st Education and Research in Computer Aided Architectural Design in Europe (eCAADe) Conference*, 2023.

[59] J. K. Witt and M. A. Riley, "Discovering your inner Gibson: Reconciling action-specific and ecological approaches to perception–action," *Psychonomic Bulletin & Review*, vol. 21, pp. 1353–1370, 2014.

Emmanouil Xylakis is a PhD student at the Institute of Digital Games, University of Malta. He received his Msc degree in Lighting design from Aalborg University, Copenhagen in 2018, focusing on the use of Virtual Reality as a design tool during the process of Architectural design. His Phd work focuses on the affective potential of our surrounding environments combining the fields of Affective Computing, Virtual environments, Environmental psychology and Human Computer Interaction.



Antonios Liapis is an Associate Professor at the Institute of Digital Games, University of Malta, where he bridges the gap between game technology and game design in courses focusing on human-computer creativity, digital prototyping and game development. He received the Ph.D. degree in Information Technology from the IT University of Copenhagen in 2014. His research focuses on Artificial Intelligence in Games, Human-Computer Interaction, Computational Creativity, and User Modeling. He has published over 150 papers in the aforementioned fields, and has received several awards for his research contributions and reviewing effort. He serves as Associate Editor for the IEEE TRANSACTIONS ON GAMES, and has served as general chair in four international conferences, as guest editor in four special issues in international journals, and has co-organized 16 workshops.



Georgios N. Yannakakis is a Professor at the Institute of Digital Games, University of Malta. He received the Ph.D. degree in Informatics from the University of Edinburgh in 2006.. He does research at the crossroads of artificial intelligence, computational creativity, affective computing, advanced game technology, and human-computer interaction. He has published over 300 papers in the aforementioned fields and his work has been cited broadly. His research has been supported by numerous national and European grants (including a Marie Skłodowska-Curie Fellowship) and has appeared in *Science Magazine* and *New Scientist* among other venues. He is currently the Editor-in-Chief of the IEEE TRANSACTIONS ON GAMES and an Associate Editor of the IEEE TRANSACTIONS ON EVOLUTIONARY COMPUTATION, and used to be Associate Editor of the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the IEEE TRANSACTIONS ON COMPUTATIONAL INTELLIGENCE AND AI IN GAMES journals. He has been the General Chair of key conferences in the area of game artificial intelligence (IEEE CIG 2010) and games research (FDG 2013, 2020). Among the several rewards he has received for journal and conference publications he is the recipient of the *IEEE Transactions on Affective Computing Most Influential Paper Award* and the *IEEE Transactions on Games Outstanding Paper Award*. Georgios is an IEEE Fellow.