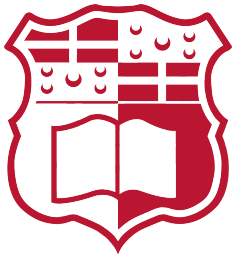

Towards a Model of Affect in Architectural Experience

Using Virtual Environments as Spatial Elicitors to Capture Real-time and Continuous Observer Feedback.

Emmanouil Xylakis

Supervised by Prof. Georgios N. Yannakakis

Co-supervised by Prof. Antonios Liapis

Institute of Digital Games

University of Malta

March, 2025

A dissertation submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

L-Università ta' Malta

FACULTY/INSTITUTE/CENTRE/SCHOOL Institute of Digital Games

DECLARATION OF AUTHENTICITY FOR DOCTORAL STUDENTS

(a) Authenticity of Thesis/Dissertation

I hereby declare that I am the legitimate author of this Thesis/Dissertation and that it is my original work.

No portion of this work has been submitted in support of an application for another degree or qualification of this or any other university or institution of higher education.

I hold the University of Malta harmless against any third party claims with regard to copyright violation, breach of confidentiality, defamation and any other third party right infringement.

(b) Research Code of Practice and Ethics Review Procedure

I declare that I have abided by the University's Research Ethics Review Procedures. Research Ethics & Data Protection form code N./A..

☒ As a Ph.D. student, as per Regulation 66 of the Doctor of Philosophy Regulations, I accept that my thesis be made publicly available on the University of Malta Institutional Repository.

□ As a Doctor of Sacred Theology student, as per Regulation 17 (3) of the Doctor of Sacred Theology Regulations, I accept that my thesis be made publicly available on the University of Malta Institutional Repository.

□ As a Doctor of Music student, as per Regulation 26 (2) of the Doctor of Music Regulations, I accept that my dissertation be made publicly available on the University of Malta Institutional Repository.

□ As a Professional Doctorate student, as per Regulation 55 of the Professional Doctorate Regulations, I accept that my dissertation be made publicly available on the University of Malta Institutional Repository.

Acknowledgements

It has been a fulfilling and long journey to complete this thesis, despite the many obstacles and challenges that I encountered along the way. I would like to express my heartfelt gratitude to my supervisors, Professor Georgios N. Yannakakis and my co-supervisor, Professor Antonios Liapis, for their patience, guidance, and continuous support throughout this journey. Their dedication during our supervisions and discussions made me feel included, interested and engaged, both within our group and in my research.

I would also like to thank my family, especially my father and mother, for enduring my emotional ups and downs throughout this period. Their patience and constant encouragement were pivotal in helping me overcome challenges and continue pushing forward. I am deeply grateful to my friends Antonis, Dimitris, Mai, Giorgos, Akis, Manos, and Stelios, and to the community of our second home on Splantzia Square. This neighborhood was a source of respite, providing a retreat to recharge and gain the motivation to return to my research with renewed determination. My friend Themis, who has made my stay in Malta a great experience.

My partner, Virginia, has been immensely patient and supportive, helping me to see beyond the immediate struggles and continually pushing me to keep going. Her encouragement has been invaluable.

I would like to thank David Melhart for his constant availability and for sharing his thoughts and expertise on common research interests and work-related matters. Additionally, I am grateful to all my colleagues at the Institute of Digital Games where our weekly GameAI meetings on Fridays provided a source of inspiration and new perspectives.

Finally, I extend my thanks to all the individuals I had the privilege of meeting at conferences, through joint collaborations, and during my time in The Hague. Their intrigue and support have played a crucial role in bringing this work to fruition.

Thank you all. This is dedicated to you.

Abstract

How do spaces make us feel? What is the perceived emotional impact of built forms? Is it possible to model this relationship? Estimating the affective impact of space has long been a challenge in understanding human-environment interaction. Most of the theories link environmental preference to instincts and evolution with preferred spaces displaying features of refuge, naturalness and ease of cognitive processing. Many studies study the effects of space focusing on specific elements and recording observer reactions. However, many of these studies rely on passive stimuli, such as photographs or static computer-generated imagery, to gather affective data. While other approaches employ invasive and specialized equipment, such as fMRI or EEG, these methods are often impractical for widespread application. Advancements in Affective Computing offer effective and non-invasive means of capturing continuous affect annotations of dynamic media like movies, games, and 360-degree content. This thesis adopts a similar perspective, treating architectural experiences as continuous and evolving media experiences, akin to those in interactive media. It conceptualizes the emotional responses elicited by built environments as unfolding over time, shaped by key spatial and temporal elements.

The aim of this dissertation is two-fold: (1) to understand the affective impact of spatial key elements under the temporal scope, and (2) to formalize and quantify this relationship. Key questions include: Can we reliably estimate the impact of spatial elements on human affect? To what extent can the affective impact of architectural experience—considering its temporal dimension—be modeled? How can we collect first-person annotation in response to spatial stimuli? To address these questions, four user studies were conducted using three types of stimuli and two approaches to affect annotation. The findings provide insights into the dynamic interplay between architectural design and human emotion, contributing to both theoretical understanding and practical applications in Architecture and Affective Computing.

Contents

List of Figures

List of Tables

Acronyms

AC Affective Computing. 4, 9, 16, 20, 37, 45, 49, 55, 86, 119

AEC Architecture, Engineering, and Construction. 2

ANN Artificial neural network. 31

CCT correlated color temperature. 26, 52

CM Consensus Matrix. 42, 43

CNNs Convolutional Neural Networks. 127

ECG Electrocardiography. 20

EDA Electrodermal Activity. 21, 25, 29

EEG Electroencephalogram. 3, 17, 20, 21, 23, 24

EMG Electromyography. 20, 25

fMRI Functional magnetic resonance imaging. 3, 20, 23, 24

FOV field of view. x, 28, 59, 60, 95, 96, 98, 125

GSR Galvanic skin response. 20, 21

HMD Head-mounted display. 11, 27, 93, 98, 128, 129

HP hyperprameter. 111

HRV Heart rate variability. 21, 23

IM Individual Matrix. 41, 42

LOSOCV Leave-one-Subject-out Cross-Validation. 111

MAM Morph-A-Mood. 15, 16, 28

MDI Mean Decrease Impurity. 46

ML Machine Learning. 10, 47

NPC Non-player character. 62, 109

PAM Pick-A-Mood. 16, 28

PL Preference Learning. 10, 38, 40

RESP Respiratory Rate. 20

RF Random Forest. ix, 39, 83, 86, 111–114

SAM Self-Assessment manikin. 15, 16, 23, 28

SDA Signed Differential Agreement. x, 44, 96–98

SVM Support Vector Machine. 39, 46, 80, 83–85

VE Virtual Environment. 2–4, 10, 11, 27, 28, 31, 35, 49, 50, 58–60, 89, 95, 103, 116, 120–122, 128

VEs Virtual Environments. 13, 17, 26–28

VR Virtual Reality. x, 17, 18, 23–29, 44, 59, 89, 91–103, 118, 120, 121, 125, 129

VSPs Video sharing platforms. 13, 30, 31

XR Extended Reality. 129

1

Introduction

How do spaces shape our emotions and influence our perceptions? Can designers and architects be informed about the potential affective impact of forthcoming changes in spaces? The complexity of determining the architecture's impact on affect has been a non-trivial topic for centuries. Several schools of thought have emerged throughout dating back to ancient Greece. *Aristotle* was one of the first to highlight the importance of creating structures that not only fulfill the need for function but also to intrigue positive emotions (Destrée, 2021). During the Renaissance, *Leon Battista Alberti* and *Andrea Palladio* studied closer this relationship between architecture and human perception, emphasizing the principles of symmetry and proportion, but also light and color, on the observer's experience of space (Matuke, 2016). In the 18th and 19th century, the Industrial Revolution led to a rapid transformation of urban landscapes, leading to a growing interest in the social and psychological effects of architecture. Architects and urban planners began to think about how to design buildings and cities that would promote human well-being.

The emotional impact of architecture was studied systematically by psychologists and social scientists in the early 20th century. The German psychologist and *Gestalt* theorist *Kurt Lewin* conducted experiments on the effects of different office environments on worker productivity. Even though his experiments were not directly related to design and spatial organization of office spaces, *the Harwood* experiments and the studies that followed thereafter, displayed consistent findings regarding the impact of natural light, fresh air, and open office plans on employee productivity, satisfaction, and well-being (Marrow, 1977). Alexander, 1977 in *"Pattern Language"* developed an approach of identifying the affective impact of patterns in urban planning and architecture, promoting the idea of *human-centered design*, while Pallasmaa, 2012 in *"Eyes of the skin: Architecture of the senses"*, take this further and encourages us to look further than vision when trying

to determine the impact of space and consider our remaining senses of touch, smell and hearing.

The late growing awareness of the importance of mental health and well-being and the emotional impact of architecture (Lomas et al., 2022), has changed the Architecture, Engineering, and Construction (AEC) practitioners approach in designing buildings that promote positive emotions and support human flourishing. This interest on the individual or the user of space transferred beyond architecture and interior design, to the design of everyday things and services introducing the user as an integral part of the design process, this process being referred to as co-designing or participatory design (Sanders et al., 2010).

Additionally, several standards and guidelines have been developed aimed at comfort, well-being, and safety in built environments. These standards cover areas such as thermal comfort, lighting, acoustics, air quality, spatial arrangement, and general human-centric design principles. The most prominent ones are Leadership in Energy and Environmental Design (Altomonte and Schiavon, 2013), also known as LEED, the WELL Building Standard (Well, 2014), ISO 16817: Building Environment Design - Indoor Environment (ISO 16817:2017,) and the Daylighting Standards and Guidelines, EN 17037 (Šprah and Košir, 2019).

Given the above, we acknowledge the contribution that emotion has acquired in the process by which we design our spaces. There is a constant need for new methodologies and perspectives that assist in understanding this relationship. The goal of this thesis is to enhance our knowledge of the relationship between space and emotion by leveraging the latest advancements in psychology and technology. Finally, the significance of this endeavor is emphasized, as it aims to benefit designers, researchers, space occupants, and policymakers.

1.1 | Motivation and Problem Statement

With this growing interest in the affective impact of space, researchers in the field have increasingly explored this relationship using Virtual Environment (VE)s. VEs offer significant advantages for studying the relationship between space and human responses, particularly due to their ability to provide highly controlled and repeatable conditions. Researchers can precisely manipulate and maintain various spatial elements, such as lighting, geometry, and scale, allowing for the isolation of specific variables to observe their effects on affective responses. This level of control is difficult to achieve in real-world environments, where external factors can interfere with the consistency of ex-

perimental conditions. Moreover, VEs enable the exact replication of spatial settings, ensuring that all participants experience the same stimuli, which enhances the reliability of the findings and facilitates comparisons across studies. This direction has become a systematic method for data collection for studies in Architectural design, Psychology and Social studies (Slater and Sanchez-Vives, 2016; Christofi et al., 2020; Banaei et al., 2017b; Schneider et al., 2018; Gómez-Puerto et al., 2018).

Most of such studies that investigate the relationship between space and affect tend to rely on intrusive methods such as Electroencephalogram (EEG) devices (Ruta et al., 2023; Banaei et al., 2020) and Functional magnetic resonance imaging (fMRI) technology (Vartanian et al., 2015). While these approaches provide valuable insights into brain activity and emotional responses, they come with several drawbacks. First, they can be uncomfortable or restrictive for participants, potentially altering the natural affective responses the study aims to capture. Second, the complexity and cost of such equipment make these methods less accessible, limiting their widespread use. Lastly, the artificial nature of lab settings and the presence of intrusive devices may reduce ecological validity, as participants' experiences may not fully reflect real-world interactions with space.

Studies that gather affective responses using non-intrusive methods often rely on real-time evaluations (Bianconi et al., 2021; Gomez-Tone et al., 2021) or post-stimulus surveys (Presti et al., 2022). These approaches capture valuable data in various forms, including free-response and forced-response formats, providing meaningful insights into how different spatial configurations impact emotion. However, many of these studies overlook the temporal dimension of architectural experience, frequently reducing their analysis to a single label or rating that represents an entire spatial condition, which may oversimplify the complexity of dynamic experiences in space.

Moreover, the number of available affect datasets within the field of architectural design remains limited. Many studies present their stimuli and findings without sharing the underlying data or material, which restricts further exploration and adaptation by other researchers. This lack of open data slows potential advancements in the field, as shared datasets could enable more robust comparisons and foster the development of more comprehensive affective models.

We the above in mind we summarize that the relationship between architectural spaces and human emotions faces several key challenges. First, there is a lack of quantitative methods to evaluate the impact of space on affective experiences, limiting the ability to systematically assess this relationship. Second, the availability of high-quality data to model the interplay between spatial design and human emotions is insufficient, creating barriers to accurate and robust modeling. Third, current methods for capturing

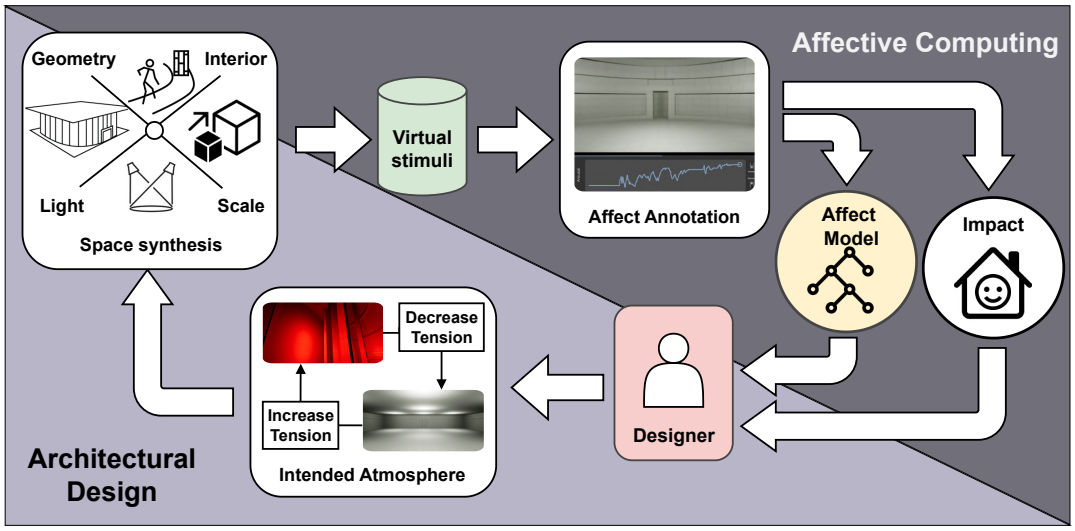

Figure 1.1: The proposed dissertation framework, from space synthesis and construction of synthetic stimuli to gathering affect annotations and affect modeling, informing the designer on the affective impact of space and back to space synthesis again.

affective responses are often intrusive, making it difficult to gather data in a naturalistic and scalable manner.

1.2 | Proposed Solution

Affective Computing (AC), i.e. *"computing that relates to, arises from, or deliberately influences emotions"* (Picard, 2000) is an interdisciplinary area, bridging computer science with psychology. The field of AC displays great examples of methods and tools in capturing of continuous user annotations under various conditions and stimuli using non-intrusive means. Most common areas of studies of AC are speech (Cowie et al., 2000), Human-computer interaction, videogames (Liapis et al., 2018; Lopes et al., 2015; Melhart et al., 2019), Marketing and Consumer Behavior (Alajmi et al., 2013)), Music Greer et al.; Greer et al., 2020; 2019, video and 360 video Fayn et al.; Xue et al., 2022; 2020b and VEs Yang and Kalantari, 2022.

Even though the field demonstrates a plethora of areas that are investigated using methods and tools for continuous affect annotation, architectural design studies here are limited. This work sees the benefits of this field in producing enriched affect datasets of environments to quantitatively study the impact of spaces.

Regarding the problem statement formulated in the previous section, this work pro-

poses a closed-loop solution (see Figure 1.1) that if first defines the primary elements of a space and the it utilizes them to construct virtual environments. The next step involves creating dynamic stimuli, allowing observers to interact with and annotate their affective states in real-time. Based on these annotations, an affective model is built, capturing the relationship between spatial conditions and emotional responses. This model then serves as a tool for designers to predict and refine the affective experience during transitions in synthetic environments. In turn, this feedback informs the space synthesis process, enabling designers to make iterative adjustments to elicit desired emotional responses from users.

It is important to note that the proposed solution is not intended to function as a closed-loop system, such as the experience-driven content generation approach introduced by Yannakakis and Togelius, 2011. While this work examines the methods used to derive affect and experience from virtual stimuli, it does not encompass content generation techniques based on data-driven approaches, which are commonly explored in procedural content generation research (Shaker et al., 2016).

The proposed solution adopts a bottom-up approach by adjusting primary elements of space and explores how specific design features influence affective responses. This method aligns with common practices in architectural and design research, where isolating and modifying environmental parameters helps to assess their impact. Through this approach, relationships between spatial features and emotional responses can emerge, be analyzed, and quantified.

1.3 | Research Questions

In an attempt to address the problem statement of this dissertation, the following research questions are formulated to explore the three main areas of this work: architectural parameters, virtual stimuli and capturing affect.

- RQ1. How can we reliably estimate the impact of architecture elements to human affect?
  - RQ1.1 Can we observe similarities across different types of virtual stimuli.
  - RQ1.2 How do key elements in architectural design affect perceived levels of arousal and pleasure (valence)?
  - RQ1.2b How do key elements in architectural design contribute to affect emotion manifestations during gameplay?

  - RQ1.3 Are absolute or relative measures of affect more appropriate measures of the impact?

  The first research question is addressed via its sub-questions focusing mainly the reliable impact of key elements in architecture, both in terms of reported affect in arousal-pleasure dimensions but also in regards to manifested affect states conveyed through expressions during gameplay.

- RQ2. How effectively can we model affective responses (arousal & pleasure) based on architectural key elements across different stimuli?

  - RQ2.1 Can we minimize temporal biases through effective data processing?
  - RQ2.2 To what extent can we model elicited arousal and pleasure of virtual spaces?
  - RQ2.3 How can we capture first-person annotations during the act of virtual exploration and what is the impact of display type during this experience?

  The second research question and its sub-questions addresses the modeling and capturing aspects of the dissertation across the three different investigated stimuli of videos, interactive virtual environments and games.

1.4 | Contributions

The following contributions of the dissertation are detailed further within the final chapter (see Chapter 8) and are as follows:

- **A continuous affect capturing methodology of non-static spatial stimuli using the PAGAN platform and RankTrace annotation method:** Addressing the gap in continuous affect annotation methodologies, this dissertation employs the RankTrace annotation method for evaluating non-static stimuli in both video-based and virtual environments. This approach advances the field by enabling a more dynamic and temporally precise measurement of affective responses.
- **An extended processing framework for continuous affect annotations of non-static spatial stimuli:** Building upon the works of Cowie and McKeown, 2010, this dissertation introduces an extended framework for processing continuous affective data captured in real-time within virtual environments. This framework refines the methodological pipeline for analyzing affect annotations, offering a structured approach for handling temporal affect data in immersive contexts.

- **Two publicly available corpora -Affrooms12 and Affrooms24- containing:**

  1. Raw continuous affect ratings (arousal and pleasure) provided by both expert and non-expert annotators.
  2. Processed affect annotations alongside their corresponding results, documented within this dissertation and the related published works.
  3. A collection of architectural walkthrough video sessions that can serve as a foundation for further research and methodological extensions in affective computing and spatial experience analysis.

- **A continuous affect capturing methodology within Virtual environments.** Expanding beyond passive stimulus-based studies, this dissertation implements the RankTrace affect annotation method within an immersive virtual walkthrough experience. Participants actively engaged with their surroundings in a **3-degrees-of-freedom (3DoF)** environment while continuously reporting their affective states (pleasure). This contribution enhances the understanding of affective experiences in interactive, exploratory virtual spaces.

- **A data collection and affect extraction pipeline for capturing elicited affect of readily available streamed content in the wild.** This dissertation introduces **OutlastAFF**, a publicly available dataset comprising approximately 8.5 hours of YouTube live-streamed playthroughs of a single level ("Asylum") from the horror game Outlast. The dataset has been processed using multi-modal affect recognition techniques, leveraging pre-trained models to extract affective responses from streamers' facial expressions, vocal features, and verbal utterances during gameplay. This contribution provides insights into the relationship between spatial parameters, gameplay events, and emotional responses, enriching research in affective computing, game studies, and virtual environments.

- **Models of Affect and Impact of Space across different Stimulus type and affect gathering methodologies.** Three models of affect were developed using **Random Forest classification**, predicting changes in affect dimensions (**Arousal** or **Pleasure**) based on variations in key spatial parameters. Additionally, a more detailed analysis was conducted to assess the impact of each spatial parameter on affective responses using **linear agreement analysis**. This approach provides deeper insights into how specific spatial features influence emotional states, considering differences in perception across various stimulus types and contexts.

1.5 | Publications

The publications that were the outcome of this work and their contribution to the dissertation's overall structure are as follows:

1.5.1 | Within the scope of this Thesis

- Xylakis, Emmanouil, Antonios Liapis and Georgios. N. Yannakakis, "Architectural Form and Affect: A Spatiotemporal Study of Arousal." In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2021

  This paper contributes to Chapter 3 and Chapter 4 and the results of the study are detailed in Chapter 5. It is the first study on continuous affect annotation of video stimuli. The resulting dataset of the study is the **Affrooms24** dataset containing affect annotations of Arousal of 3 expert annotators.

- Xylakis, Emmanouil, Antonios Liapis and Georgios. N. Yannakakis, "Affect in Spatial Navigation: A Study of Rooms." In *IEEE Transactions on Affective Computing*. IEEE, 2024.

  This paper contributes to Chapter 3 and Chapter 4 while its results are detailed in Chapter 5. It is the second study on continuous affect annotation of video stimuli. **Affrooms12** is the resulting dataset of this work that contains continuous affect annotations of Arousal and Pleasure, from 76 annotators gathered through crowdsourcing platforms.

- Xylakis, Emmanouil, Ali Najm, Despina Michael-Grigoriou, Antonios Liapis, and Georgios 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.

  This paper contributes to Chapter 3 and Chapter 4 and the study's results are presented in Chapter 6. The paper focuses on using interactive Virtual environments as affect elicitors allowing users to annotate during their experience.

- Xylakis, Emmanouil, Antonios Liapis and Georgios. N. Yannakakis, "The Scream Stream: Multimodal Affect Analysis of Horror Game Spaces." In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2024.

  This paper contributes to Chapter 3 and Chapter 4 and the study's results are presented in Chapter 7. The paper documents the study on using *Let's play* videos

of streamed gameplay of Horror games, as means to extract affect regarding the impact of game spaces and play.

1.5.2 | Outside the scope of this Thesis

- Barthet, Matthew, Chintan Trivedi, Kosmas Pinitas, Emmanouil Xylakis, Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. "Knowing Your Annotator: Rapidly Testing the Reliability of Affect Annotation." .In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*. IEEE, 2023

  Even though the research aim of this paper is determined as out of scope, it contributes in the general discussion on minimizing annotator bias for continuous annotation tasks. The study proposes a quality assurance methodology for continuous annotation tasks, aimed at determining . Assuming that the annotation tasks rely on audiovisual stimuli (videos), this work proposes and evaluates two tests: a visual and an auditory QA test. As the focus of this dissertation lies on affect annotation of visual stimulus in both interactive and non-interactive environments, this work is not discussed further in the chapters that follow.

1.6 | Dissertation Structure

Figure 1.2 illustrates the structure of the methods chapters. In more detail the dissertation is structured as follows:

- **Chapter 2** reviews the relevant literature in three primary areas, *Architectural design*, *Affect Modeling* and *Applications*. The first section provides an overview of key architectural design concepts, focusing on design elements and the various media (e.g., images, videos, virtual environments) used in studies to explore their impact on affective responses. This is followed by a discussion of spatial representations commonly used by scholars in the study of architectural spaces. The second section offers a theoretical background on affect modeling across different types of media and stimuli, with an emphasis on methods for capturing affect, including self-report annotations and affect recognition techniques. The chapter concludes with an exploration of relevant tools, datasets, and environments that were used to address similar topics.
- **Chapter 3** outlines the key data processing methods, algorithms, and modeling techniques used widely in AC studies and employed in this dissertation. This

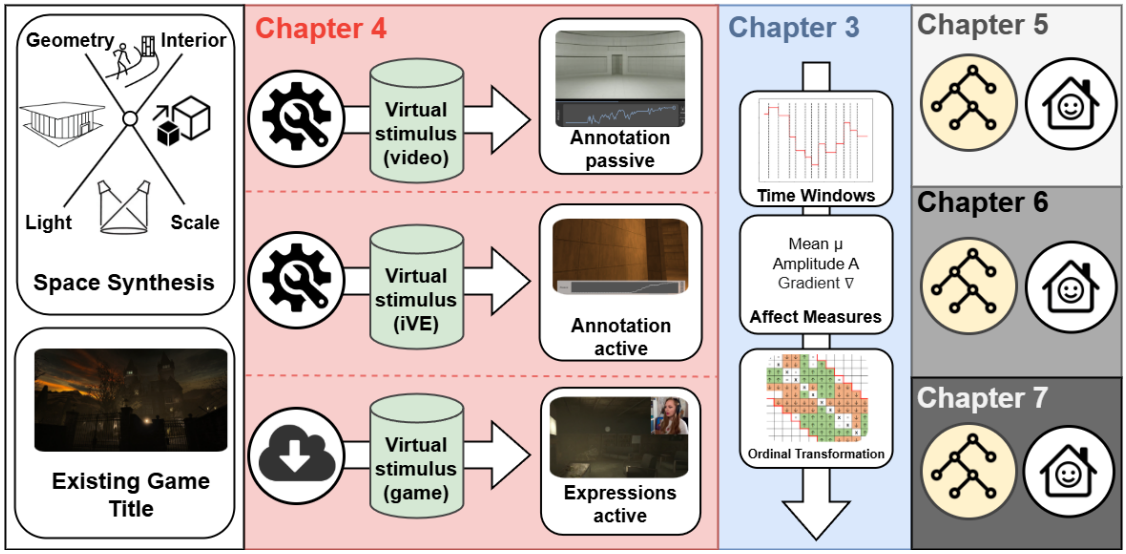

Figure 1.2: Dissertation methodology chapters structure from stimuli construction to data collection and processing, leading to model building and assessing space impact.

chapter is structured around the general processing pipeline that is used throughout the four studies, covering steps such as data pre-processing, feature extraction, inter-rater assessment and modeling. This chapter also discusses extensively modeling techniques for predicting affect, including Machine Learning (ML) approaches such as Random Forests, Ranking, Preference Learning (PL) and the evaluation measures used to assess model performance.

- **Chapter 4** focuses on the data collection methodology across the four studies. Each study is described in terms of its experimental design, participant recruitment, and the specific nature of the affective data collected. The chapter provides a detailed account of the methodologies used to gather affective responses, from pre-recorded video stimuli to real-time interactions in VEs and affect expression in games, highlighting the diversity of contexts in which affect is captured.

- **Chapter 5** presents the results from the first two studies on affect annotation using pre-recorded videos of synthetic room traversals, referred to as *Affrooms24* and *Affrooms12* throughout the dissertation. *Affrooms24* served as a pilot study using expert annotators, while *Affrooms12* utilized a crowd-sourced sample of non-expert participants. This chapter explores how interior spatial parameters influenced arousal and pleasure ratings, with machine learning models, particularly Random Forests, used to establish the relationship between affect dimensions and

spatial features.

- **Chapter 6** examines affect annotations during interactions with a VE using two distinct display formats: a Head-mounted display (HMD) and a Desktop Display. The chapter presents results from a comparative study where participants engaged in in-lab sessions, providing affect annotations while using both display types. The study also evaluates participants' experiences in terms of usability, distraction, and presence, while analyzing how different environmental parameters influenced affective responses across the two media.

- **Chapter 7** focuses on the final study, which introduces the *Outlast Asylum Affect Corpus*. This study diverges from the previous ones by analyzing free-form expressions (facial, vocal, and linguistic) as indicators of affective responses during gameplay in a horror game context. The chapter investigates how various spatial conditions within the game influenced players' expressed affect and models the relationship between these spatial features and emotional responses.

- **Chapter 8** is the final chapter of the dissertation. This chapter concludes this work and outlines the main contributions and limitations. Limitations for each of the four user studies and more general obstacles related to the overall approach of the dissertation. Finally, the extensibility section addresses some of the limitations and highlights the continuation of this work, proposing different directions for this research.

Additionally, to improve readability and organization of this dissertation, *GPT 4-o*[1] was used strictly for editing purposes. Its role was limited to improve linguistic clarity and refine the document's overall structure. At no point was it employed for generating new content, ensuring the originality and integrity of the research presented.

1.7 | Summary

This chapter outlines the key questions, challenges, and motivations that shape this dissertation. It highlights the gaps in the study of space and affect, particularly the neglect of temporal dynamics in affect annotation methods and the scarcity of affective datasets. To address these issues, this work proposes the investigation of temporally-based stimuli and the use of continuous affect annotations to reveal how key spatial

[1] `https://platform.openai.com/docs/models`

elements influence affect. The goal is to develop affect models that can inform and support designers during the design process.

The research questions are framed to focus on the stimuli and the reliable collection of affect annotations, with a particular emphasis on the temporal dimension. This chapter also includes a list of related publications and projects that contributed to the overall research. Furthermore, it outlines the theoretical framework underpinning the study and provides a road map for the chapters that follow.

2

Related work

This dissertation explores the affect-inducing capabilities of space by examining its individual elements and combining them with theories of emotion and applications of affect modeling. The current chapter provides a comprehensive overview of key studies in the fields of space and affect modeling.

Section 2.1 introduces the theoretical foundation, focusing on the most widely used representations of emotions. It discusses the advantages and limitations of each approach, providing a balanced view of the emotional models commonly applied in affect research. Section 2.1.1 presents the primary frameworks and tools for affect capturing, detailing studies that range from continuous to discrete emotion capture. This section also delves into how emotions are annotated by both humans and machines, highlighting the methodologies used to quantify affective responses in a reliable manner. Section 2.1.2 addresses the critical considerations involved in the treatment of affect data. It emphasizes the distinctions between nominal, interval, and ordinal data, offering a detailed review of how different studies handle these data types to answer their core research questions. Additionally, this section examines methods that enhance reliability and inter-rater agreement through varying data treatment techniques. Section 2.2 shifts the focus to the spatial parameters that impact affect, identified through an extensive review of studies in architecture and affective response. This section also outlines the methods used to construct spatial representations, which are key to drawing meaningful inferences between space and affect. Finally, Section 2.3 explores the applications used across various studies to elicit and capture affect. These applications utilize a wide range of stimuli, from videos and stereoscopic images to more immersive mediums such as Virtual Environments (VEs) and video games. Additionally, Section 2.3.3 presents how data collection can occur "in-the-wild" using Video sharing platforms (VSPs), and the considerations when capturing these type of affect data. The section highlights how

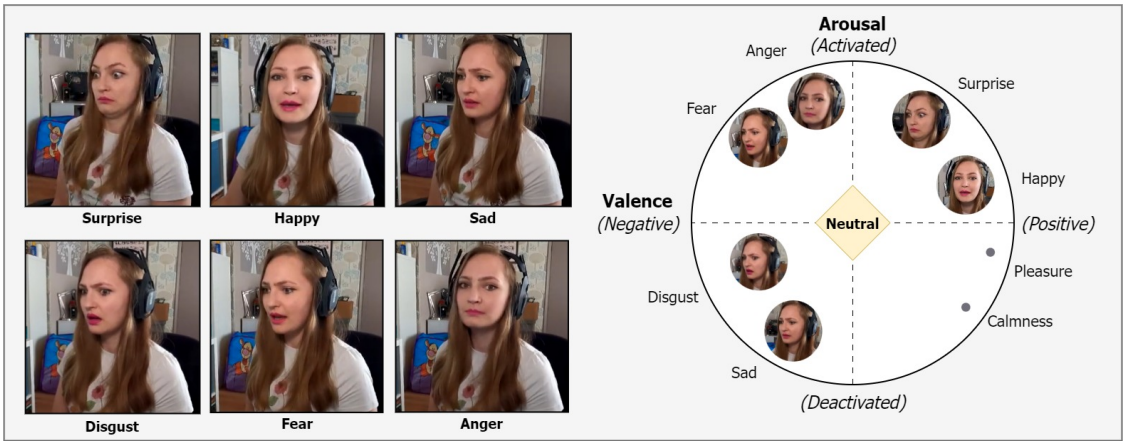

Figure 2.1: Discrete Affect labels (*left*) and Affect dimensions *(right)*.

these diverse stimuli contribute to understanding the affective impact of space in different contexts.

2.1 | Affect Theory Background

This section outlines the theory and models of representing, capturing and treating affect across various stimuli, not only for architectural design research but also in virtual worlds and games research. The focus of the dissertation lies mainly on the dimensional models of affect of arousal and valence, emphasizing on ordinal approaches, representations and treatments of affect.

The two main theoretical models commonly found in affective research are based on either **discrete emotional categories** or **dimensional approaches** (Yannakakis and Martinez, 2015; Lopes et al., 2017b; Harmon-Jones et al., 2017). Both approaches of depicting emotions are illustrated in Figure 2.1, using snapshots taken from the *Outlast Asylum Affect Dataset* (see Chapter (6)).

When emotions are described in a **discrete or categorical manner**, researchers often follow **Ekman's theory of basic emotions** (Ekman, 2004). According to Ekman, there are six core emotions categories—happiness, anger, fear, disgust, sadness, and surprise—that are considered biologically innate and universally recognized. These emotions are thought to occur distinctly, separate from one another, each representing a unique emotional experience. While this model has been foundational in understanding human emotion, it assumes that emotions exist as discrete states, which can limit its ability to fully capture the complexity of real-world emotional experiences. In reality,

emotional responses are often more nuanced and cannot always be neatly classified into one category.

To address these limitations, many studies have sought to enrich **Ekman's categorical framework** by adding additional emotional states or introducing dimensions within these basic categories. For example, tools like the **Self-Assessment manikin (SAM)** (Bradley and Lang, 1994) or or **Morph-A-Mood (MAM)** (Desmet et al., 2016) offer more flexible representations. These tools enable researchers to assess emotions on a continuum rather than restricting them to fixed labels, providing a more nuanced and comprehensive understanding of emotional experiences. SAM, in particular, measures emotions along dimensions such as valence and arousal, offering greater insight into the intensity and nature of a person's emotional state, while MAM allows for the visualization of dynamic emotional transitions.

Dimensional models typically describe emotions within a two-dimensional space of **Arousal** and **Valence**, or sometimes a three-dimensional space with the additional dimension of **Control** or **Dominance**. The most widely used dimensional frameworks include **Russell's Circumplex Model of Affect** (Russell, 1980), which operates on the arousal-valence axes, and **Mehrabian's Pleasure, Arousal, Dominance (PAD)** framework (Mehrabian, 1996), which adds a third dimension to capture the sense of control or dominance over an emotional situation. Emotions here are plotted within a continuous space, with arousal indicating the level of activation or intensity, and valence representing the positive or negative nature of the experience. For example, emotions like excitement are placed at high arousal and positive valence, while sadness occupies low arousal and negative valence. The idea behind these models is that all emotions can be expressed as coordinates within these dimensional spaces, allowing for a more dynamic and flexible representation compared to categorical models.

In this dissertation we choose to work with dimensional models of affect to access more fine-grained information than the categorical counterpart. Having access to the degree or volume a stimulus might induce, is the foundation for building valid models of affect.

2.1.1 | Affect Annotation Methods

In practice, the ways in which affect states can be captured fall under two main categories: **manual self-reports** and **objective physiological measures** (Yannakakis and Togelius, 2018); Pinilla et al., 2020. Self-reports collect subjective user feedback or experience to stimuli either as *free* or *forced* response. Free response approaches like think aloud and open ended interviews El-Nasr et al., 2016 allow participants to express their

opinions, affective states, or experiences in an unstructured way, providing more depth and insight than structured questionnaires or forced responses, but can be demanding to analyze. Forced-response reports use structured questionnaires and annotation tools prompting participants to respond following an experiment protocol defined by the researcher. With this method, the resulting data are much easier to process but can be proven limiting for the participants to express themselves.

The SAM tool, introduced by Bradley and Lang, 1994, offered a visual, non-verbal tool that enables participants to rate their emotional state across three dimensions: valence, arousal, and dominance. It uses a series of cartoon-like figures to represent emotions, making it easier for users to express their feelings without needing to rely on complex verbal descriptions. By simplifying the rating process, SAM helps reduce ambiguity and increase the reliability of affect data. The Pick-A-Mood (PAM) annotation tool by Desmet et al., 2016, was designed to enhance emotion annotation by offering a selection of recognizable cartoon faces that represent different moods. Participants "pick" a mood that closely matches their current emotional state. This tool is particularly useful in settings where quick, intuitive responses are needed, and it offers a more direct form of emotional reporting compared to traditional dimensional models. Taking these methods further, the MAM tool developed by Krüger et al., 2020, introduced an interactive interface where users can adjust the facial features of a cartoon figure to match their emotional state. By allowing continuous adjustments to facial expressions, MAM provides a dynamic and personalized way of capturing emotions, which adds depth to the emotional data collected and addresses the limitations of both categorical and dimensional models. It bridges the gap between static discrete labels and the fluid nature of emotions in real-world scenarios.

The types of affect data captured by researchers are typically interval, ordinal, or nominal. Physiological studies, which rely on bodily measurements, generally capture interval data. In contrast, self-reported annotations of emotional experiences can fall into any of the three categories. Traditionally, psychologists have captured emotions and affective states in nominal or interval forms, as seen in models like Russell's Circumplex of Affect (Russell, 1980) and Ekman's basic emotions (Ekman, 1992). However, recent advances in AC have shifted toward ordinal data.

The literature on emotion theory highlights the advantages of viewing emotions in a relative rather than absolute manner, which tends to yield higher reliability. *Adaptation Level Theory* (Helson, 1964) posits that our perception of experiences is relational, influenced by our memory of recent experiences. Consequently, annotators can provide more accurate feedback on their affective states when they consider their emotional *'anchors'* in relation to more recent experiences. This makes ordinal data, which reflects

these comparative judgments, more desirable when measuring affect and analyzing the collected data in emotion modeling.

2.1.1.1 | Methods and Tools for Affect annotation

The methods and frameworks for collecting user annotations vary according to the available data types that can be collected and the questions that researchers seek to answer. Architectural studies that seek to capture the affective impact of the built space make use of a plethora of tools acquiring both categorical and interval annotations. Schneider et al., 2018 on their work in *VREVAL* sought to collect both categorical and interval data within VEs during participant's navigation. Using VR displays and controller input, users were prompted to navigate and report their experience in free-form, using descriptive labels and 5-point likert ratings. In the study of Shemesh et al., 2017 a similar setup with VR and EEG was chosen but this time participants had to provide their feedback in post-experience sessions with likert scale responses and pairwise comparisons. This post-stimuli approach to data collection is a method that researchers choose when they seek to have an uninterrupted experience aiming at lowering the cognitive effort while collecting physiological data. Ruta et al., 2019 compared real-time and post-stimuli reporting of wall projected images of spaces acquiring both pairwise-preference ratings and liker ratings regarding familiarity, preference and complexity. Findings reveal the importance of *the mere exposure effect* (Zajonc, 1968) inline with *adaptation level theory*, suggesting that valence ratings are affected by our over time exposure, as we adapt to a stimulus, our response to it becomes less intense. As such collecting post-stimuli reports is a common practice among architectural studies seeking to avoid stimuli over-exposure in order to collect more reliable annotations from the participant's side. Presti et al., 2022 followed this procedure in their experiment protocol allowing raters to experience a virtual space and interact with it for a fixed duration of 12.5 seconds. Immediately after, participants were instructed to input their arousal and valence ratings for each space on a Visual Analog Scale (VAS).

Continuous Affect annotation: One area that hasn't not been fully explored by researchers studying affect and architecture is the one of collecting continuous self-reports. Continuous self-reports has been explored as a method extensively by media researchers in works of film (Fayn et al., 2022), video games (Melhart et al., 2019; Melhart, 2021), VEs (Xue et al., 2021; Zhang et al., 2020), serious games research (Christofi et al., 2020), speech and sound (Cowie et al., 2000; Parthasarathy et al., 2016). These works seek to study how affect estimates unfold over time while giving access to more fine grained data rather than a single point or single preference on a given stimuli. Con-

tinuous data annotations come usually in the form of dimensional annotations (Cowie et al., 2000). The *Feeltrace* annotation tool (Cowie et al., 2000) is one of the first tools that was developed to capture continuous dimensional affect on the Activation-Evaluation space (or similarly defined by Russel's work as the Arousal-Valence space). Additionally, continuous tools for capturing discrete affect descriptions were developed and used across linguistic studies like *ANVIL* by Kipp, 2001 and *ELAN* by Wittenburg et al., 2006.

Acknowledging the importance of anchoring and the relative nature of emotion elicitation, the *Affectrank* annotation tool (Yannakakis and Martinez, 2015) was developed to capture rank data between 8 discrete affect categories on a two-dimensional affect space. Similarly inspired by this approach the *RankTrace* annotation tool (Lopes et al., 2016) was developed, allowing the capturing of affect changes across a single dimension, aimed at minimizing the required cognitive effort of multiple dimensions. Additionally, the RankTrace tool includes a graphical representation of the previous affect ratings in the form of a one-dimensional continuous signal overlayed on top of the stimulus, serving as reference for the rater during the annotation, instead of constraining them to assess stimuli in a fixed, absolute manner. Lastly, its "limitless" or unbounded feature serves as a way to track relative changes throughout the produced trace allowing for ordinal data treatment, a most desired approach when analysing data of affect.

Similarly *RCEA* (Zhang et al., 2020) and *RCEA-360 VR* (Xue et al., 2021) were developed as continuous annotation methods of "emotion intensity" for immersive media. Both methods prompt raters to input emotion intensity estimates during their experience using VR controllers. Affect rating intensity is reported back to the rater using two peripheral vision techniques with changes in shaded geometry's color (Halolight) or geometry's size (Dotsize). Lastly, based on the assumption that that the affect space is unipolar the the Full Throttle was developed to address the needs for continuous annotations of 2 unipolar dimensions during the process of watching captivating movie clips. The *Full Throttle* annotation method (Fayn et al., 2022) uses a two-handed reporting system with two separate graphical representations allowing for gathering of non-bipolar affect ratings and suggests more usable and accurate ratings when compared to mouse method as input system.

While affect annotation of interactive experiences like in the cases of *VREVAL*, *RCEA* and *RCEA-360 VR* suggest that continuous self-reports can be collected reliably, there is a factor to be consider and that is the effort required to interact with the synthetic experience. In the case of affect annotation in video games whereby the medium requires the user's full attention, the simultaneous reporting of affect data is most of the time not possible. Under these conditions *stimulated recall* (Lankoski et al., 2015) is proposed whereby annotators initially interact with a game while their gameplay is being

recorded. Once the gameplay session is over, the annotator is being shown his/her recorded gameplay and is asked to recall their experience and annotate accordingly. Lopes et al. utilized RankTrace in their study in the same way, asking participants to interact with their developed Horror game and after that annotate their experience of Tension within the game. In their work on the *AGAIN* dataset, Melhart et al., 2021a, employed the stimulated recall method to gather arousal ratings across 9 games of 3 different genres (Racing, First person shooters and Platformers). Similarly as in the work of Lopes, annotators were given the opportunity to first interact with the environment and annotated their affect at a later stage using stimulated recall. Stimulated recall is also investigated by Yang and Kalantari, 2024 for their Continuous Uncertainty Annotation tool (CUA), and being compared with an alternative real-time tool (RCUA). In their experiment, participants where instructed to annotate their levels of uncertainty during a navigation task in closed space and reported their levels of uncertainty using the RCUA annotation tool with *5 Hz* annotation frequency. During the navigation task participants recorded first-person footage with cameras (GoPro Max) positioned at chest level. The recorded footage was used at the second phase of this experiment where the same participants reported once again their uncertainty levels using the CUA method. Results display the validity of both real-time and stimulated recall approaches with the RCUA method displaying superior performance at capturing detailed dynamics of the human experience. An overview of the reviewed annotation interfaces are depicted in Table 2.1.

2.1.1.2 | Third-person Annotation for Ground Truth

Collecting affect ratings from immersive and intense experiences like video games can be a challenging task due to participants' deep engagement with the synthetic environment. The methods discussed above provide effective solutions to overcome these challenges. However, one limiting factor that these approaches might face, and specifically the stimulated recall method, is the underestimation of the felt sensation due to memory gaps. This is being referred to as the *memory-experience gap.* (Miron-Shatz et al., 2009). Miron-Shatz et al., conducted two separate studies among female participants (N= 810, Study 1, and N= 615, Study 2) and examined how reconstructed experiences of previous episodes aligned with reports acquired during the actual events. Their results show a clear gap in terms of negative experiences as being less intense than they actually were, or conversely, participants might recall positive experiences more fondly than they were in reality.

To overcome such limitations, *a third-person annotation* method is employed whereby

Annotation Interface	Mode	Session	Study
RankTrace	Dimensional	RT	Lopes et al., 2017b
PAGAN	Dimensional	RT	Melhart et al., 2019
Gtrace	Dimensional	RT	Cowie et al., 2013
Feeltrace	Dimensional	RT	Cowie et al., 2000
AffectRank	Dimensional	RT	Martinez et al., 2014
RCEA	Dimensional	RT	Zhang et al., 2020
RCEA-VR	Dimensional	RT	Xue et al., 2021
RCUA / CUA	Dimensional	RT/Post	Yang and Kalantari, 2024
Geneva emotion wheel	Mixed	Post	Sacharin et al., 2012
VAEW	Mixed	Post	Abukhodair et al., 2024
Emojigrid	Mixed	Post	Toet and van Erp, 2019
Morph a Mood (MAM)	Mixed	RT	Krüger et al., 2020
Pick a Mood (PAM)	Discrete	RT	Desmet et al., 2016
PrEMO	Discrete	RT	Crippa et al., 2012
Full throttle	Dimensional	RT	Fayn et al., 2022
EMuJOY	Dimensional	RT	Nagel et al., 2007

Table 2.1: An overview of annotation interfaces used in self-reporting studies, indicating their mode (Dimensional, Discrete or mixed) and their support for real-time annotation.

expert annotators (experienced gamers, game designers or researchers) rate the experience of the player (Yannakakis and Togelius, 2018). An additional advantage of utilizing third-person annotations is the availability of multiple perspectives of affect annotations on the same exact stimuli (i.e. the same video of a player playing super mario) allowing for a better approximation of a **ground truth**. Multiple studies look into third person annotation with the aim of estimating a ground truth, by employing either a large sample of raters through crowdsourcing (Barthet et al., 2023) or limiting the sample size to include only experts raters (Barthet et al., 2023; Mavromoustakos-Blom et al., 2023).

2.1.1.3 | Affect Capturing via Physiological Input

In addition to the collection of affect report methods via subjective approaches as documented in the previous sections, AC researchers collect objective affect responses via physiological input. The most common methods that studies use include brain imaging techniques such as EEG measurements and fMRI scans, Galvanic skin response (GSR), Electrocardiography (ECG), Electromyography (EMG), Respiratory Rate (RESP) and eye-tracking.

Brain-imaging methods analyze the various regions of the brain with either installed electrodes (EEG) or by measuring the blood flow in the brain (fMRI), and seek to under-

stand which regions of the brain are activated during an architectural experience. Signals that are acquired via EEG are usually analyzed based on defined frequency bands with the delta band at 1–3Hz, the theta band at 4–7Hz, the alpha at 8–12Hz, the beta at 13–40Hz, and gamma at 40 Hz and above (Kim and Kim, 2022).

GSR sensors, also known as Electrodermal Activity (EDA) measures changes in skin conductance due to shifts in a person's psychological state. By applying a constant voltage through electrodes, the current flow is recorded, reflecting two types of conductance: tonic (baseline, unaffected by the environment) and phasic (responsive to environmental events). Phasic responses occur when the sympathetic nervous system triggers sweat glands, increasing skin conductance. Studies commonly use GSR wearable devices when investigating tension and arousal during gameplay in horror game titles (Graja et al., 2020; Lopes and Boulic, 2020; Makantasis et al., 2021).

Autonomic nervous systems (ANS) are used as non-intrusive methods compared to brain-imaging techniques and include measurements using Electrocardiography and wearable devices. A common metric used throughout most of the studies is Heart rate variability (HRV) which is either coupled with other biomarkers like EEG (Marín-Morales et al., 2018) or EDA (Formiga et al., 2022; Makantasis et al., 2021)

2.1.2 | Treating Affect

In the study of affect, it is crucial to consider how emotional data is treated, as this influences the accuracy and validity of the analysis. As we described above, affective data usually are described into three main types: Nominal, Interval, and Ordinal. Nominal data categorizes emotions into discrete labels, such as "happy," "sad," or "angry," without implying any order or magnitude between them. Ordinal data ranks emotions in a specific order, suggesting a relative intensity or preference among emotional states, but without assuming equal intervals between them. Interval data, on the other hand, treats emotions as points on a continuous scale, where the difference between values is meaningful, enabling more precise measurements of affective intensity.

Lately though, several studies bring forth incentives suggesting that *Interval* and *Nominal* approaches to treating affect data can be problematic. Yannakakis et al., 2017, highlight several works ranging from media studies, marketing and psychological studies and neuroscience examples that display the superiority on the use of *preference learning* and ordinal approaches. This suggested approach to an ordinal analysis within their scope classifies two main approaches: a *first-order ordinal annotation* where a ordinal scale (or dimension) exists during data collection, f.x. likert scales or pairwise comparisons and *second-order ordinal annotation* where collected data come in the form of

interval or nominal type but are converted to an ordinal scale.

In architecture studies where the affective impact of space is investigated, we see predominately the use of *first-order ordinal annotation*, which come most of the time in form of pairwise comparisons of a simultaneous exposure to two different spatial configurations (Shemesh et al., 2017; Ruta et al., 2019). Other research has highlighted the benefits of using mixed approaches that combine these traditional methods. For instance, Harmon-Jones et al., 2017 discuss the advantages of integrating both dimensional (Interval) and discrete (Nominal) models of emotion. This mixed approach acknowledges the complexity of affect, recognizing that emotions can be both qualitatively distinct and quantitatively measured. By leveraging the strengths of both Nominal and Interval data, researchers can more effectively capture the nuances of emotional experiences, overcoming the limitations of relying solely on one type of data.

In this section, we outlined the dominant theories of affect and reviewed several methodologies for capturing and processing affective responses. Given our focus on continuous affect annotation, we highlight the scarcity of such approaches in architectural design studies, which tend to prioritize the capture of physiological signals. Moreover, we emphasize the importance of treating affect relatively, as this leads to more reliable affect data. The four studies presented in this dissertation address these gaps by collecting continuous, dimensional affect annotations alongside observed expressions, ultimately supporting relative data processing. The following section presents the dominant parameters of space, and outlines the methods in determining their affective impact.

2.2 | Definition of Architectural Elements

To explore how design elements influence affective responses, architectural design research often involves identifying, defining, and adapting the parameters that make up the built environment. This approach allows researchers to uncover and quantify the relationships between spatial features and the emotions they evoke. In this chapter, we establish a common vocabulary by defining the primary elements of space, following a similar approach to Ching, 2023. The literature review identifies five key categories of spatial parameters that are commonly investigated in user studies to assess their affective impact: spatial contour and shape, scale and proportions, openness and obstructiveness, material properties and patterns, and illumination characteristics.

The user studies reviewed in this dissertation are limited to those conducted within the last two decades (2004-2024) and required to include at least one form of subjective

assessment, such as annotations, questionnaires, free responses, think-aloud protocols, or interviews. Some studies also incorporated objective measures, including biofeedback, EEG, and HRV. However, since the focus of this dissertation is on subjective assessments, the results presented here pertain mainly to that area.

2.2.1 | The Affective Impact of Space

The shape and overall **geometry of a space**—whether this features curves, sharp angles, or irregular forms— can have a significant impact on feelings of comfort, tension, or curiosity (Bertamini and Sinico, 2021). Rounded or organic-like forms are often associated with relaxation, while angular designs can prompt feelings of alertness or discomfort. Numerous studies have studied *curvature* of spatial elements not only in terms of affective response but also as a parameter of objects' shape (Bertamini and Sinico, 2021; Dazkir, 2009; Gómez-Puerto et al., 2018). Designers often seek to emphasize curvature in their works, due to its association with reassurance compared to our dislike of angularity and the perceptual threat that this conveys (Ruta et al., 2019). Curvature has also been investigated as a decorative feature in interior openings or in façade design (Ruta et al., 2019; Chamilothori, 2019).

For instance, Ruta et al., 2019, examined the relationship between artistic expertise and preference for curvature. In their study, 24 female participants completed three different tasks that included —pairwise preference comparisons, rankings of psychological variables, and preference rankings— while comparing four different types of façades (curved, mixed, angular and rectilinear) projected on a wall. The findings of this study highlight a general preference towards curved and mixed properties of façade design.

In another study, rather than using self-reports Shemesh et al., 2017 used EEG measurements of 42 participants while in virtual environments viewed through a VR headset to study responses to curvature, irregularity and rectilinearity. Among the 42 participants, EEG results showed consistent differences between experienced and inexperienced designers. While inexperienced participants preferred curved spaces, experienced designers tended to favor rectilinear spaces. Complementary rankings supported these observations. Similarly, Banaei et al., 2017b compared two methods for measuring arousal responses in interior scenes, using the SAM (Bradley and Lang, 1994) and EEG. SAM results showed positive correlations between curvature and arousal, and negative correlations between arousal and rectilinear or angular elements.

Further research by Vartanian et al., 2013 used fMRI to examine the effects of curvature and rectilinearity in interior spaces with variations in ceiling height and openness of space. In this study, 18 participants evaluated preferred scenes and later selected

approach- or avoidance-based responses to these scenes during fMRI scans. The results indicated a clear preference for curved over linear designs

Ceiling height and spatial **scale** of an interior can convey feelings of freedom or confinement (Zhang et al., 2023). Zhang et al., 2023 studied how ceiling height can influences specific emotional categories within art gallery settings using 360-degree VR panoramic scenes. A total of 183 participants were included in the study and exposed to virtual art galleries with varying ceiling heights, and their emotional responses were collected using self-report measures. The research comprised two studies: the first measured absolute emotional ratings across different ceiling heights, while the second asked participants to choose ceiling heights based on specific emotions. The findings revealed that ceiling height significantly affected discrete emotions, with lower ceilings evoking greater fear and anger, and higher ceilings enhancing feelings of joy and increased arousal. The effects on emotions such as sadness, surprise, and disgust were more variable. This research underscores the value of using editable VR environments in studying the psychological impact of architectural design, offering critical insights into how spatial dimensions can shape emotional experiences in art gallery settings.

This tendency for larger volumes to trigger a higher arousal is also explained in Niedenthal, 2009 where the experience of awe is conveyed by the Gothic setting of churches and castles in the game *Resident Evil 4* (Capcom, 2005). The change of volume either from low-ceiling rooms to higher or the opposite is a tool that in many cases is employed by video game designers and architects in order to accentuate a change in spatial relationship, usually from a transitional "no-space" to a space; this process is identified as an *arrival* by Totten, 2014. Larger spaces have been associated with increased high-frequency oscillations in EEG bands, particularly in the *beta* and *gamma* bandwidths (Bower et al., 2022). However, smaller rooms were generally perceived as more pleasant, calmer, and safer, except when exposed to threatening sounds (Tajadura-Jiménez et al., 2010).

Interior configuration, or **spatial complexity**, is another key element explored by researchers. Vartanian et al., 2015, investigated how perceptions of enclosed versus open spaces influence beauty judgments and decisions about whether to approach or avoid a space. Using photographs of spaces that represented these two extremes, they collected both subjective evaluations and fMRI scans to observe brain activity while participants viewed the images. The results showed that enclosed spaces were strongly associated with avoidance decisions and activated brain regions involved in fear processing. This suggests that reduced visual and locomotive permeability in enclosed spaces may trigger negative emotional responses.

In another study Jang et al., 2018, examined how visual complexity in a fashion

store affects emotional responses. Using self-reports and physiological measures, the researchers created two store layouts in 3D: a low-complexity grid layout and a high-complexity free-form layout. Twenty-four participants evaluated these spaces on a 4K monitor while their EDA and facial EMG were recorded. The results indicated that high visual complexity negatively impacted pleasure in participants with low fashion involvement, although this effect was less pronounced in those with high fashion involvement. Regardless of involvement level, EDA measures consistently showed that higher visual complexity led to increased arousal.

Other studies emphasize the role of visual complexity in shaping preference judgments. Configurations that balance ease of processing with a certain degree of complexity tend to be more appealing. Additionally, research shows that "unusual-ness" and complexity can positively influence arousal, enhancing architectural engagement and memory (Gregorians et al., 2022).

Color appearance and materials is another important factor explored in studies that investigate their affective impact on occupants. Garip and Seymen, 2021 examined children's reactions to concrete materials in a school setting. Thirty-three children, aged 6–7, viewed classroom designs through VR. The study found that material and lighting variations did not significantly influence the children, as they tended to focus more on identifiable objects in the room rather than the background materials. Coşgun et al., 2021 conducted a study with 298 students to assess wall coverings—wood, metal, and concrete (in two tones)—in café environments across 11 semantic bipolar dimensions. Light-colored wall coverings were generally viewed more favorably than dark-colored ones, and wooden wall coverings were perceived as warmer compared to concrete and metal. The results also revealed that gender and design familiarity influenced preferences, with male participants showing a stronger preference for physical materials than female participants and design-oriented students. In another study,Gomez-Tone et al., 2021 recruited 22 students to rate their preference for materials during VR navigation. The responses highlighted differences based on students' study year, with results closely aligning with the responses of professional architects. Finally, Lipson-Smith et al., 2021 studied the impact of colored walls in virtual settings, including living rooms, waiting rooms, and empty rooms. Results from 100 participants, each experiencing a VR simulation of these rooms, showed significant correlations between two annotation models: a dimensional annotation for valence and a discrete annotation using the *Pick-A-Mood* scale (Desmet et al., 2016).

Research has shown that **lighting parameters** significantly impact task performance and cognitive function in various settings. Studies show that higher levels of illuminance and color temperatures tend to enhance performance in cognitive tasks, mem-

ory, and problem-solving (Knez, 1995; Konstantzos et al., 2020). For example, blue-enriched light and higher correlated color temperature (CCT) have been found to boost concentration, processing speed, and memory in students (Lekan-Kehinde and Asojo, 2021). Additionally, optimal contrast ratios and specific spectral tuning—such as red or blue wavelengths—can further improve task performance (Konstantzos et al., 2020). However, lighting effects may differ between genders, with mood playing a mediating role in cognitive performance (Knez, 1995). In work environments, observers modify their posture in response to lighting conditions to maintain task performance (Rea et al., 1985). Future studies are now focusing on additional lighting parameters such as vertical illuminance, directionality, daylight provision, and outside views to gain a more comprehensive understanding of how lighting affects task performance and cognition (Konstantzos et al., 2020).

The concept of "double dynamic lighting" is a recent area of exploration. Hansen et al., 2022a studied task lighting inspired by natural daylight inflows, incorporating varying combinations of direct and diffuse lighting to respond to changing sky conditions and daylight levels. In their 3-month study with four participants, lighting systems combined colder diffuse lighting from ceiling panels with warmer spotlighting at tilted angles, creating specific light zones at workspaces. While short-term effects of this setup were inconclusive, long-term results indicated improvements in visual comfort, atmosphere, and work engagement. Additionally, the study developed five dynamic light settings to respond to seasonal and daily changes (Hansen et al., 2022b). Lighting's effects on circadian rhythms and non-visual impacts have also been investigated in studies (Sen et al., 2017; Flyvholm et al., 2016). These studies align with guidelines on non-visual effects of lighting that support the natural day-night cycle (Rea et al., 2010).

Several recent studies choose to study these effects of lighting parameters also within VEs (Marples et al., 2020; Chamilothori, 2019) as with pre-rendered static images (Rockcastle and Andersen, 2014). Researchers that use VEs as test-beds, often focus on brightness, color, and light distribution (Joosten et al., 2012) to study affective responses, preferences, and attention-related tasks (Marples et al., 2020; Rockcastle et al., 2017). Lighting effects are also well-studied in video games, where they impact player performance and gameplay (Graja et al., 2020; El-Nasr et al., 2009). VEs and digital twins provide ideal platforms for studying these parameters, as they allow for real-time adjustments and observations of the resulting impacts. This feedback loop, where input parameters are adjusted and outcomes observed, greatly facilitates the design process (Kim et al., 2018a). For instance, in the *"Brighter Bronshoj"* case study, Kim et al., 2018b, designed and implemented an interactive lighting installation in a VR environment, demonstrating how game engines enable designers to test and refine their ideas in responsive and

dynamic settings. The following section reviews key studies that utilize immersive VEs to evoke and measure a range of affective responses, providing insights into how these environments contribute to our understanding of emotional reactions to spatial stimuli.

In this section we presented the prevalent spatial parameters that studies commonly investigate. Table 2.2 summarizes studies and trends in the field of environment and affect, highlighting spatial features, experiment protocols (Objective and Subjective affect capturing), stimulus type and session category regarding affect annotation (Post-stimulus annotation or Real-Time annotation).

2.3 | Affect and Virtual Environments

In recent years, VEs have become powerful tools for studying and influencing affective experiences. VR, video games, and video-sharing platforms provide immersive and interactive spaces that enable researchers to gather real-time data on emotional responses. These platforms allow for controlled manipulation of environmental elements, offering new insights into how people perceive and react to various spatial and visual stimuli. This chapter explores three key areas where affect is examined: virtual environments, video games, and video-sharing platforms, highlighting the distinct methodologies and findings that contribute to our understanding of affect in digital contexts.

2.3.1 | Affect Reporting within Virtual Environments

With the latest trends in game engines and tools that develop VEs, a plethora of environments is designed and developed with the aim of gathering affect while interacting with a VE or experiencing synthetic stimuli. This type of annotation is usually studied in terms of validity, being challenged mainly in terms of usability factors.Voigt-Antons et al., 2020, tasked 18 participants to compare two different annotation mechanisms on two different display media –HMD vs Desktop display– while viewing 360° video content. The first annotation mechanism was retrospective (participants reported affect after a session). The second method was real-time and continuous, encouraging participants to input their affect input on a 2D grid overlaid on top of the video content. The results show that no significant differences were found between the two annotation mechanisms, but significant differences were found regarding the sense of presence between the two displays, which as expected scored higher for the HMD mode.

Toet and van Erp, 2019, tasked 40 participants to annotate 360° videos displayed in VR on the arousal-valence dimensions using *Emojigrid*, a self-annotation reporting tool for immersive media. Results indicated high inter-rater agreement for the valence di-

mension and moderate agreement for the arousal dimension. In the study of Krüger et al., 2020, the development of the MAM tool sought to provide a pictorial representation of affect states to be embedded within VEs. MAM aimed to improve the accuracy of discrete models such as PAM (Desmet et al., 2016) and SAM (Bradley and Lang, 1994) by introducing interpolations between emotional states. As with *Emojigrid*, MAM displayed higher inter-rater agreements in the valence scale ratings compared to arousal. Lastly, Xue et al., 2021 designed and compared two methods for collecting continuous dimensional input within VR environments, aimed at reducing workload and distraction. For their two proposed methods, *HaloLight* and *Dotsize* (opacity and size of filled circle), they considered peripheral visualization techniques to avoid superimposition of the actual stimuli. Results showed both techniques being consistent with discrete labels.

The VREVAL framework and tool demonstrated in Schneider et al., 2018, aimed at the simultaneous collection of way-finding, spatial experience and qualitative feedback data, during an architectural experience within a VE. Authors evaluated the tool with the participation of 20 design students, and highlighted the tool's ability to assist in identifying problematic scenarios in early design phases. Additionally, the participating students reported that the tool could positively impact user-centered approaches and user-centric methodologies, especially in the areas of co-design and place-making projects. Other works that aim at collecting user reports during an architectural experience, emphasize on studying the impact of spatial features of generated geometries within VEs, In Gómez-Puerto et al., 2018, VR display was employed as means to navigate and interact with a set of six synthetic rooms, varying in 5 design characteristics from scale, to texture, to architectural style. The study aimed at comparing *digital twins* to their *real world* equivalents in terms of perceived sensations. While experiencing the virtual world and each of the different rooms, participants verbally reported affect from a set of predefined labels, for each of the 5 design categories, with results showing great similarities in induced affect states between real world and digital twins. Marín-Morales et al., 2018 designed four variations of virtual rooms meant to elicit four different affect states with different valence-arousal combinations according to the circumplex model of affect (Russell, 1980). The experiments 15 participants rating the four rooms confirmed the intended ratio of arousal-valence. The test occurred in VR, presenting rooms with the adjacent dimensions of the SAM (Bradley and Lang, 1994) embedded within the participant's FOV.

2.3.2 | Affect in Video Games

The emotional effect of spaces has also been studied in the domain of digital games and level design. Game developers and researchers often gather and analyze player feedback to understand how various game elements contribute to the overall experience. According to Liapis et al., 2018, these elements can be categorized into six facets: level design, visuals, audio, gameplay, rules, and narrative. The emotional resonance of spaces in video games is studied mainly through the facet of level design (Totten, 2019), but ambience is also impacted by the game's sounds, visuals and gameplay (Liapis et al., 2014).

In "The Underwood Project", McCall et al., 2022, explored how spatial uncertainty could be used to evoke emotional responses in a virtual environment. They employed level design techniques to create a predefined "tension arc" by manipulating factors such as lighting, scale, and the use of specific 3D models, audio, and interactive elements. The study measured participants' affect during gameplay through physiological data (heart rate) and post-game ratings of tension, asking participants to rate whether certain rooms felt "frightening," "creepy," or "unpredictable." The results showed that tension was primarily influenced by elements like hiding places, nearby hostile agents, blocked paths, and darkness. Similarly, Niedenthal, 2009, emphasized the role of lighting, obscured vision, and large-scale environments in creating tension and uncertainty in horror games.

To further study the effects of VR and in-game interactions on anxiety, Ferreira et al., 2023, developed a custom virtual environment following horror game design principles; they designed multiple scenarios for different tension levels by manipulating environmental factors such as lighting brightness, contrast, and dynamics, as well as environmental audio, triggered events, and hostile agents. Players' tension levels were captured during gameplay via three different types of biomarkers (electrocardiogram signals, EDA activity and respiration recordings), followed by a post-game questionnaire. Results showed that the scenario with the highest player tension had reduced lighting levels, flickering lights, and introduced the player to hostile agents. This study highlighted that environment parameters play an important role in manipulating player tension, but these parameters have to be viewed within the context of the game.

Similarly, Graja et al., 2020, tasked participants to play the popular horror game *P.T.* (Konami, 2014), capturing in-game recordings of EDA response and post-game self-reports of tension. Game features that were investigated were changes in lighting, player actions and sound. Their results suggested that, despite few cases where in-game biomarkers matched self-reports, there were promising tendencies between events re-

lated to lighting and sound changes. The authors highlighted that the order in which effects (e.g. sound changes) were arranged had a strong impact on emotional responses. As Boonen and Mieritz, 2018 pointed out in their player agency model, darkness can evoke a sense of uncertainty and tension, but in the case of *Amnesia the Dark Descent* (Frictional Games, 2010), darkness is also a tool to be used by the player in order to avoid hostile agents.

2.3.3 | Affect in the Wild

In addition to using virtual worlds and video games to study the relationship between space and affect, new methods have been developed to gather affective data in non-controlled environments, known as "affect in the wild" approaches. Several datasets have emerged from these efforts, including AffectNET (Mollahosseini et al., 2017), AFFWild (Kollias and Zafeiriou, 2018) and SFEW (Dhall et al., 2011). These approaches benefit from the vast amount of data available through the internet in general, social media platforms or VSPs. VSPs have become powerful tools for informing design and understanding the impact of media content, with video games being one of the most popular types of media streamed on these platforms. As a result, researchers and developers are exploring how to extract valuable information from these platforms to enhance game design.

YouTube[1] and *Twitch*[2] provide real-time feedback through audience participation features, such as live chat, comments, and reactions. For instance, when a streamer plays a game live, their audience can interact in real time, providing developers with immediate insights into how specific game mechanics, level designs, or story elements resonate with players. By monitoring these interactions, developers can identify the strengths and weaknesses of their games and make necessary adjustments. In one study, Melhart et al., 2020a, proposed a method for predicting engaging moments during gameplay by analyzing *Twitch* chat interactions. Focusing on the battle royale game *PUBG* (KRAFTON, 2017), they collected chat logs and in-game telemetry data from five popular streamers. Their prediction model achieved an accuracy of 84%, which was generalizable across different streamers and play styles. Similarly, Song et al., 2021, developed a model for identifying enjoyable moments in streamed content by analyzing viewer responses in the chat panel. They found that popular moments were characterized by high utterance frequency, short messages, and incoherent chat content, with features such as emotional expressions and audience sentiment playing key roles.

[1] `https://youtube.com/`
[2] `https://www.twitch.tv/`

Striner et al., 2021 explored emerging themes such as pacing, community, and player agency in games designed with audience participation in mind. Their work demonstrated the potential of VSPs as platforms for designing interactive experiences and how these platforms could be used to inform game design by analyzing emerging concepts from player feedback.

By observing how players react and interact with game environments during live streams, developers gain valuable insights into player engagement. They can see how streamers navigate levels, interact with characters, face challenges, and react to in-game events. This feedback helps developers understand the player experience, identify frustrations, and improve overall gameplay. Roohi et al., 2019 trained an Artificial neural network (ANN) to assign emotional labels to gameplay videos using a multi-modal expression streams,such as facial expressions, transcript sentiment, vocal emotion, and low-level audio features, gathered from the *Youtube* platform. Their model achieved over 70% accuracy in assigning custom emotional labels and 80% accuracy in detecting the five most intense moments of each video.

Live streaming sessions also serve as a form of "crowd-sourced" testing. When thousands of viewers watch a streamer play, they can quickly identify bugs, glitches, or design flaws, enabling developers to address issues that may have been missed during internal testing. Guglielmi et al., 2022 introduced *GELID*, an approach that detects anomalies and bugs during gameplay from videos extracted from VSPs. Their methods automates the detection of problematic segments, determining the type of anomaly and assigning it to a particular game context, thus assigning this to the appropriate development team for resolution. Similarly, Lin et al., 2019, proposed a ranking approach to identify gameplay videos likely to contain bug reports, helping developers collect context-rich bug information and improve game performance.

In this section we look deeper into the advancements of VEs and video games and how these are exploited by researchers seeking to capture continues affect both in forced (affect annotations) and manifested expressions, either in lab-setting or in the wild. We outline the drawbacks and benefits of each methodology and provide a foundation in developing environments with the aim of eliciting and capturing affect.

2.4 | Summary

This chapter outlined the theoretical background of **emotion research** and **architectural design**, highlighting relevant studies and applications that examine the affective impact of space. The studies included depicted a spectrum of architectural experiences for af-

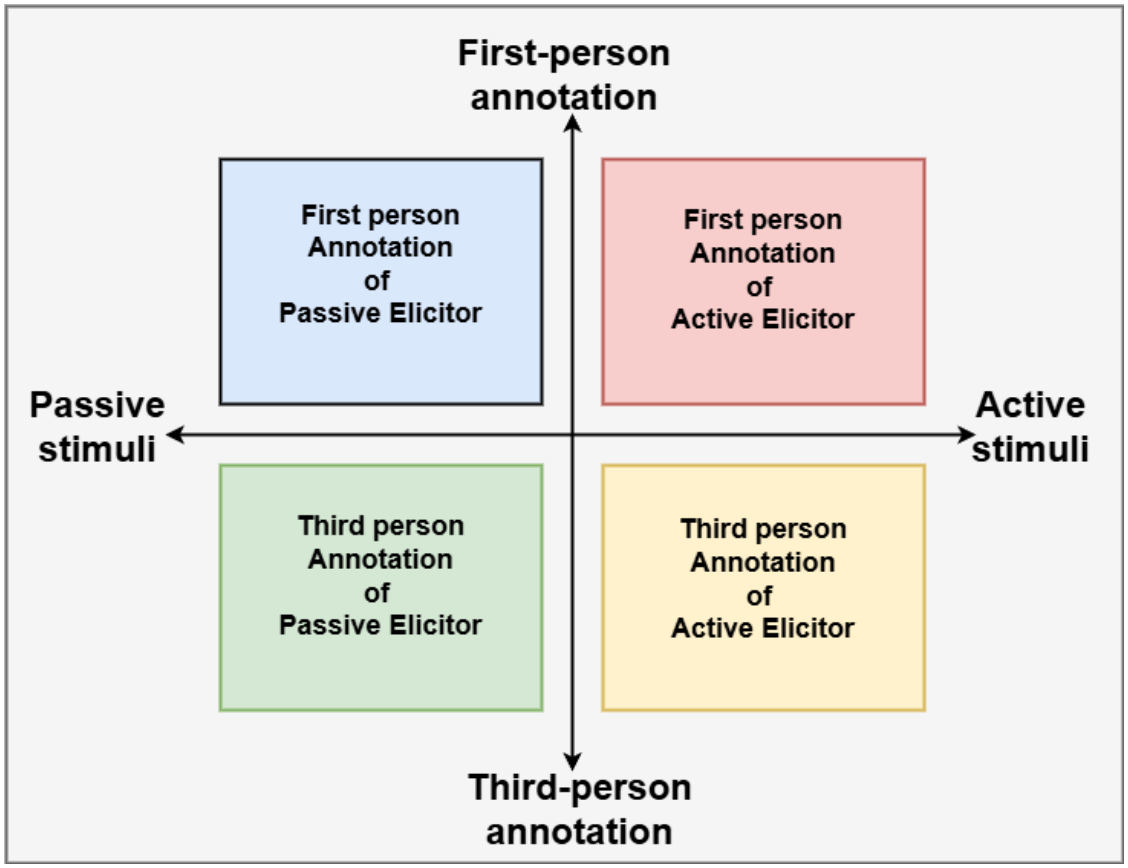

Figure 2.2: Framing Stimuli type to Affect Capturing methods. The three methods studied in this dissertation marked with blue, red and yellow and the out-of-scope method of "Third-person of Passive Elicitor" marked with green.

fect gathering, using both static (e.g., photographs, static renders) and dynamic elicitors (e.g., interactive virtual environments and video games). Additionally, affect gathering methods vary, ranging from first-person reports (where individuals describe their own affective states) to third-person reports (where another party assesses the affective state of the perceiver). We view the relationship between **stimulus type** and **affect-capturing methods** as a crucial aspect of studying the emotional impact of space, and we identify four distinct areas where these two dimensions intersect (see Figure 2.2). This dissertation focuses on three of these areas through four user studies. The fourth area, highlighted in green curvy lines in Figure 2.2 involves capturing affect through third-person annotation of a passive elicitor, which is not explored within this work. This approach seems more appropriate in the study are of films and affect (Almeida et al., 2021), due to the nature of the stimulus.

The next chapter details the processing framework used in this dissertation, focusing on the treatment of continuous affect data, how affect labels are derived in relation to spatial parameters, and the methods used to train a model predicting affective states based on these spatial parameters.

Table 2.2: Overview of Environment and Affect studies highlighting stimulus, spatial parameters and session type.

Study	Elements	Objective	Subjective	Stimulus	Type
Franz et al., 2005	Windows,Inter., Height, Area	-	Likert survey	Projections	Post
Meyers-Levy and Zhu, 2007	Height	-	Survey, Tasks	Real room	Real-Time
Vartanian et al., 2015	Height, Enclosure	FMRI	Approach-Avoid,Preference	Photos	Real-Time
Shemesh et al., 2017	Curvature, Symmetry	EEG	Preference,Free resp.	VR	Post
Banaei et al., 2017b	Form, Interior	EEG	SAM, PAD	VR	Real-Time
Schneider et al., 2018	Openess, Tasks, Brightness	Navigation	Survey,Discrete labels,Free resp.	VR	Real-Time
Ergan et al., 2018	Windows,Inter., Lighting	-	Preference,Free resp.	Photos	Real-Time
Marín-Morales et al., 2018	Lighting,Text., Geometry,Inter.	EEG, ECG	SAM	VR	Real-Time
Chinazzo et al., 2021	Daylight, Lighting	-	Semantic opposites,Likert	VR	Real-Time
Coburn et al., 2019	Natural patterns, Int., Ext.		Likert survey	Photos	Real-Time
Vartanian et al., 2019	Interior, Curvature	-	Likert,Approach-Avoid	Photos	Real-Time
Banaei et al., 2020	Form, Interior	-	SAM	VR	Real-Time
Lipson-Smith et al., 2021	Wall Color	-	Valence, Mood	VR	Post
Gomez-Tone et al., 2021	Scale, Materials, Enclosure	-	Free response, Discrete labels	VR	Real-Time
Garip and Seymen, 2021	Texture, Materials, Color	-	Free response, Preference	VR	Real-Time
Coşgun et al., 2022	Texture, Materials, Color	-	Bipolar opposites survey	Photos	Real-Time
Gregorians et al., 2022		-	VA ratings, Likert survey	Videos	Post
Formiga et al., 2022	Curvature	EDA, HRV	SAM	VE	Post
Presti et al., 2022	Scale, Windows, Color	-	VA ratings	VR	Post
Ruta et al., 2019	Curvature, Facades, Windows	-	Pairwise-preference	Projections	Post

3

Methodology

This chapter outlines the framework used to process the affective responses elicited by the environmental parameters presented in this dissertation. The primary focus of this dissertation is the capturing and analysis of *continuous data of affect* in the form of 1-dimensional signals, which reflects participants' emotional responses over time. Each study begins with a single continuous trace of affect data that goes through a series of processing steps to extract meaningful insights.

Figure 3.1 provides an overview of the general processing framework that was used across all four participant studies included in this dissertation. The framework covers every stage, from data processing to the final acquisition of affect labels. These resulting affect labels are either used as inputs for *modeling tasks* or as part of *agreement analysis*. By following this structured approach, the research aims to standardize the handling of affect data across multiple studies, facilitating a deeper understanding of the relationship between environmental stimuli and emotional experiences across various **dynamic stimuli** (videos, interactive virtual environments, and video games).

3.1 | Processing Continuous Affect Signals

All *forced-response* affect signals are captured using the *RankTrace* annotation protocol (Lopes et al., 2017b). For the *Affrooms24* and *Affrooms12* studies (See Chapter 5), we used the PAGAN annotation platform (Melhart et al., 2019) on videos of pre-recorded traversals of room sequences. *AffroomsMR* study (see Chapter 6) implemented the *RankTrace* protocol into the VE, allowing for affect recording to be captured while the rater is exploring synthetic spaces. Finally, the *OutlastAFF* study (see Chapter 7), focused on *manifested affect* rather than forced ratings. For this, affect signals were derived through affect recognition models using facial emotion recognition and voice emotion recogni-

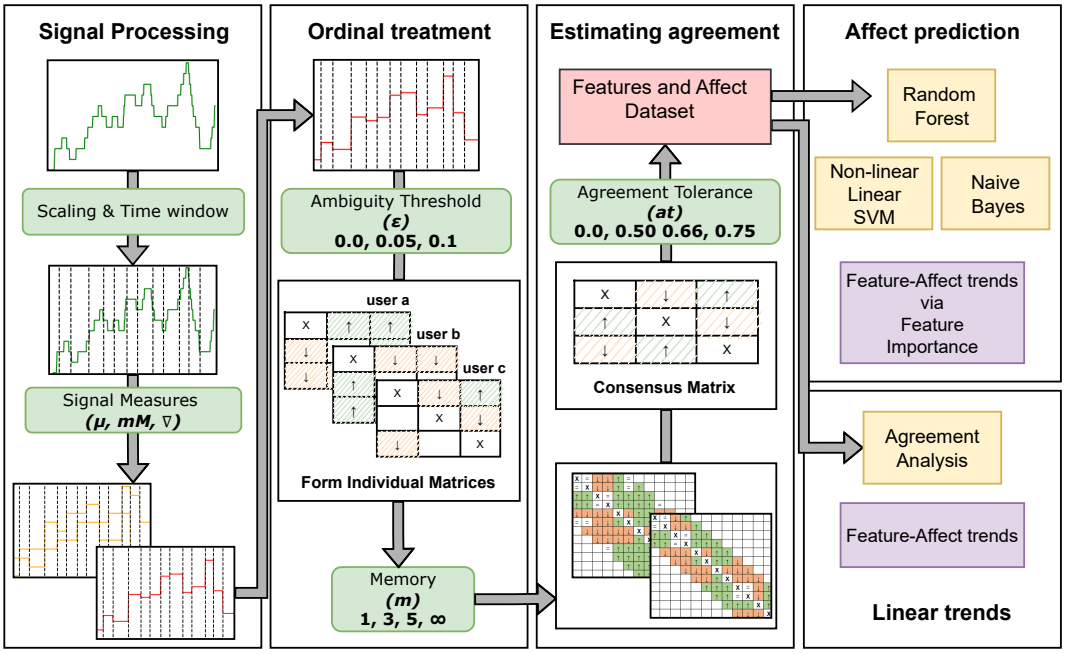

Figure 3.1: General Data Processing pipeline, from cleanup, scaling and room assignment to relative treatment, estimating inter rater agreements and finally modeling and spatial parameter impact

tion techniques. Further details on the data collection pipeline can be found in Chapter 4.

Scaling for all acquired affect signals occurs on a per-rater, per-room sequence basis using [0-1] min-max normalization. This ensures that **each signal is normalized within a consistent range, from 0 to 1, for each rater and room sequence,** regardless of the original range of values. This normalization process removes variability due to differences in individual rater scales or room contexts, allowing for fair comparisons across participants and conditions. It ensures that the relative differences in affective responses are preserved, while the absolute values are adjusted to a uniform scale, thus improving the robustness of subsequent analyses.

Throughout the dissertation the*sampling rate* of affect annotations was set to *250 ms* as a finer sampling rate proved to be unnecessary for the specified task of continuous rater annotations of interior synthetic spaces. As for the affect signals that derived from the emotion recognition models in Study 4, the sampling rate was set to *1 second*, aligning with the resolution used by the chosen models.

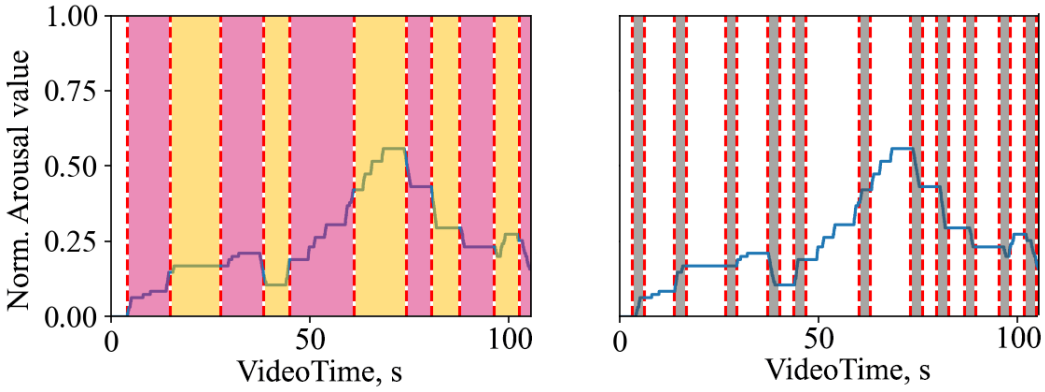

Figure 3.2: Two different approaches of signal splitting. Room windows *(left)* and Arrival windows *(right)*.

3.1.1 | Room Assignment

To understand the impact of each spatial parameter, this dissertation adopts a *room-based approach*, segmenting the continuous signal at designated timestamps based on when each room was entered. This is a common method in AC studies analyzing continuous signals (Parthasarathy et al., 2016; Camilleri et al., 2017a; Lopes et al., 2017b). This segmentation into 'room bins' allows for pairwise comparisons of the bins, treating them relatively. The *a room-based approach* results in windows of varying sizes, as each room is navigated in a different manner (see Figure 3.2 *left*). Additionally, a second approach—'arrival windows'—focuses on capturing affective responses during the experience of entering a space for the first time, as described by Totten, 2014, (see Figure 3.2 *right*). An *arrival* is defined here as the transition from one room to the next, with *fixed* time windows created around the moment the player enters a new room. Specifically, if t_e is the timestamp (in seconds) when the player enters the room, we define an arrival time window as $[t_e - 1, t_e + 2]$ seconds. Even though we considered both time windows for the pilot study, arrival windows were not further studied for the *Affrooms12*, *AffroomsMR* and *OutlastAFF* corpora.

3.1.2 | Measures of Affect

Common practices regarding the processing of the continuous signal suggest that absolute measures yield less reliable data compared to relative measures (Camilleri et al., 2017a). We choose therefore two *relative* and one *absolute* measure for our affect signals, which are computed using all data points for each room, and are compared throughout the dissertation and across the chosen stimuli. Figure 3.3 depicts an annotator's affect

trace for a single sequence and the corresponding visualization for the affect measures of room affect mean, amplitude and gradient. Their detailed description is as follows:

- **Affect mean (μ)**, is an absolute measure of affect and is calculated as the mean value of all points of the annotation within the room window (sampled at 4Hz). This measure gives us a general trend of the emotional state of the viewer within the room.
- **Affect amplitude (A)**, is a relative measure and is 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 and a small indicates no significant variation for the affect signal was registered.
- **Affect gradient (∇)**, is another relative measure of affect and it is calculated as the sum of differences between consecutive data points (sampled at 4Hz) within the room window. 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; This measure is an indication of the frequency of affect change within the time window.

3.2 | Ordinal Treatment with Preference Learning

In section 2.1.2 we highlight the superiority of ordinal treatments over absolute when processing data of affect, resulting in more reliable and valid affect labels. *PL* is a machine learning paradigm focused on understanding these ordinal relationships through preferences, either of individuals or agents over a set of items or alternatives. PL aims to model the underlying preferences implicit in data, and has been used extensively in personalized recommendation systems, decision support tools, and optimization algorithms tailored to individual preferences (De Gemmis et al., 2009).

PL algorithms typically take a relative assessment approach, where they learn from pairwise comparisons between items rather than absolute ratings or labels. This approach enables the algorithms to capture the relative preferences or rankings of items, allowing for more nuanced and fine-grained modeling of individual preferences. By focusing on a relative relationship between items, algorithms can effectively handle scenarios where absolute judgments are difficult or impractical to obtain, such as in sub-

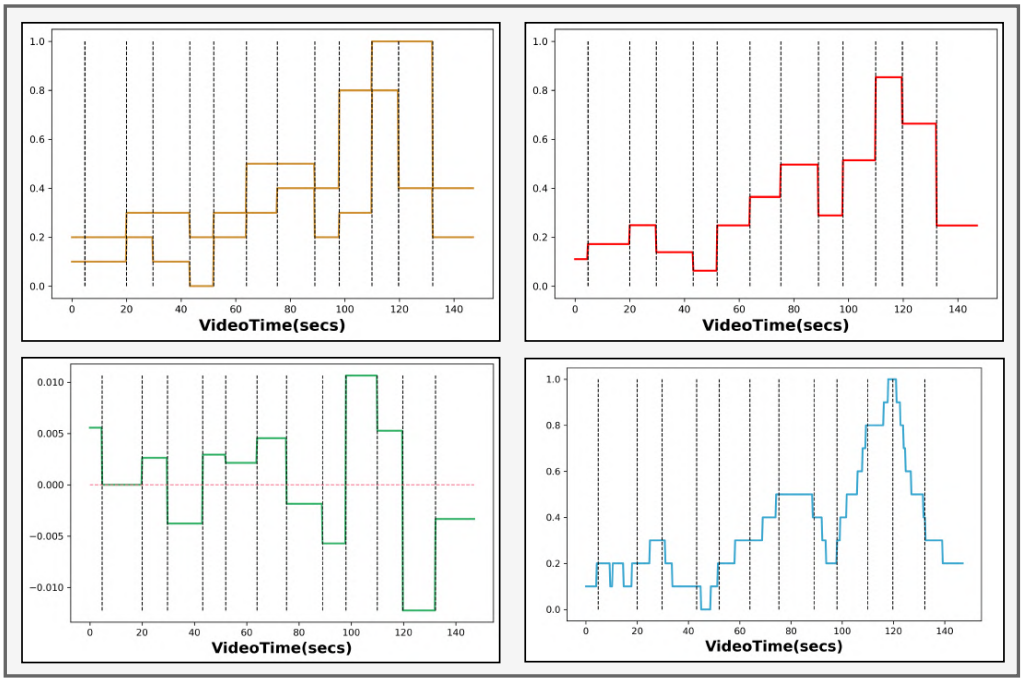

Figure 3.3: A single participant affect trace split across room bins *(bottom right)* and its corresponding measures of Amplitude (*top left*), Mean *(top right)* and Gradient *(bottom left)*

jective or context-dependent decision-making tasks, (Camilleri et al., 2017b; Yannakakis et al., 2017; Lopes et al., 2017b; Melhart et al., 2020b).

Pairwise preference learning is an approach aimed at predicting preferences between pairs of items. It focuses on learning the relative preference between two set of items, **A** and **B** and the objective is to predict which of the two is preferred based on their respective feature vectors, X_A and X_B. Rather than learning an absolute rating or score for each item, pairwise preference learning models **the relative relationship** between the pair. Typically, a model learns a function $f(X)$ that represents the "preference score" for each item based on its features. The prediction task then becomes the scores of two items: if $f(X_A) > f(X_B)$, item **A** is preferred over item **B**, and vice versa. Mathematically, the goal is to learn a function f such that:

$$f(X_A) > f(X_B) \implies X_A \text{ is preferred over } X_B$$

The model is typically a classifier (e.g., Support Vector Machine (SVM), logistic regression, RF, or neural networks) that takes the difference between the feature vectors of

two items as input, i.e., X_A-X_B. The classifier learns a decision boundary that separates pairs where A is preferred over B, and vice versa. Once trained, the model can predict the preference between any pair of items. It outputs a probability or a binary prediction (A > B or B > A) for each pair.

This approach is particularly useful in applications like ranking or recommendation systems, where the goal is to determine relative preferences rather than predict exact values. By learning from observed comparisons between pairs of items, the model can reflect the underlying preferences in the data. In the context of this dissertation, these pairwise comparisons between items occur on a room level (as described above), whereby within a single sequence, all rooms are compared with the remaining rooms of a rater's session.

3.2.1 | Relative Transformation for Affect Labels

Two integral parameters that are part of the processing framework are, the **Uncertainty (or Ambiguity) threshold** and the **Memory setting** and are generally used across studies to eliminate annotator bias in PL Datasets. These measures are used during our *relative transformation* step, where the affect measures for a single room are compared with the next rooms in the same sequence and ordinal relations are derived.

The **Uncertainty threshold**, that we refer throughout this dissertation also as the ϵ *value*, is an important step during the relative transformation and determines what we constitute as an affect change and thus needs to be further explored. Martinez et al., 2014, defined this measure as *minimum distance* and studies have worked with this parameter to reject noisy or ambiguous pair relationships. Depending on which affect measure is used (mean, gradient, etc.), various thresholds are explored (5% or 10%). If an absolute affect change does not exceed the set threshold, it is not considered and thus removed from the dataset. As such the following 3 cases apply to determine if the relationship between a pair is positive, negative or ambiguous:

$$M(room_A) - M(room_B) > \epsilon \tag{3.1}$$

$$M(room_B) - M(room_A) > \epsilon \tag{3.2}$$

$$|M(room_B) - M(room_A)| < \epsilon \tag{3.3}$$

Here, $M(X)$ represents the affect measure for Room X, where X can be any room (e.g., $room_A$ or $room_B$), and ϵ is the threshold that must be exceeded for the relationship to be considered non-ambiguous. With this process the relationship between room pairs

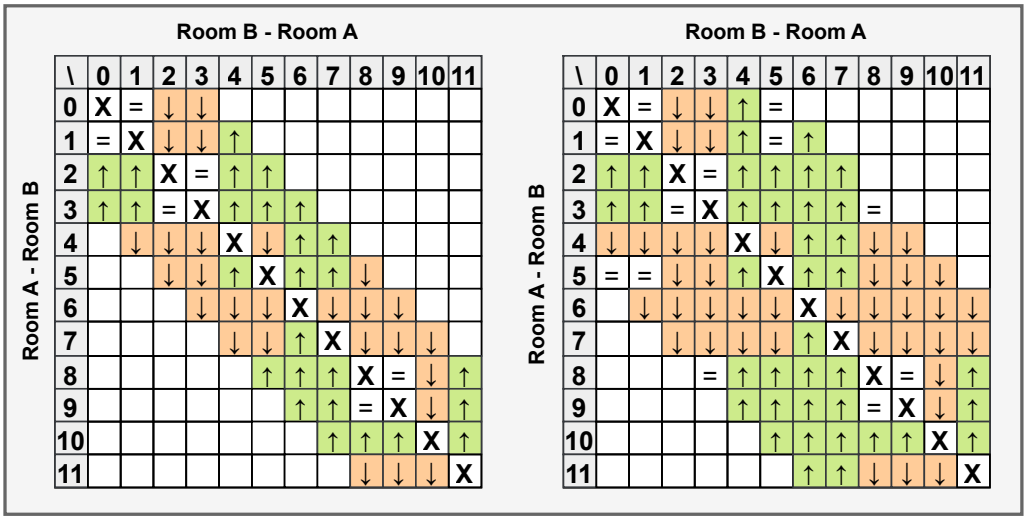

Figure 3.4: Examples of memory capacity m considered in IMs for a sequence of 12 rooms and memory settings of: $m = 3$ (left), $m = 5$ (right). Positive relations are marked with up-arrow, negative relations with down-arrow and ambiguous with equal sign.

derive, and labeled as negative (or downwards) in *Eq. 3.1*, positive (or upwards) in *Eq. 3.2* or ambiguous in *Eq. 3.3*.

Thus, the uncertainty threshold's ϵ goal is twofold: one to derive a relative trend among pair comparisons, and two to detect the minor effects as a measure for reducing uncertainty. All user studies included in this dissertation use a threshold of $\epsilon = 5\%$ for the normalized signals within the [0-1] range. In the Affrooms12 study (see Section 5.2), three different ϵ levels (0%, 5%, and 10%) are used to evaluate their impact on a Random Forest classifier's performance.

Following the terminology of the QA approach (Cowie and McKeown, 2010), these ordinal relationships between room pairs in the same room sequence are stored in an *Individual Matrix (IM)* (see Figure 3.4). Lastly, to create balanced datasets, we take into account not only the upper part of an IM as proposed in the QA approach, but also the lower part of the matrix marking the opposite trend relationship of the pair. For example, when the pairwise comparison relationship between rooms A and B is marked as positive then the comparison trend between B and A is marked as negative (see Fig. 3.4). Ambiguous labels are rejected, thus the predictive task becomes a binary classification task. As a result of the above, this processes ensures a 50% random guess baseline for all our predictive models.

The **Memory setting** is the next parameter used to control bias in rater's annotations

during a session. When populating an IM, all relationships between each room and the remaining rooms are retained for a single user. Although the IM captures all relevant information for one annotation session and affect measure, it is not always necessary to compare affect across all rooms in the sequence (higher memory setting). Raters may exhibit short-term memory patterns in their annotations, which may not remain consistent throughout the entire trace. Therefore, it can be more relevant to compare pairs of rooms that are closest to each other within the session.

To cater for this, the memory setting can be adjusted to compare neighboring rooms rather than rooms that are far apart within a session (referred to as the "single memory" setting). The *Affrooms24*, *AffroomsMR* and *OutlastAFF* studies use solely consecutive room comparisons, meaning each room is compared with the next room in the sequence. In contrast, the crowdsourcing study (*Affrooms12*), uses four different memory parameter settings: consecutive rooms, memory 3, memory 5 and memory ∞ (all rooms against all rooms), and explore how non-expert affect ratings could be affect by temporal biases.

As the memory setting increases, the dataset size grows significantly, potentially resulting in more ambiguous affect labels (see Figure 3.5, right). The choice between single or a higher memory setting can generally depend upon the nature of the stimuli or the expertise of the rater. For more demanding stimuli, such as video games, a single memory setting is commonly used (as seen in the works of Melhart et al., 2021a, and Lopes et al., 2017b). Conversely, for less intense stimuli, such as plain audio, or listening to discussions during human interactions, a higher memory setting could be appropriate (as observed in the studies by Sethu et al., 2019 and Cowie and McKeown, 2010).

3.3 | Inter-rater Agreement and Measures of Reliability

The final processing step before deriving the target affect labels is the inter-rater agreement analysis. To estimate how much raters agree with each other on the same stimuli (i.e., the same rooms and room sequences), we analyze their IMs and combine them into a single Consensus Matrix (CM), one for each room sequence. The CM is constructed similarly to the **QA framework** (Parthasarathy et al., 2016), and requires a minimum of three raters to have provided annotations for the same room sequence. Room sequences that did not contain at least three valid rating sessions were discarded.

The CM includes only the samples that meet a predefined **agreement tolerance (a_t)** threshold, meaning that raters are considered in agreement if their ordinal labels within an IM bin are not conflicting (e.g., one rater reporting upward trends while another

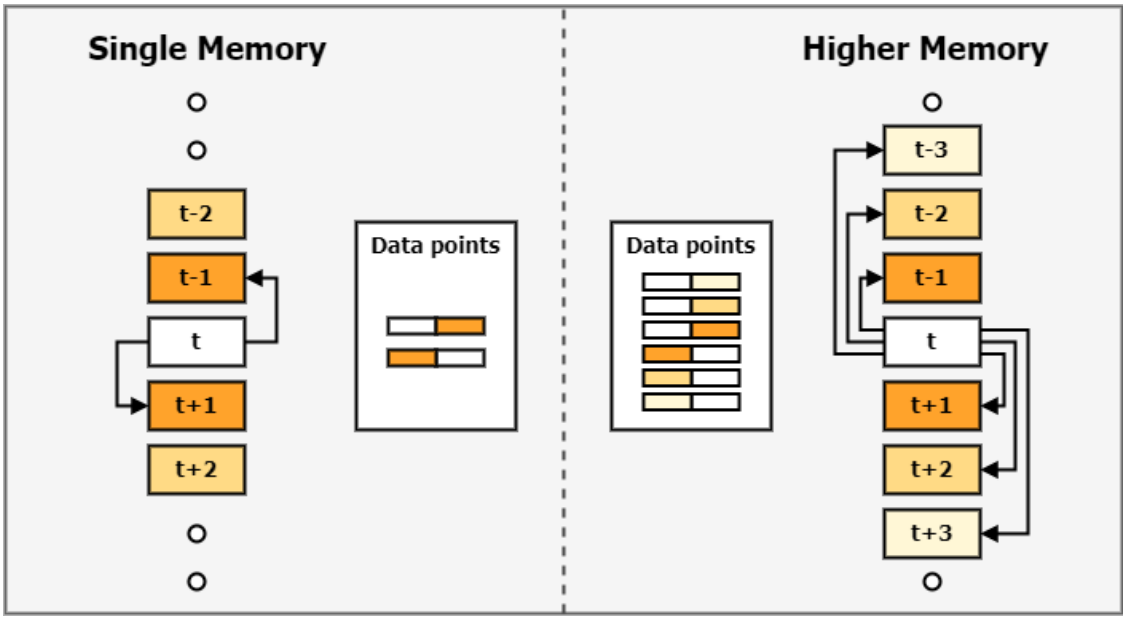

Figure 3.5: Single Memory setting *(left)* and Higher Memory setting comparison *(right)*. The former considers solely consecutive instances in a sequence of rooms, the latter compares instances multiple time-frames apart resulting in more data points. In the example here a single room is compared with 3 rooms apart in the same sequence.

reports downward trends). A label is assigned in the CM if it exceeds the agreement ratio set by the (a_t) threshold. This dissertation explores different (a_t) values above 50% (simple majority), 66%, 75%, 100% (all raters agree) and a baseline of (a_t) = 0%, where all user data is considered valid. It is important to note that filtering based on inter-rater agreement is applied only for the video sessions (Chapter 5), where we are certain that all raters are exposed to the same stimuli, making inter-rater agreement meaningful. However, in studies involving interactive environments (Chapter 6 and Chapter 7), where we have no control over what the rater perceives, this method is less relevant.

Reliability is crucial when evaluating the consistency of annotations made by multiple raters. It measures the degree to which different annotators provide consistent labels for the same data, which is essential for ensuring that the labels are reliable and not the result of random disagreement. Various statistical methods are used to quantify this reliability, each with unique properties and applications. The most common measures of inter-rater agreement used throughout AC studies are the following.

Cronbach's Alpha (Cronbach, 1951) is a measure of internal consistency, often used to assess how closely related a set of items are as a group. In the context of inter-rater

reliability, it is used to evaluate how consistently different raters assess the same set of items. A higher value (closer to 1) indicates better agreement and consistency among the raters.

SDA (Booth and Narayanan, 2020) is a measure that quantifies the extent of agreement between raters by focusing on the signed difference between ratings. This method is particularly useful when comparing continuous or ordinal data, as it takes into account both the magnitude and direction of the differences in ratings, providing a more nuanced measure of agreement. The SDA measure is adopted in Chapter 6 as a method to quantify intra-rater agreement between continuous annotations for two different media displays (VR and desektop).

$$\text{SDA} = \frac{1}{N-1} \sum_{t=2}^{N} \delta \left[\text{sgn}(x_t - x_{t-1}), \text{sgn}(y_t - y_{t-1}) \right] \tag{3.4}$$

Krippendorff's Alpha (Krippendorff, 2018) is a versatile reliability coefficient that works for various types of data, including nominal, ordinal, interval, and ratio. It accounts for the possibility of agreement occurring by chance and is robust even in the presence of missing data. Krippendorff's Alpha is commonly used in content analysis and works well for both small and large sample sizes.

Cohen's Kappa (Cohen, 1960) measures the agreement between two raters while accounting for the possibility of chance agreement. It is commonly used for categorical data and provides a more conservative estimate of agreement compared to simple percent agreement. Values of Cohen's Kappa range from -1 to 1, where 1 indicates perfect agreement, 0 suggests no agreement beyond chance, and negative values imply systematic disagreement.

The **Qualitative agreement** (QA) proposed by Cowie and McKeown, 2010, is a framework designed to create more reliable labels from continuous affect signals. The method involves the following steps:

1. **Generate Relative Labels:** The signal is split into fixed bins, and pairwise comparisons are made between predefined windows of the signal. This is done in a per-rater, per-signal manner, using the methods described earlier in Section 3.2.1.

2. **Estimate Inter-Rater Agreement:** The next step is to measure the level of agreement between different raters. This allows the identification of labels where raters are most in agreement, while discarding labels that do not meet the agreement criteria. The agreement level is controlled by an *agreement tolerance* parameter, which can be set at different thresholds: 50% (at least half the raters agree), 75%, or 100% (all raters agree).

Study	Measures	Time Windows	Memory	(ϵ %)	(a_t %)
Affrooms24	(μ, A, ∇)	Arrival & Room	1	5	66, 100
Affrooms12	(μ, A, ∇)	Room	1, 3, 5, ∞	0, 5, 10	0, 50, 66, 75
AffroomsMR	(μ, A, ∇)	Room	1, 3, 5, ∞	0, 5, 10	0
OutlastAFF	(μ A, ∇)	Room	1	5	0

Table 3.1: Processing framework parameter settings and methods for estimating spatial feature impact across the four studies.

3. **Train Rank-Based Classifiers:** In the final step, a ranker is trained using the relative labels and their corresponding feature vector differences. This uses the labels that meet the agreement criteria to build a classifier.

Table 3.1 provides an overview of the framework's parameter settings that we used for each user study. Affect Measures and Time Windows as described in Section 3.1, Memory and Uncertainty threshold ϵ in Section 3.2 and finally Agreement tolerance a_t as described in this section. The next section outlines the machine learning methods that were considered for building classifiers in the context of *Preference Learning*.

3.4 | Learning to Rank

As pointed out in Section 3.2, we take an ordinal approach to affect treatment using *Preference Learning*, whereby we seek not to learn from relative assessments and ranks rather than absolute ratings, commonly happening in traditional classification tasks. This task is commonly referred as *Label ranking* (Fürnkranz et al., 2008)is a common approach in AC, with studies utilizing a plethora of algorithms tasked to learning the underlying order of the affect labels.

One of the most common techniques is the use of Ensemble methods such as **Random Forests**. Random Forests are a popular supervised learning algorithm utilizing an ensemble learning technique renowned for their versatility and robustness across a spectrum of machine learning tasks. Many studies have displayed their robustness when compared with Support Vector Machines and Deep learning techniques. Rooted in decision tree algorithms, Random Forests harness the collective wisdom of multiple individual trees to deliver accurate predictions and effectively handle complex datasets. By aggregating the predictions of numerous decision trees trained on random subsets of the data, Random Forests mitigate over-fitting tendencies and enhance generalization performance. Random forests include a range of hyper-parameters configuring the learning task and their implementation within this thesis is achieved via the scikit-learn

Python package Pedregosa et al., 2011, and the hyperparmeters 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.

Additionally, Random Forests are a viable option throughout studies for their interpretability by producing Feature importance vectors commonly by **mean decrease impurity**. *Mean Decrease Impurity (MDI)* in random forests is a method that measures the importance of features by calculating the average decrease in impurity (e.g., Gini impurity) across all decision trees when a particular feature is used for splitting nodes. As a result each feature receive a MDI value which is normalized between 0 and 1, with higher values indicating the contribution for each feature on this classification task, ultimately producing an ordinal relationship of all features from the most important to the least important features.

Ensemble approaches like Random Forests and Decision trees have been used by several studies that seek to rank labels (de Sá et al., 2017), displaying robust results. de Sá et al., 2017 proposed Label Ranking Forests, an ensemble method combining different decision trees for label ranking. These approaches leverage the strengths of random forests, such as scalability and parallelizability (Zhou and Qiu, 2018), while addressing the challenges of preference learning and label ranking. The studies collectively demonstrate the effectiveness of random forest-based methods in handling complex preference and ranking tasks.

SVMs are another powerful supervised learning technique used for binary and multi-class classification tasks. SVMs aim to find the hyperplane that best separates the data points of different classes in the feature space. In a binary classification task, SVMs seek to find the hyperplane that maximizes the margin between the closest data points of the two classes, known as support vectors. A Rank Support Vector Machine (RankSVM), is a variant of the traditional SVM algorithm that is specifically designed for learning to rank tasks (Joachims, 2002). While traditional SVMs are primarily used for binary classification tasks, RankSVM is tailored for scenarios where the goal is to learn a ranking function that orders a set of items based on their relevance or preference. Parthasarathy et al., 2016 trained a RankSVM model using data from the SEMAINE database (McKeown et al., 2011), which contains audiovisual dyadic interactions. Their goal was to predict affect rankings across four dimensions: arousal, valence, power, and expectation using their extensive Qualitative assessment (QA) approach.

Naive bayes is a simple yet effective probabilistic classifier commonly used for binary classification tasks. It is based on Bayes' theorem and makes the "naive" assumption that features are conditionally independent given the class label. Despite this simplifying assumption, Naive Bayes often performs well in practice, especially in text clas-

sification tasks. In binary classification, it calculates the posterior probability of each class based on the input features and assigns the class with the highest probability to the input instance. Aiguzhinov et al., 2010 proposed an adaptation of the Naive Bayes classifier for a label ranking task. Unlike previous ranking approaches, which rely on multiple pairwise comparisons to predict labels, this method aimed to predict the order of labels in the test set. The results demonstrated that the Naive Bayes ranker could achieve similar, and in some cases superior, performance compared to traditional SVM rankers.

All three classification techniques are implemented on the *Affrooms12* dataset and details regarding their configuration, results and are documented in Chapter 5. A Random Forest preference learner is used also on the *Outlast Asylum affect dataset* and and further information about its implementation and prediction results can be read in Chapter 7.

3.4.1 | Evaluation

All ML models that are built during the course of this dissertation study are evaluated using the four most common performance measures of accuracy, f1-score, precision and recall. Accuracy measures the ratio of correctly predicted instances to the total number of instances. It provides a straightforward assessment but is not sufficient in cases imbalanced datasets. As our methodology relies in creating balanced datasets – as highlighted in Section 3.2.1 – we primarily report accuracy scores for all models and discuss our findings based on this performance metric. Precision indicates the proportion of true positive predictions among all positive predictions made by the model. A high precision score means the model makes fewer false positive errors. Recall (or Sensitivity) measures the proportion of actual positives that the model correctly identifies. It is useful in assessing the model's ability to capture all relevant instances. Finally, the *f1* score represents the harmonic mean of precision and recall, offering a balanced measure of performance, especially in cases with imbalanced datasets.

Another evaluation method other than the measures of performance is the *Train-Test* split method and the *Holdout* validation method. The train-test method is the most straightforward evaluation technique, where the dataset is divided into two parts: a training set and a testing set. The model is trained on the training set and evaluated on the test set. This method provides a quick estimate of how well the model can generalize on unseen data. However, it can lead to biased results if the split does not represent the overall dataset well (Pal and Patel, 2020). On the other hand, the holdout method involves splitting the dataset into three parts: a training set, a validation set (for tuning hyperparameters), and a test set (for final evaluation). This approach allows for more

comprehensive assessment but requires sufficient data to ensure each subset is representative (Pal and Patel, 2020).

We evaluate all models performance within this work with the accuracy measure, except *Affrooms12* dataset (see Section 5.2) where all four measures are employed, as we make a thorough comparison of the Random Forest classifier with additional classification methods. We choose accuracy as the main performance indicator as it is a straightforward approach to assess a model trained on balanced datasets and binary labels, looking only at the ratio of correct predictions to the overall number of test predictions.

3.5 | Summary

This chapter detailed the processing methodology used throughout the dissertation. Following the main framework illustrated in Figure 3.1, we outlined each step involved in processing the affective signals to extract ordinal affect labels. In the next chapter we present a detailed description of the data collection pipeline that we followed for each study.

4

Data Collection

Advancements in the study of space and affect, as explored in Chapter 2 display a shift from static content –photographs, projections and rendered frames– to videos (Gregorians et al., 2022; Xue et al., 2021) and more more dynamic content, such as interactive virtual environments (Schneider et al., 2018; Gómez-Puerto et al., 2018; Marín-Morales et al., 2018) and games (Ferreira et al., 2023; McCall et al., 2022). However studies that investigate dynamic content focus mainly in acquiring physiological (Banaei et al., 2017a; Banaei et al., 2017b; Banaei et al., 2020; Shemesh et al., 2017; Marín-Morales et al., 2018). These methods usually pose challenges due to their intrusive nature, high costs, or the requirement for participants to be present in controlled lab settings. Non-intrusive means for gathering affect such as participant reports, are considered by many but most of these fail to acknowledge the temporal dimension (Ergan et al., 2018; Franz et al., 2005; Coburn et al., 2019 ; Vartanian et al., 2015 ; Ruta et al., 2019). In most cases where user annotations are gathered, these mostly come in the form of post-experience evaluations. These retrospective evaluations tend to capture affective states in discrete segments that could underestimate the felt sensation due to memory gaps, as pointed in Section 2.1. Other studies that seek to capture the real-time impact of spaces embed affect reporting mechanisms within VEs like in Marín-Morales et al., 2018, where participants were instructed to report their affective states at discrete moments with a single annotation point. By doing so these approaches are oblivious of any emotional insights that could unfold in time.

In the past decade the field of AC has dramatically evolved and a large number of studies show great examples of methods and frameworks regarding the capturing of continuous user annotations across a range of different stimuli like video and 360 video (Fayn et al.; Xue et al., 2022; 2020a; Greer et al., 2020; Greer et al., 2019) using music or sounds (Ridley et al., 2024) as main stimulus or VEs (Yang and Kalantari, 2022).

Table 4.1: Summary of the properties from the four datasets, outcome from four different subject studies.

Properties / Dataset	**Affrooms24**	**Affrooms12**		**AffroomsMR**	**OutlastAFF**
Nr. of Rooms	24			16	39
Stimuli type	**Passive**			**Active**	
Nr. of Sequences	20	55		31	16
Nr. of rooms in Sequence	24	12		16	93
Duration per Sequence (mins)	3.2	2.1		4	≈ 33
Affect Annotations					
Annotation	First-person (Forced)				Third-person (Free)
Annotation method	RankTrace				Pre-trained Aff models
Annotation environment	PAGAN			VE Embedded	
Avg session duration (mins)	62.3	14.1		4	-
Database duration (hours)	≈ 3	17.2		≈ 4	≈ 10
Affect Label	Arousal	Arousal	Pleasure	Pleasure	Arousal
Nr. of room Sequences	60	224	215	62	16
Nr. of Subjects	3	39	37	31	16

Additionally, the late advancements in machine learning and artificial intelligence has enabled a plethora of works to study manifested expressions with the use of pre-trained emotion recognition models.

This dissertation aims at creating Spatial affect corpora to address the gaps in non-static stimuli. The review in Chapter 2 framed 4 study areas, conveying the interplay between stimulus and affect-capturing methods, as illustrated in Figure 2.2. Four different affect datasets were the result of four different studies sought to capture continuous affect states from users; see Table 4.1. Within this chapter each study's experiment conditions, procedure and data collection pipeline is documented.

4.1 | Affect Annotation Methods

Motivated by the lack of comprehensive studies to address the objectives above, four studies were conducted varying in affect capturing methods and on different types of stimuli. The first three studies follow a forced-response approach using a continuous affect reporting method while the fourth extracts manifested affect in the form of free-response. Two of the three forced-response studies implement a "First-person annotation of Passive elicitor" (Affrooms24 and Affrooms12 datasets) using pre-recorded videos as stimuli and Study 3 follows a "First-person annotation of Active Elicitor" method (AffroomsMR dataset), using an interactive VE as stimulus. The last study uses the "Third-person annotation of Active elicitor" and uses pre-trained affect recognition models to extract manifested affect data. The stimulus here is a *horror game* title and affect is captured during gameplay, from videos that were streamed online by play-

ers (Outlast Affect dataset). More specifically the pre-trained models capture affect via facial expressions, and process both the utterances (i.e. what is said) and audio information of the voice (i.e. how it is said) for affect recognition.

In Chapter 2 we defined Passive elicitor as a non-reactive stimuli that is presented to the rater for feedback report and Active elicitor as stimuli that is interactive. Both these approaches are used to capture continuous affect changes in a dimensional manner regarding environment changes. What is not covered as a 1st-person affect annotation method is the *Stimulated Recall* situated between a passive and active stimuli, aimed at reducing cognitive load during the annotation process. The reason for this is that the required task of spatial exploration features minimal interactions, thus it is considered as an appropriate experimental protocol for this Dissertation. The stimulated recall method is usually applicable in more dynamic scenarios, like in video games (Melhart et al., 2019; Lopes et al., 2017b) where first-person forced-response affect annotations were studied during gameplay.

Section 2.1 presented the most common methods for reporting affect across single and multiple dimensions. As pointed out, single dimension reporting comes with the advantages such as ease of use, faster processing from the rater's side and better explainability (Clerico et al., 2016;Lopes et al., 2017b). This dissertation uses single dimensions reporting throughout all self-reporting studies. The RankTrace affect reporting tool of Lopes et al., 2017b was chosen as the main method for affect annotation in this dissertation. RankTrace differs from other uni-dimensional methods as it is unbounded, allowing for each rater to report without the limitation of an absolute scale. This unbounded design is based on *anchoring effect* and *adaptation level theories* as discussed in Chapter 2.1. Lastly, RankTrace's continuous annotation method, encourages raters to annotate relatively as earlier ratings appear in the form of a developing trace.

4.2 | Stimuli & Space Parameters

Based on the review study documented in Section 2.2 four key features were selected regarding their contribution to the appearance of space, either in relation to the *room's geometry*, its *interior configuration* or its *ambient lighting setting* affecting its appearance. In terms of room geometry we study the effect of **room contour curvature** and **ceiling height** while for its **interior complexity** the presence of occluding elements of walls and columns. For each of these properties two extremes are considered either as being present or absent in each room. In the case of room size, low ceilings represent small size (absence) and high ceilings represent large size (presence). Figure 4.1 shows all four pos-

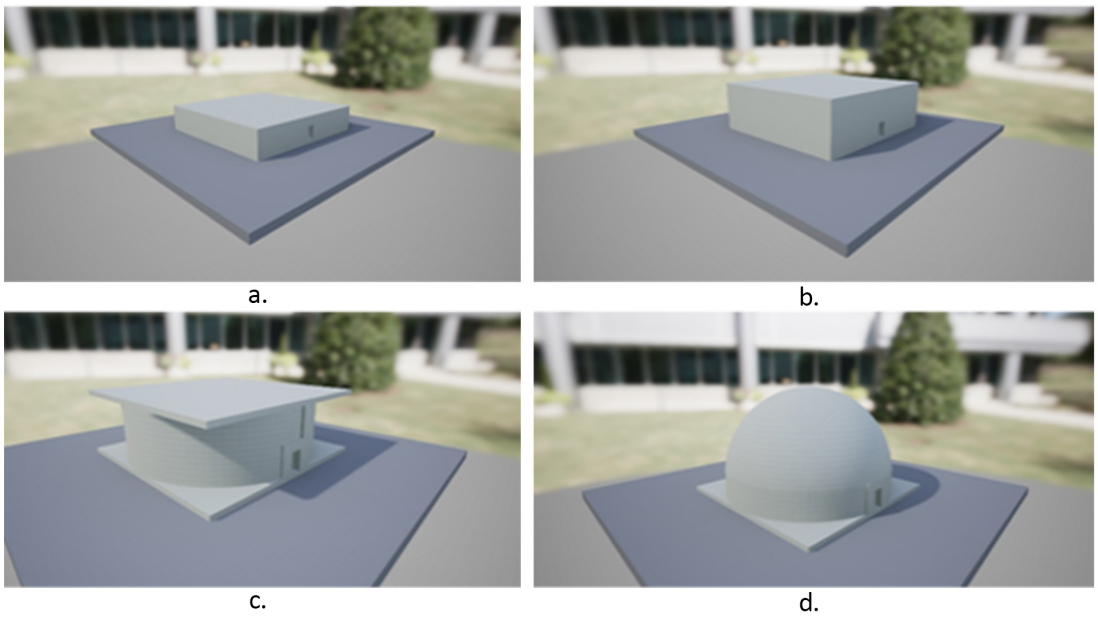

Figure 4.1: External view of rooms for combinations of size and curvature: a) rectilinear with small size, b) rectilinear with large size, c) curved with small size, d) dome with large size.

sible room geometries we use for our synthetic stimuli. Combinations of curvature and size affect the room's form with small size and curvature resulting in cylindrical rooms (see Figure 4.1 c) while large size and high curvature resulting in a dome-like structure (see Figure 4.1 d). Interior complexity introduces boundaries and obstructions during navigation, with complex spaces having symmetrically placed columns and two walls in the middle of the room that obstruct both visibility and the path of the user navigating through the room. As such the aforementioned room encoded features of curvature, complexity and size are represented in a binary format of absence (0) and presence (1); see Table 4.2 for considered room features and their encodings. Each feature is categorized if its contribution lies in the room's geometry, interior setting or appearance (ambient lighting setting).

Following the paradigm of previous studies on the effect of color in navigable spaces as in Niedenthal, 2009, we explore different ambient illumination color settings. *Affrooms24* and *Affrooms12* studies use a three-color setting of blue, white and red while for the *AffroomsMR* study we choose a two-color setting of warm (2500K) and cold (6500K) CCT settings, replicating the conventional light emitting diodes under extreme conditions. Light sources were distributed within each room so as to provide well lit general

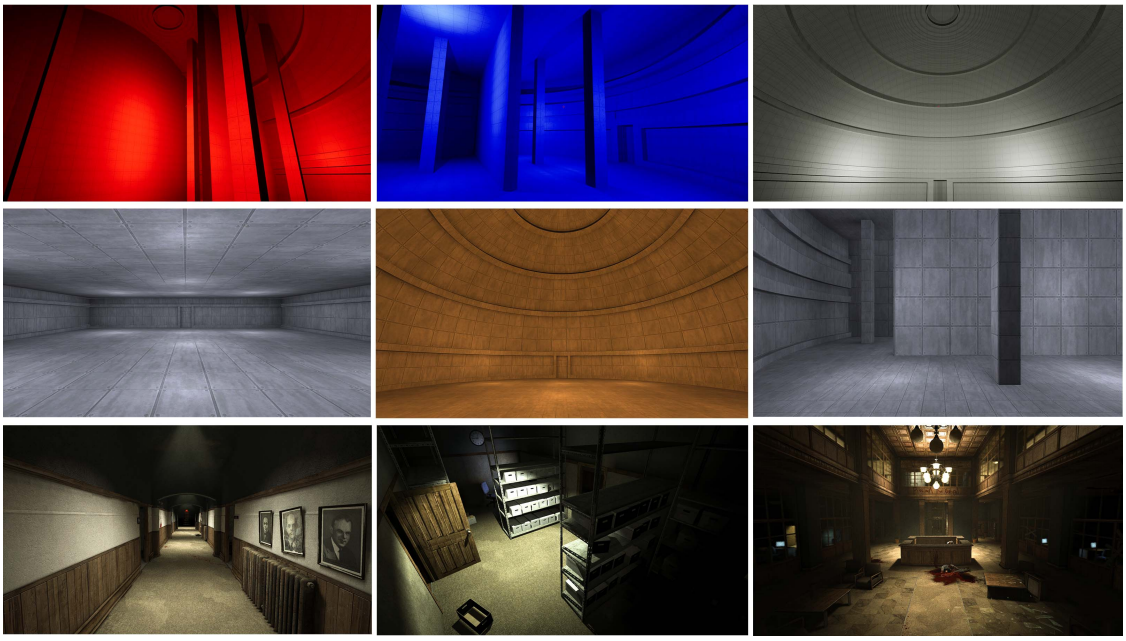

Figure 4.2: Views of rooms on the *Affrooms12* & *Affrooms24* datasets (top row), *AffroomsMR* dataset (middle row) and *Outlast Asylum Affect* dataset (bottom row)

illumination, eliminating strong contrast throughout each room. The corresponding encoded states for the ambient illumination with the three colored setting are 1 for red, 0 for white, -1 for blue, and for the two-color setting 1 for warm and 0 for cold, to track changes in illumination.

With the above configurations the resulting number of rooms for *Affrooms24* and *Affrooms12* are 24 and for the *AffroomsMR* 16. Each room has the same exterior dimensions, consisting of 20 meters width by 20 meters depth. Depending on the size and curvature parameter combinations, heights range from 3 to 12 meters. All resulting rooms as part of the dataset, from a first-person view, are shown in Fig. 4.2.

In the *OutlastAFF* study we explore games as a case of potential informants regarding the impact of space in manifested emotions. This study is based on the horror title *Outlast* developed by Red Barrels[1] and released for PC in 2013. *Asylum* is the first level of Outlast and includes 41 different rooms across three floors varying in several aspects of room's form, interior complexity and lighting appearance. Rooms are assigned eight *spatial* descriptors that for the most part, describe the room's geometry, its navigability (blocked paths), its contents (e.g. empty, neatly arranged furniture, or chaotically strewn debris and corpses), and the level and color of the illumination (see Table 4.2). More-

[1]https://redbarrelsgames.com/games/outlast/

Table 4.2: The 15 features describing spatial and game properties and their values. The numbers in parentheses denote the encoded values used to measure differences between adjacent rooms.

	Feature	Values	Type
Affrooms	Contour curvature	False (0), True (1)	Geometry
	Ceiling height	Small (0), Large (1)	Geometry
	Occluding elements	False (0), True (1)	Interior
	Light color	Blue (0), Neutral (1), Red (2)	Appearance
	Light temperature (Aff16)	Cold (0), Warm (1)	Appearance
Outlast Spatial	Area size	Small (0), Medium (1), Large (2)	Geometry
	Ceiling height	Low (0), Medium (1), High (2)	Geometry
	Light contrast	None (0), Uneven (1), Even (2)	Appearance
	Light levels	Dark (0), Dimly Lit (1), Bright (2)	Appearance
	Light (color) temperature	Warm (0), Cold (1)	Appearance
	Blocked path	False (0), True (1)	Interior
	Empty room	False (0), True (1)	Interior
	Interior arrangement	Chaotic (0), Ordered (1)	Interior
Outlast Game	Hiding place	False (0), True (1)	Gameplay
	Triggers present	False (0), True (1)	Gameplay
	Battery present	False (0), True (1)	Gameplay
	Note present	False (0), True (1)	Gameplay
	Cutscene	False (0), True (1)	Gameplay
	Event	False (0), True (1)	Gameplay

over, rooms may contain gameplay elements, or trigger new interactions. These *game* features are also labeled, as they affect the gameplay affordances of each room. Seven game features are labeled (see Table 4.2), including the presence of hiding places, batteries, triggers that allow the player to continue their level traversal (e.g. keys, levers), notes, and events (e.g. interactive jump scares or audio cues) or cut-scenes in the room. Even though the *OutlastAFF* study extends the investigated features and in some cases with additional settings, it does not investigate the contribution of room's curvature, a spatial parameter rarely considered by level designers especially in the subdomain of horror titles as curvature is usually linked to more calming and pleasurable experiences (Ruta et al., 2019; Shemesh et al., 2017).

4.3 | First-person Annotation of Passive Stimuli

Affrooms12 and *Affrooms24* studies feature affect reporting on videos recordings of traversals of synthetic pre-generated sequences of rooms, with the former employing expert

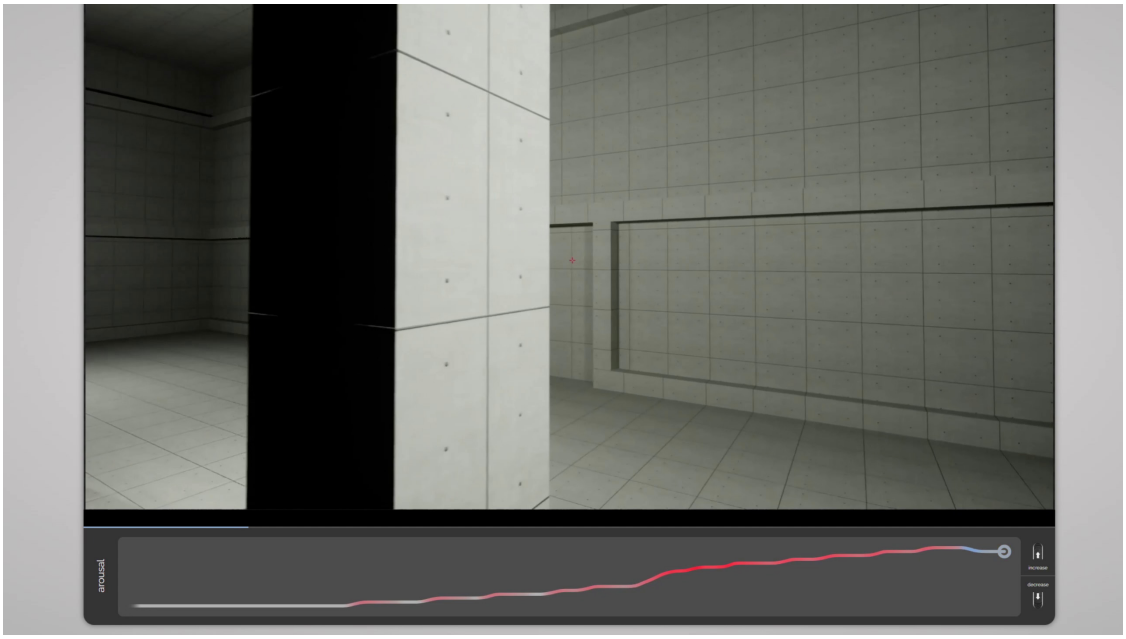

Figure 4.3: Screenshot of PAGAN during Arousal annotation: navigation video (top) and continuous arousal annotation (bottom)

annotators and the latter non-expert annotators. Both studies use the *RankTrace* annotation protocol (Lopes et al., 2017b) to capture annotators' ratings on affect changes. The Arousal dimension was chosen for the *Affrooms24* study while for the *Affrooms12* study we conduct both Arousal and Pleasure rating sessions. All data was collected within the *PAGAN* video annotation platform (Melhart et al., 2019). Figure 4.3 shows an instance of PAGAN during annotation whereby users could control the degree of affect change with their mouse wheel while watching a video of room navigations.

4.3.1 | Experiment Protocol

Affrooms24 study employed three participants experienced in *PAGAN* and the *RankTrace* annotation tool to annotate 20 different videos that included all the 24 room variations in random ordering. All participants were male, aged between 22 and 36, research staff at University of Malta with expertise on artificial intelligence, AC and digital games. The whole session for each user took about an hour to complete, while each participant was given the option to pause the session at any moment throughout. All appropriate consent was acquired by the participants and no personal data were retained. Each session was initiated with a corresponding definition of the affect label that is being rated by the annotator, a required step to minimize participant bias for potential misinterpretation

of the required task. Thus participants that were part of the arousal sessions received the following definition of arousal in the context of space appraisal:

> *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.*

Affrooms12 study aimed at including a broader sample recruiting 100 participants through *Amazon's mTurk* crowd-sourcing platform. Details regarding the recruited sample and background follow in Chapter 5. This study's experiment protocol was altered to incur less fatigue by including shorter spatial navigation videos (12 rooms each) and 6 sequences were required from participants to complete their session. These 6 sequences were picked at random from 55 pre-generated videos contained within the *AffRooms12* dataset. Annotations for both arousal and pleasure dimensions were collected assigning each participant to affect group at random. Figure 4.4 depicts the data collection pipeline for the crowdsourcing study. Participants that were picked for the Arousal sessions received the same Arousal definition as in Affrooms24 study and participants in the Pleasure sessions received the following definition:

> *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.*

4.4 | First-person Annotation of Active Stimuli

Capturing of real-time affect estimates during spatial exploration is an intricate task. The two preceding studies have concentrated on collecting continuous affect annotations from pre-recorded videos of spatial explorations. As described in Chapter 2, this method is termed throughout the dissertation as "first-person annotation of passive elicitor". With this method, interaction with the virtual world has already been performed and recorded, with annotators later tasked to indicate their affective responses to these stimuli interactions. This method offers several advantages, such as a reduced cognitive load for the annotator (by eliminating interactivity), and consistent stimuli across all annotators (everyone rates the same traversals and field of view recordings). This consistency facilitates the examination of inter-annotator agreements and enhances the

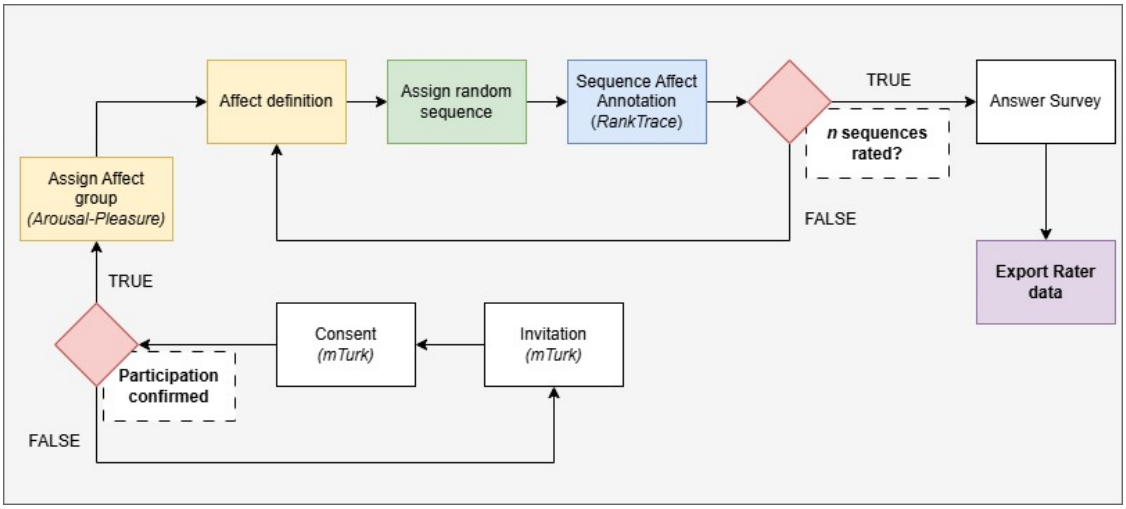

Figure 4.4: Data collection pipeline for the Affrooms12 study.

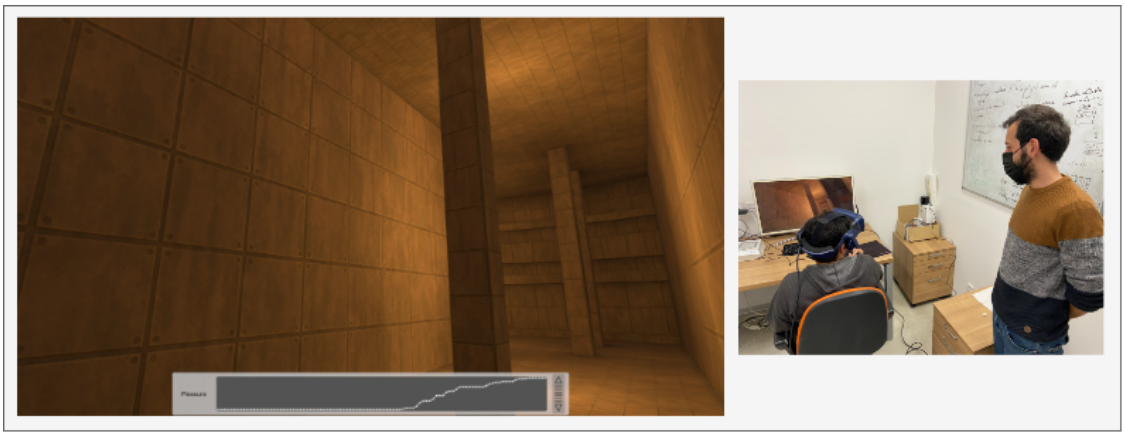

Figure 4.5: *Left*: RankTrace annotation embedded within the Virtual Environment, *Right* VR affect annotation session

validity of the data, while potentially forming a more reliable ground truth estimation based on these ratings (Yannakakis and Togelius, 2018).

While first-person annotation with passive elicitors has its merits, it also presents certain limitations. Reduced immersion and personal experience, resulting from the lack of direct interaction, can lead to less emotionally profound and authentic annotations. This diminished immersion may compromise the quality of the annotations, as they become less relevant to the individual rater and may introduce noise into the data. These limitations are addressed by the use of a method called *Stimulated Recall* (Shaker et al., 2012; Lankoski et al., 2015).

Stimulated Recall is an alternative to passive elicitation, where participants interact with a more dynamic environment. As described in Chapter 2 the stimulated recall protocol is widely use across interactive media where a lot of effort is spent in interactions with the synthetic environment i.e. video games. This method allows emotions to manifest during the interaction, followed by an annotation session where participants recall and annotate their affective experiences. The stimulated recall method bridges the gap between immediate emotional response and subsequent reflection, offering a more nuanced understanding of affective experiences while reducing the require load involved in affect annotation simultaneous interaction. As pointed out by Yannakakis and Togelius, 2018, drawbacks of this approach to collecting affect responses can face discrepancies between the "the experiencing self" and "the remembering self" (Yannakakis and Martínez, 2015) with a phenomenon known as the **experience memory gap** (Miron-Shatz et al., 2009).

One approach to minimize this limitation is with the embedding of the affect reporting mechanism within the VE. This type of method, has been explored by studies across various media of 360 video content (Xue et al., 2021; Fayn et al., 2022) and interactive architectural walkthroughs (McCall et al., 2022). Such methods enhance the immediacy and relevance of the emotional responses, providing a more direct and immersive way of capturing affect responses. Additionally, depending on the annotation task and the content that is annotated by the rater, this methodology can be considered more cognitive demanding when compared to the aforementioned approaches.

4.4.1 | Study Description

In the *AffroomsMR* study we investigate affect reporting within the VE (see Figure 4.5). This approach allows the gathering of real-time affect data during interaction with the VE. Using the natural sensorimotor contingencies of VR (Christofi et al., 2020), the spatial perception of the viewer is enhanced while processing environmental stimuli. This is achieved here via traversals of randomized sequences of different rooms on predetermined paths at a fixed movement speed, allowing solely the control of the camera (3-degrees of freedom). This ad-hoc fixed navigation ensures consistent duration for all annotators within each room and fixed room stimuli configuration for the same sequence. For the study we target two different media of display (Desktop and Head Mounted display) and compare acquired pleasure annotations.

To fulfill the needs of current and future experiments an application was developed within the Unity environment[2]. The application enables the session coordinator

[2]https://unity.com/

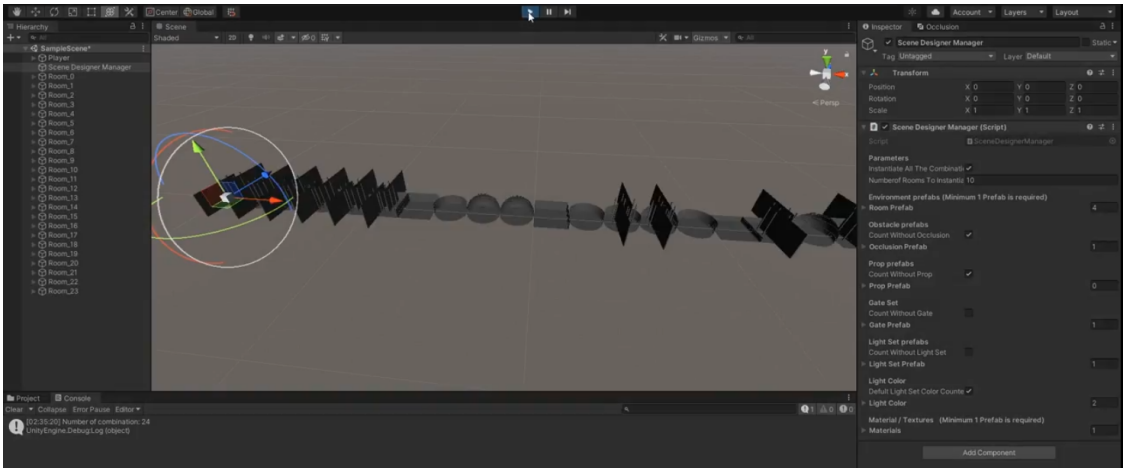

Figure 4.6: Generating Random Room Sequences for each participant session.

to generate a randomized sequence of all 16 rooms described above (see Figure 4.6). The room sequence remains consistent for both the Desktop and VR sessions. Similarly as in the first two user studies, the *RankTrace* annotation method is used to gather one-dimensional continuous and unbounded ratings of affect. *RankTrace* has been implemented within the VE with the annotation chart fixed at the lower-central part of the view both for VR and desktop views, overlapping the content; see Figure 4.5. While the participant is annotating, the entire chart history is visible to act as reference of previous ratings. Changes in pleasure are captured via mouse scroll movement (scrolling up: pleasure increase; scrolling down: pleasure decrease) and displayed on the chart. The avatar's location and FOV are registered every 100 milliseconds, while affect annotations are registered at every change. After each session a participant's data file includes the following: Timestamp, Participant ID, Display mode (VR-Desktop), annotated Pleasure value, Room ID, Location and Camera's pitch, yaw and roll values.

4.4.2 | Experiment Protocol

Figure 4.7 depicts the experiment protocol of the study. To eliminate any potential bias of medium order, each participant was randomly assigned to one of two groups: VR session first followed by the desktop session, or the opposite order. Each session starts with a pre-session survey and a consent form, assignment of stimuli group and a definition of Pleasure. Participants here receive the same definition for Pleasure as in the *Affrooms16* study; see Section 4.3.

Each participant carries out a trial run before starting with the session. During the

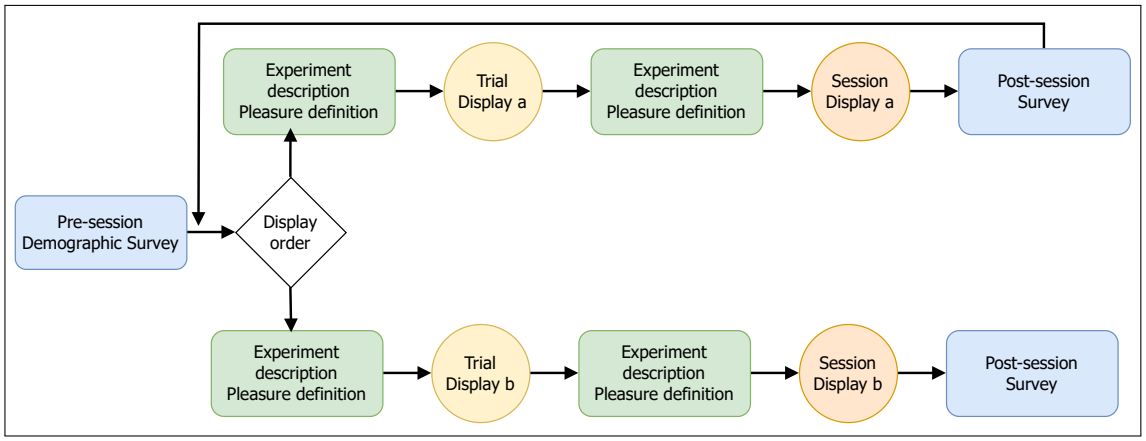

Figure 4.7: Experiment protocol flowchart

session, participants are encouraged to comment and think out loud but also to express if they want to leave the session. Once the session is finished, participants are instructed to fill out the post session survey. The same steps are then repeated once more for the second medium of display. For each session, the following data is collected: pre- and post-session questionnaire responses, room parameters, FOV pitch and yaw angles and pleasure annotations via the use of VE-embedded RankTrace. The main findings and results of this study are detailed in Chapter 6.

4.5 | Third-person Annotation of Active Stimuli

The previous studies focused on affect annotation methods that follow *forced-response* methodology. In the *OutlastAFF* study, the fourth and last study of this dissertation, we study manifested affect within synthetic spaces. Additionally, we explore video games as a potential case aimed at contributing to the existing body of knowledge on the interplay between emotion and space. Here we use an 'in-the-wild' approach of collecting affect responses from videos of gameplay streamed online (i.e. *Let's Play videos*). The focus is on environment variables and affect reactions, captured through third-person annotations and emotion recognition models analyzing streamers' facial and vocal reactions. Following our initial affect annotation and stimuli relationship diagram (see Chapter 2) this method is categorized as *3rd-person annotation of active stimuli*.

Figure 4.8: One of many "Let's play" moments that part of the corpus while traversing the Asylum level, including face camera overlay.

4.5.1 | Study Description

The outcome of this study is the *Outlast Asylum Affect Corpus* (OutlastAFF) . This dataset comprises streamed video content of the first level (the Asylum) of the 2013 popular survival horror title Outlast, (Red Barrels, 2013). Sixteen popular streamers (10 Male & 6 Female) were selected, whose gameplay and reactions were streamed via YouTube, resulting in 8.5 hours of annotated video gameplay. The multi-modal data includes annotated game events (e.g. room changes, jump-scare events, cut-scenes, game states), moment-to-moment facial expression labels, linguistic affect labels (what has been said) and paralinguistic affect ratings (how it has been said). The next sections describe the methodology of acquiring these aforementioned affect labels.

4.5.2 | Outlast Asylum Affect Corpus

Outlast is a horror game developed by Red Barrels and released for PC in 2013 and game consoles in 2014. In the game, the player takes the role of an investigative journalist and navigates a dilapidated psychiatric hospital overrun by homicidal patients. The game is played via a first-person camera perspective, and the player can move, jump, climb, crouch but not defend against or attack enemies. While encounters with enemies are generally sparse, the player can only outrun or hide from enemies. If a player dies, they

re-spawn at the most recent checkpoint; checkpoints are hard-coded by the level design. The plot of *Outlast* is mostly conveyed through dialogue with the few sane Non-player character (NPC)s remaining in the hospital, from notes strewn around the hospital, or from their own camcorder recordings which trigger "self-reflection" as voice lines of the player's character. The virtual environment tends to be fairly dark, prompting the player to make use of the night vision capabilities of their camcorder, which has limited power and must be recharged with batteries. Overall, *Outlast* tends to rely on jump scares and audio cues to trigger strong, visceral reactions which do not last very long. However, the background audio and environment design (with blood and gore) is likely to keep players alert and aroused. It is worth noting that jump scares are usually interactive (i.e. the player can move away from them) or non-interactive cut-scenes (where the game controls the player's actions for a short duration).

Asylum is the first level of *Outlast* and thus sets the mood and dangers of the game. The level includes 41 rooms across three floors. The player's trajectory, at least in the first half of the play-through, is predetermined through locked doors and barricaded hallways. This design pattern is common in tutorial levels, allowing the game designers to control which parts of the mechanics or story are shown to the player and in which order. This is also convenient for our affect corpus, as the order of streamers' reactions is expected to match. Each floor contains diverse rooms, but the basement especially features many narrow and dark spaces. The Asylum level features two hostile NPCs, while many NPCs are neutral bystanders but contribute to the eeriness via audio cues and jump scares.

4.5.3 | Properties of the Raw Video Data

The video data was collected via YouTube, in the form of 16 complete runs of the *Asylum* level of *Outlast* from the 16 YouTube streamers. Streamers were chosen for their popularity (each channel has thousands of subscribers) with special attention towards diversity (we can not ascertain the gender identities of the streamers, but we aimed to capture as diverse a population as possible). Some runs through the *Asylum* level were split up to 4 separate videos which were merged in terms of their labeling during data processing (see Section 7.1). Whether combining multiple videos or processing one video, segments introducing the game and the stream were removed. All processed data have a full-screen view of the game (containing its own audiovisual content) and a face camera overlay of the streamer along their own vocal narration of the play-through (see Figure 4.8). Complete play-throughs in the dataset lasted between 22 and 48 minutes.

4.5.4 | Manual Labeling of Features

We labeled each video in the Outlast Asylum Affect Corpus in terms of timings when events occurred in each video. Since this study is mostly invested on the impact the virtual space has on affect, each play-through is split according to the room the player is in (i.e. based on timings when the player entered and exited the room). Rooms are assigned eight *spatial* descriptors (see Table 4.2) that, for the most part, describe the room size, its navigability (blocked paths), its contents (e.g. empty, neatly arranged furniture, or chaotically strewn debris and corpses), and the level and color of the illumination. Moreover, rooms may contain gameplay elements, or trigger new interactions. These *game* features are also labeled, as they affect the gameplay affordances of each room. Seven game features are labeled (see Table 4.2), including the presence of hiding places, batteries, triggers that allow the player to continue their level traversal (e.g. keys, levers), notes, and events (e.g. interactive jump scares or audio cues) or cut-scenes in the room. We treat each room visit as one set of such values for the entirety of the room visit. We note that players may revisit rooms they were in before: if the conditions are different (e.g. lights that were previously on are now off, or the player already picked up the battery in this room) the features are labeled accordingly.

4.5.5 | Generating Multi-modal Labels of Affect

The available modalities in the Outlast Asylum Affect Corpus that attempt to capture the streamer's affective state are the streamer's facial expressions and their voice. We leverage established pre-trained models for capturing affect via facial expressions, and process both the utterances (i.e. what is said) and audio information of the voice (i.e. how it is said) for affect recognition. The resulting three data streams (see Fig. 4.9) are sampled at 1Hz throughout the entire video, and capture specific emotional dimensions relevant to horror gameplay. Specifically, we are interested in *arousal* levels and, from categorical emotions Ekman, 1992, we are interested in *fear* and *surprise* as the most targeted by this type of game: fear from disturbing imagery such as blood and gore and surprise from jump scares. These models have not been further fine-tuned in the task of recognizing affect from content related to gameplay or more specifically the context of Survival Horror. The raw data points with extracted affect per modality can be found in Chapter 7.

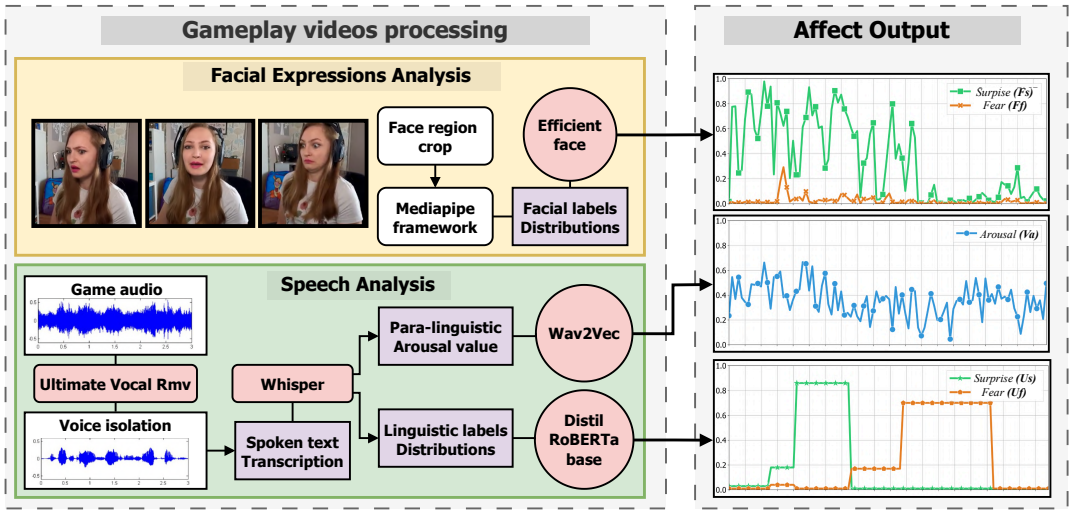

Figure 4.9: Affect labels from different modalities of the streamer's face camera and audio, derived through pre-trained models. Face camera depictions are from YouTube user AnidaGaming; image used with the streamer's permission.

4.5.5.1 | Facial Expression

For this modeling task, a preliminary step is carried out whereby streamer's face region on each video is cropped (see Figure 4.9). Categorical emotions are assigned via Google's Mediapipe (Google, 2023), using the 'efficient face' pre-trained model of Zhao et al., 2021. This model outputs probabilities within $[0, 1]$ for seven labels: anger, disgust, fear, joy, sadness, surprise, neutral. Since the model takes frames as input (with 30Hz sampling rate), we average the probabilities per second (1 Hz sampling rate) and retain probabilities for fear (F_f) and surprise (F_s). This procedure is essential to mitigate the impact of mislabeled frames, which primarily arise due to sub optimal lighting conditions and camera-related anomalies on the streamers side. This step is particularly crucial given that the existing dataset was compiled in uncontrolled, real-world setting and not in lab.

4.5.5.2 | Streamer's Voice (Para-linguistic Data)

To analyze affect in the player's speech, we first isolate the streamers' vocal cues from the game with the *'Ultimate Vocal Remover'* application (Anjok07, 2023; Takahashi and Mitsufuji, 2017). This crucial stage guarantees that the extracted linguistic (what has been said) and para-linguistic (how it has been said) vocal streams are free of game-related audio, ensuring purity for further analysis further down the pipeline. We then

apply Speech Emotion Recognition on the isolated vocal cues using *Audeering's*[3] fine-tuned version of the wav2vec 2.0 model (Wagner et al., 2022; Baevski et al., 2020). This model outputs arousal, valence, and dominance values at 1 Hz sampling rate. We retain only the arousal values (V_a), the most relevant dimension for the study and the Horror games genre. Additionally, affect from speech recognition models display higher results for the arousal and dominance dimensions. Similarly as in the case with *Audeering's* model, the paralinguistic stream achieves better scores in Arousal dimension rather than the Valence and dominance (Wagner et al., 2022).

4.5.5.3 | Streamer's Utterances

One of the differences between traditional players and streamers is that the latter constantly converse with their audience. Thus, streamers' utterances can be information-rich, although the context of these utterances may be different from the current game context (such as discussing the streamer's day or commenting on something an audience member said). We use the isolated vocal cues from Section 4.5.5.2 and transcribe them as text using OpenAI's Whisper (Radford et al., 2022). Then, the transcription is processed via the Emotion English DistilRoBERTa-base (Hartmann, 2022) model which returns probabilities within $[0, 1]$ for seven labels (anger, disgust, fear, joy, sadness, surprise, neutral). As in Section 4.5.5.1, we aggregate the probabilities per second at 1 Hz sampling rate and retain only probabilities for fear (U_f) and surprise (U_s).

Para-linguistic Emotion Analysis: As a next step Speech Emotion Recognition (SER) is applied on the isolated vocal cues. Using Audeering's fine-tuned version of the wav2vec 2.0 model (Wagner et al., 2022; Baevski et al., 2020), the paralinguistic stream is extracted (how the speech is delivered). Specifically, the model produces Arousal, Valence, and Dominance values. Notably, the Arousal component, is more accurately captured by Audeering's model when compared to Valence and Dominance.

Transcription and Emotion Analysis: The third affect stream focuses on the linguistic content, what has been said, converting spoken text to emotion labels. This is achieved by leveraging two advanced pre-trained models. OpenAI's Whisper (Radford et al., 2022) is used on the isolated vocal cues described above, for transcribing the streamers' spoken content. Subsequently, Jochen Hartmann's model (Hartmann, 2022) assesses the transcribed text, retrieving the probability distributions on Ekman's six basic emotions.

Through the process outlined above, three distinct affect streams are generated. Both the Facial and Linguistic affect streams yield seven different emotion labels along with

[3]https://www.audeering.com/

their respective probabilities. However, only the labels of Surprise and Fear are employed for these streams. For the Paralinguistic stream, alternatively referred to as the Vocal stream, the focus lies on the dimension of Arousal. This approach allows the study to concentrate on Tension and Arousal, which are particularly relevant and fitting for the context of the Survival Horror genre.

4.6 | Summary

The present chapter described the methodology of acquiring continuous affect responses from four subject studies that are part of this dissertation. In each section we detailed the main differences between these studies and the featured *stimuli* and *annotation method*. Although these studies differ in their approaches, they all address the primary objective of modeling the affective impact of various spatial conditions. The following chapters present the results and details of these studies. Chapter 5 presents the results of *Affrooms24* and *Affroom12*, using the 1st-person annotation of Passive stimuli. Chapter 6 covers the results of *AffroomsMR* study which uses 1st-person annotation of Active stimuli. Finally, Chapter 7 discusses the results of *OutlastAFF* study that uses a third-person annotation on active stimuli, with horror video games serving as our case study.

5

Annotation of Videos

This chapter presents the results of the first two participant studies of this Phd dissertation. These studies follow a forced response method via continuous and unbounded annotations of Arousal and Pleasure (Valence). The affect data was collected using the Ranktrace annotation tool (Lopes et al., 2017b) and the PAGAN affect annotation platform (Melhart et al., 2019). Participant annotations were gathered on pre-recorded videos depicting spatial walkthroughs traversed in a *First-person manner* as described in Chapter 4.

The literature reviewed in Section 2.2 encompassed a diverse array of methods for gathering affect feedback across various stimuli, particularly emphasizing those investigating the impact of spatial experiences. As we highlighted in Section 2.2, the review identified significant gaps pertaining to *capturing* observer feedback and *representing* spatial stimuli while considering *temporal* factors. Addressing these gaps, the *Affrooms12* and *Affrooms24* studies conducted within the scope of this dissertation were centered on collecting continuous affect annotations in response to environmental parameters of 24 different rooms (see Figure 5.1). This involved utilizing pre-recorded navigational videos of synthesized spaces. As described in Chapter 4, these studies employed a "first-person annotation of passive elicitor" approach. Notably, these two use studies diverge from the approach adopted in the third user study (documented in Chapter 6), employing a "real-time first-person annotation". In this latter approach, annotators directly interact with the virtual environment during affect annotation, offering immediate feedback on their affective experience.

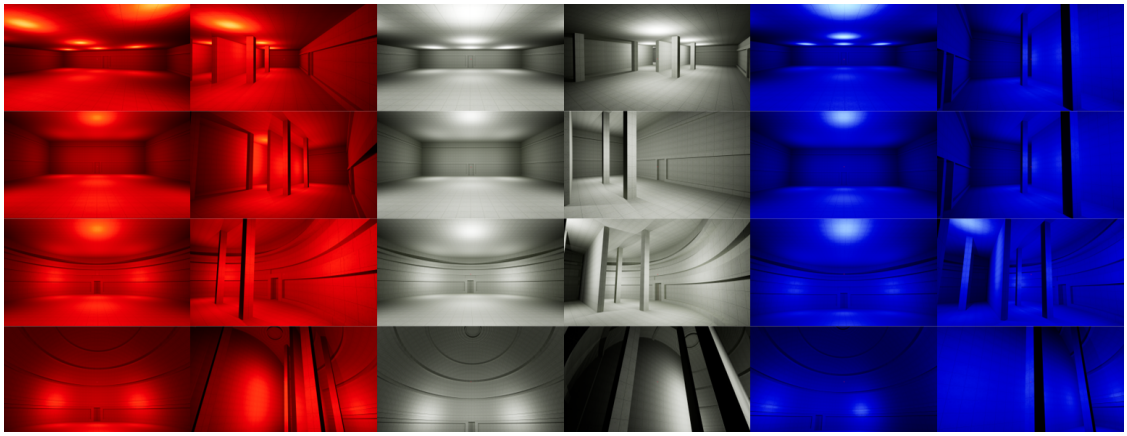

Figure 5.1: Views of the 24 rooms examined in the *AffRooms24* and *Affrooms12* corpora

5.1 | Expert Annotators

The first study on passive stimuli was conducted using only three experts and an extensive data collection protocol in order to generate as many combinations of rooms as possible for the selected sample size; see Section 4.3 for more details regarding the study's data collection pipeline. Arousal is the chosen dimension for this study, a relevant concept in architectural design that is linked to *interest* and *saliency* on regions of a specific scene (Chamilothori, 2019;Karmann et al., 2023). Figure 5.2 depicts the acquired arousal traces using the *PAGAN annotation platform*, for the three expert annotators on the same walkthrough video. The depicted traces for a single walkthrough can already display a somewhat similar annotation behavior present for two participants—the arousal traces in blue and orange— which are analyzed further below.

5.1.1 | Affrooms24 Analysis

The first steps of the analysis include: a) normalization of the collected arousal traces in a per participant per video manner, b) the window assignment processes for both types of window sizes (arrival windows and room windows) and c) the computation of affect measures within each window considering both relative (gradient and amplitude) and absolute (mean) measures of affect. On these derived measures for all rooms and arrivals, the pairwise transformation process results in acquiring comparisons between all windows with their adjacent window for each corresponding walk-through sequence. These pairwise comparisons (or *deltas*) are initially analysed across windows and affect measures. Since room features are categorical, determining whether the fea-

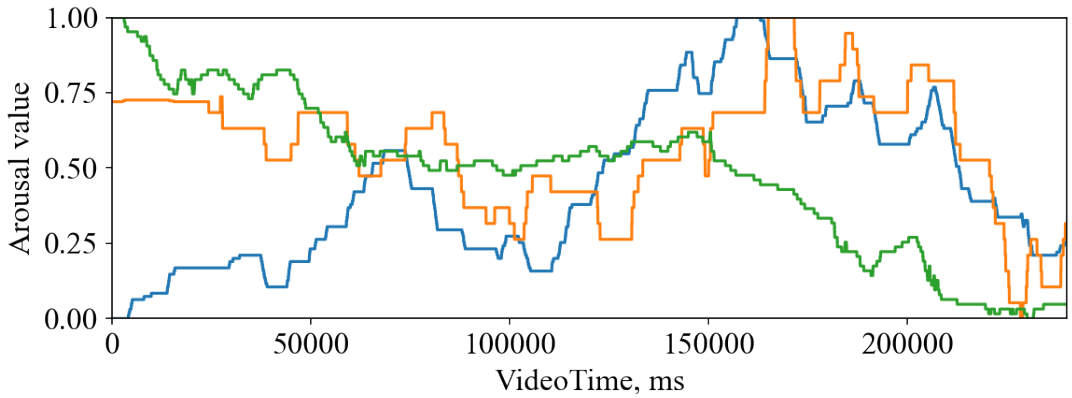

Figure 5.2: Arousal traces for the three expert annotators on the same video of the Affrooms24 dataset.

ture changes is straightforward. More details regarding the processing pipeline are described in Chapter 3. In the 20 videos recorded (i.e. 23 transitions per video or 460 transitions in total), there are 219 changes in curvature, 238 changes in size, 325 changes in light color, and 243 changes in complexity between consecutive rooms. Arousal measures changes are values within $[-1, 1]$ and thus what constitutes an affect change between the mean arousal values of two consecutive rooms is to be determined by the *Uncertainty threshold* in the next step of this analysis. (see Section 3.2.1). Figure 5.3 displays the distributions of these *delta* comparisons between affect measures for the two types of window sizes before applying an uncertainty threshold. What can be initially observed by comparing these two different windows and their corresponding affect measures is that there is a general trend of Normal distribution, with the Amplitude measure for room windows displaying a slight positive skewness. Additionally looking at the differences between participants, we can distinguish one participant (highlighted in orange) displaying milder affect changes than the other two, that display more similar affect changes with each other. These idiosyncratic tendencies in the signal are acknowledged throughout the processing pipeline, from scaling, to applying the uncertainty threshold.

As secondary step in this analysis is to determine if there are any linear correlations between room feature change and affect measure change within the two proposed windows of *Room window* and *Arrival window*. For this *Spearman's rank correlation coefficient* is used fitting the ordinal treatment of the dataset. Table 5.1 depicts the results of this analysis. What can be initially inferred by looking at Table 5.1 is that regardless of the chosen window of analysis, Occlusions demonstrate the highest correlation between the two variables, with Arousal Mean displaying a mild positive correlation of 0.33 for the

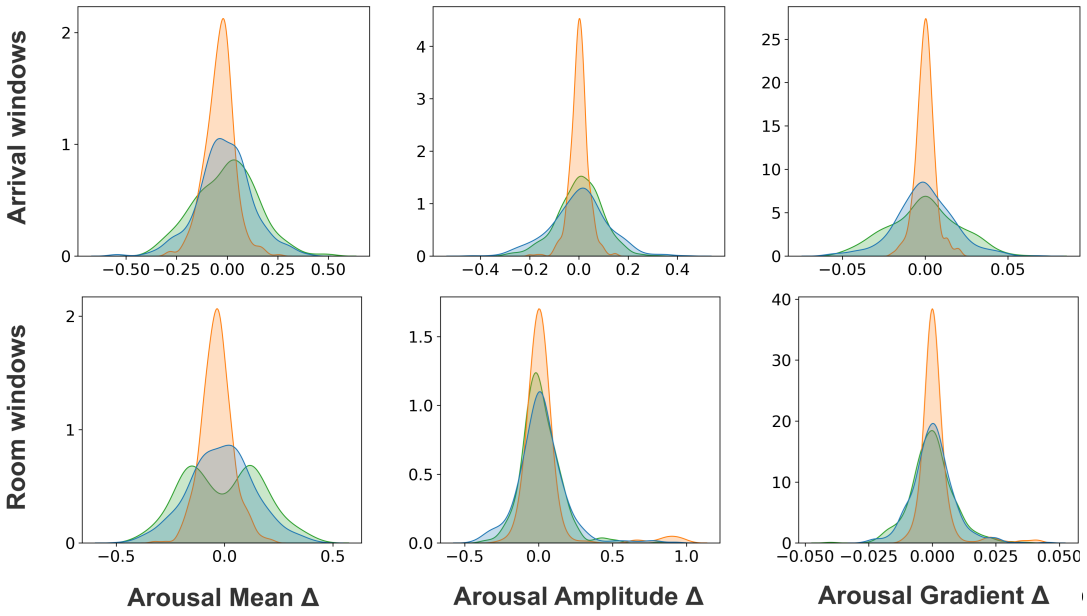

Figure 5.3: Distribution of Affect measures deltas for Arrival windows (top) Room Windows (bottom) for each participant.

whole room and while in the case of arrival windows the Gradient measure displays the highest ρ value of 0.37. Following Occlusions, Curvature changes between rooms display fair correlation strength for the same measures and windows with mean in room window yielding 0.21 and arousal gradient within arrival windows 0.26. Height displayed very weak results with correlations between affect measure change and feature change close to 0.1 and Illumination displaying the weakest relationship out of the 4 selected parameters, with results close to 0. Significance level for the correlation analysis is set at $p = .05$. These initials results on the *Affroosm24* dataset did not display conclusive findings regarding a linear relationship between affect measure changes and feature changes for the two chosen windows, but merely uncovered which measures could perform well under the selected window of analysis. For the features of Occlusions and Curvature using the whole room as processing window, affect mean did produce stronger correlations, while for the Arrival windows the gradient measure displayed the best indications. These preliminary findings could suggest that the gradient measure can perform better under smaller window sizes. We choose to investigate further arousal mean for room windows and arousal gradient mean for arrival windows in the subsequent steps of this analysis. Additionally, the next section we use two reliability

Table 5.1: Spearman's rank correlations between Room Feature change and Affect Measure change. Bold values highlight significant relationships at p=0.05.

Room window				
Affect Measure	Curvature	Height	Illumination	Occlusions
Arousal Mean	**0.21**	**0.07**	**-0.04**	**0.33**
Arousal Amplitude	**0.08**	0.02	0.06	**0.3**
Arousal Gradient	**0.11**	**0.09**	0.02	**0.12**
Arrival window				
Arousal Mean	**0.1**	0.01	-0.03	**0.16**
Arousal Amplitude	**0.1**	**0.08**	-0.02	**-0.23**
Arousal Gradient	**0.26**	**0.11**	-0.06	**0.37**

measures (agreement tolerance and uncertainty threshold), with the aim of uncovering the relationship between ratings of arousal and spatial feature changes.

5.1.2 | Agreement Analysis

Preliminary analysis considered all arousal measure changes regardless of their degree. This can prove problematic if minor affect changes are considered which could have been introduced by rater uncertainty, an expected phenomenon across continuous rater annotations, especially for longer rating sessions. We introduce an uncertainty threshold ϵ as a measure of reducing the present noise in the rater signal, whereby if the absolute difference between arousal measures of consecutive rooms is below this threshold then it is considered that there is no change in arousal; see Section 3.2.1 for the detailed description. The $\epsilon = 0.05$, i.e. 5% of the value range of each arousal trace is determined for this study. From one room to the next where a specific design feature changes, considering that there is an arousal increase or decrease is marked as an *arousal shift*. If the design feature is absent in the previous room but is present in the next room (or the color warmth increases by at least one step) and arousal increases, then this is marked as *agreement*; if arousal decreases this is marked as a*disagreement*. Similarly if the design feature is present in the previous room but is absent in the next room (or the color warmth decreases by at least one step) and arousal decreases, this is marked as an *agreement*, and if arousal increases it is marked as *disagreement*. Through this process all instances where the change in a design feature has a corresponding change in arousal are enumerated following a count analysis. A high agreement ratio means that the presence of a design feature leads to higher arousal (higher number of positive counts), while a ratio of arousal shifts over the total number of feature changes in transitions means that annotators have a reaction when this particular feature changes (most impactful

Table 5.2: Changes in arousal annotations matching with room properties, per annotator and based on agreement between annotators. Significant agreements or disagreements are shown in bold. Arousal shifts note the ratio of instances where there was arousal changes for the corresponding feature change. N displays the number of datapoints (changes in mean arousal) remaining, also as ratio over all 460 transitions.

	Curvature	**Room Size**	**Color Warmth**	**Complexity**
annotator A ($N = 413$, 90%)				
agreements	**58%**	55%	54%	**97%**
disagreements	42%	45%	46%	3%
arousal shifts	91%	89%	89%	94%
annotator B ($N = 351$, 76%)				
agreements	**83%**	**57%**	**56%**	**60%**
disagreements	17%	43%	44%	40%
arousal shifts	81%	76%	76%	77%
annotator C ($N = 250$, 54%)				
agreements	49%	52%	50%	50%
disagreements	51%	48%	50%	50%
arousal shifts	57%	51%	53%	56%
At least 2 annotators agree on room arousal change ($N = 254$, **55%**)				
agreements	**75%**	54%	**59%**	**87%**
disagreements	25%	46%	41%	13%
arousal shifts	61%	56%	55%	57%
All 3 annotators agree on room arousal change ($N = 49$, 11%)				
agreements	**87%**	61%	51%	**94%**
disagreements	13%	39%	49%	6%
arousal shifts	14%	10%	11%	15%

features).

Table 5.2 shows the agreements between mean arousal changes and design feature changes, per annotator. It is evident that some annotators were less prone to shift their arousal annotations between rooms (e.g. observing the overall arousal shifts of annotator C). On the other hand, annotator B is fairly consistent, and the presence of every spatial feature is more often associated with increased arousal than not. Significance of the ratio of agreements versus disagreements is calculated based on the binomial distribution of all arousal changes when the spatial feature changes, assuming a 50% probability that the changes may be in agreement. Significance is established at 95% confidence. It is evident that different annotators provide traces with different degrees of granularity, while some extremes (e.g. 97% agreement in terms of complexity for annotator A) raise some concerns discussed in Section 5.3. Table 5.3 shows each annotators' agreements be-

Table 5.3: Agreement between sign of the arousal gradient and changes in room properties during an arrival. Significant agreements or disagreements are shown in bold. Arousal shifts note the ratio of instances where there was arousal changes for the corresponding feature change. N displays the number of datapoints (non-zero arousal gradients) remaining, also as ratio over all 460 transitions.

	Curvature	**Room Size**	**Color Warmth**	**Complexity**
annotator A ($N = 432$, 94%)				
agreements	**57%**	53%	53%	**96%**
disagreements	43%	47%	47%	4%
measure shifts	95%	95%	94%	98%
annotator B ($N = 347$, 75%))				
agreements	**84%**	**58%**	**56%**	**60%**
disagreements	16%	42%	44%	40%
measure shifts	80%	75%	75%	74%
annotator C ($N = 260$, 57%)				
agreements	53%	50%	49%	47%
disagreements	47%	50%	51%	53%
measure shifts	85%	90%	88%	90%
At least 2 annotators agree on sign of arousal gradient ($N = 309$, 67%)				
agreements	**73%**	51%	**56%**	**80%**
disagreements	27%	49%	44%	20%
measure shifts	72%	69%	67%	69%
All 3 annotators agree on sign of arousal gradient ($N = 89$, 19%)				
agreements	**84%**	59%	59%	**90%**
disagreements	16%	41%	41%	10%
measure shifts	22%	21%	20%	20%

tween arrival gradients and changes in design features. Unsurprisingly, results follow a similar pattern as with changes in mean arousal, with annotator B showing significant impact of all design features on arousal change during the arrival time window and annotator C being more ambiguous in their annotations. Notably, annotator C displayed clearer patterns than with mean arousal changes. Additionally, the number of instances where arousal gradient was non-zero when a feature changes is increased over the times mean arousal changes (see "arousal shifts" entries in Table 5.2). This may indicate that focusing on the arrival windows and their arousal gradient could provide more concise data, although we cannot discount other effects on data processing such as the different thresholding procedures for the two signals.

5.1.2.1 | Inter-rater Agreements

Calculating the instances where at least two annotators are in agreement, and matching them with changes between consecutive rooms, 254 arousal shifts remain out of a total of 460 room transitions for room mean arousal and 309 shifts for arrival mean gradient, which is a good sample for data analysis. Table 5.2 includes the agreements, disagreements and arousal shifts with each spatial feature change in consecutive rooms for instances where at least two annotators agree. It is evident from Table 5.2 and Table 5.3 that higher complexity and higher curvature leads to higher arousal, with warmer colors also coinciding with higher arousal but to a lesser degree. A surprising outcome, as it is expected that features of color would have a more noticeable effect than all features of form.

For completeness, an additional analysis on those instances where all three annotators agreed displays even more pronounced scores, however the number of shifts where all three annotators have a non-trivial increase or decrease is decreased substantially to 11% for mean arousal and 19% for gradient. This means that most of the data are lost for the sake of complete inter-rater agreement, acknowledging this weakness for this sample size.

5.2 | Crowd-sourcing Non-expert Annotations

Crowd-sourcing as a method for participant recruitment in experimental research harnesses the power of diverse populations, significantly broadening the test sample to capture a wider array of perspectives and enriching the ethnographic diversity for a study. By tapping into this large pool of participants from varied backgrounds and expertise, crowdsourcing ensures a richer, more representative dataset, which is essential for forming robust and inclusive models. This approach, offers an expansive and multifaceted view of human behavior and responses, crucial for developing accurate predictive models that reflect the complexity of real-world scenarios especially in the areas of modeling affect (Camilleri et al., 2016; Melhart et al., 2021a; Rozado et al., 2022; Shaker et al., 2012).

The study on the *Affrooms12* dataset expands significantly the work of the *Affrooms24* study in several ways. Firstly, by introducing a broader collection of shorter spatial navigation videos that incur less fatigue and cognitive load from participants. Second, a broader sample of the population, through crowd-sourcing platforms was included, and collected annotations from 71 participants in total. These participants were not experts in the annotation protocol, unlike the three expert annotators in the previous

study. Third, unbounded time-continuous annotations of both arousal and pleasure dimensions were collected. Fourth, inspired by Cowie and McKeown, 2010 and Sethu et al., 2019 this study extended the processing frameworks levels on each parameter, estimating their ability in forming more reliable and valid datasets of affect ratings. Finally, beyond one-to-one mapping between each design feature and affect reported of the previous study, here we leverage supervised learning that combines the different spatial features together to train models of arousal and valence from spatial transitions and detect the most impactful design parameters via the trained model's feature importance vectors.

5.2.1 | Participants

A total of 100 individuals contributed to this study using Amazon's Mechanical Turk crowd-sourcing environment and the PAGAN annotation platform. In contrast to the first study, this study explored the additional Valence (or Pleasure) dimension, with 50 participants reporting on their Arousal and the remaining 50 reporting on their Pleasure. Before the session, participants completed a survey that collected background information, including gender, familiarity with arts, design, architecture, and their experience with video games and virtual worlds. Familiarity was rated on a 5-point likert scale, where 1 indicated low familiarity and 5 indicated high familiarity. The gender distribution was 66.4% male and 33.6% female. Participants reported a high level of familiarity with video games, with 50.9% rating themselves at level 5, 33.6% at level 4, 12.1% at level 3, and only 3.4% at lower levels. Similarly in familiarity with arts, design and architecture, 37.1% rated themselves as 5, 33.6% at 4, 12.1% at 3 and 17.3% for the lower levels.

5.2.2 | Affrooms12 Analysis

In Affrooms24 we looked into comparisons of four essential design features in adjacent rooms within a single walkthrough sequence and retrieved their agreements and disagreements regarding reported affect changes of the annotators. This was achieved via a relative approach rather than analyzing absolute affect measurements (Yannakakis et al., 2017) for each room. Similarly, we follow here the same relative approach with additional features for the classification task. Since navigation times in the recorded videos differ between rooms (especially in the case of rooms with many occlusions), the total duration (in seconds) spent traversing both rooms is included as an additional input, exploring how stimuli duration can impact affect changes. This results in 12 input

features related to design parameters and 1 feature input for the total room traversal duration.

A number of steps were taken to clean up the dataset. First, incomplete sessions and duplicates were removed. Second, interactions with the annotation tool were checked: since RankTrace can be used in a continuous fashion, whenever a user interacts with it (e.g., registers a change in affect), this is logged. Following the literature on time-continuous affect annotation (Melhart et al., 2021a), traces with fewer than 10 interactions in total (i.e., affect changes) were removed, considering the user was idle and not meaningfully interacting with the study. Lastly, navigation videos that were not annotated by at least 3 participants were also rejected, as inter-rater agreement is important for the analysis. After this cleanup process, the dataset contained 224 annotated videos of arousal from 39 participants and 215 annotated videos of pleasure from 37 participants. Results for the two different tasks of classification and agreement analysis follow. For the classification task, the obtained results focus on *Random Forests* for both affect labels using the model's accuracy (the number of correct predictions divided by the total number of predictions), a fitting measure suited for balanced datasets in binary classification tasks, followed by the train dataset size in parentheses. For the linear estimation task, results are presented as the percentage of agreements and disagreements between affect measure change and feature change for three different inter-annotator agreement thresholds, followed by a comparison with the expert group.

5.2.3 | Agreement Analysis

In addition to the classification task on the *Affrooms12* dataset, an additional analysis similar to the first study is conducted, looking into agreement ratios between feature changes and affect measure changes to compare experts and non-expert annotators. This analysis did not exclude any annotators, unlike in the classification task where 3 annotators for each affect label were excluded for hyper-parameter tuning. The uncertainty threshold epsilon is set to 5% to eliminate signal noise, and inter-annotator agreement thresholds are set to 0% (no consensus between raters), 66%, and 75%. Lastly, only the consecutive pairs in each video are considered (memory setting 1). These settings were selected to make a comparable case with the Affrooms24 study.

Tables 5.4 and 5.5 display the results from the one-to-one agreement ratios on the crowd-sourcing *Affrooms12* dataset. What can first be observed for both Arousal and Pleasure affect labels is the general divergence for both Agreement and Disagreement counts throughout the four design features and the chosen Affect metrics (as agreeing and disagreeing values get further from 50%, the stronger the relationship) as inter-

Table 5.4: Agreement and Diagreement between Affect stat and Feature change, with 0%, 66% and 75% inter-annotator agreement tolerance. Bold highlights significant measures at 0.05 p value

Arousal					
Metric	Trend	Curvature	Room size	Color Warmth	Complexity
Inter-annotator AT 0% (N = 3566)					
Mean	Agreeing	**53%**	48%	**55%**	**53%**
	Disagreeing	47%	**52%**	45%	47%
Amplitude	Agreeing	52%	**54%**	47%	**69%**
	Disagreeing	48%	46%	**53%**	31%
Gradient	Agreeing	**54%**	**57%**	49%	**79%**
	Disagreeing	46%	43%	51%	21%
Inter-annotator AT 66% (N = 1098, 31%)					
Mean	Agreeing	**55%**	40%	**65%**	**55%**
	Disagreeing	45%	**60%**	35%	45%
Amplitude	Agreeing	**56%**	**63%**	46%	**94%**
	Disagreeing	44%	37%	**54%**	6%
Gradient	Agreeing	**61%**	**66%**	50%	**97%**
	Disagreeing	39%	34%	50%	3%
Inter-annotator AT 75% (N = 536, 15%)					
Mean	Agreeing	**69%**	47%	**70%**	**63%**
	Disagreeing	31%	53%	30%	37%
Amplitude	Agreeing	**64%**	**71%**	41%	**98%**
	Disagreeing	36%	29%	**59%**	2%
Gradient	Agreeing	**66%**	**73%**	50%	**97%**
	Disagreeing	34%	27%	50%	3%

annotator agreement thresholds are applied. A considerable portion of the dataset is lost in the process as the (a_t) increases, and a strict rule is applied requiring at least 3 annotators per pair (pairs with fewer than 3 annotators are discarded in this step and do not count towards concordances). For the 66% (a_t) threshold, 31% and 35% of the dataset (for Arousal and Pleasure ratings respectively) is retained, while for the 75% (a_t) threshold, 15% and 21% of the dataset remains. What can be observed here initially is that, when compared to the retained dataset size of the expert group in Table 5.2, when two raters agree, 55% of the dataset is retained, displaying a significant difference when compared with the crowd-sourced non-expert group.

Regarding the linear connections between design features and affect measures, there is a noticeable effect that occluded rooms have on the annotation manner in both the Arousal and Pleasure sessions, displayed mainly in the amplitude and gradient of the binned signals, where annotators register higher affect changes and more frequently.

Table 5.5: Agreement and Diagreement between Affect stat and Feature change, with 0%, 66% and 75% inter-annotator agreement tolerance. Bold highlights significant measures at 0.05 p value

Pleasure					
Metric	Trend	Curvature	Room size	Color Warmth	Complexity
Inter-annotator AT 0% N=3522					
Mean	Agreeing	49%	52%	45%	49%
	Disagreeing	51%	48%	**55%**	51%
Amplitude	Agreeing	**54%**	50%	48%	**68%**
	Disagreeing	46%	50%	52%	32%
Gradient	Agreeing	52%	**54%**	48%	**82%**
	Disagreeing	48%	46%	52%	18%
Inter-annotator AT 66% N=1228, 35%					
Mean	Agreeing	45%	**54%**	35%	49%
	Disagreeing	**55%**	46%	**65%**	51%
Amplitude	Agreeing	**60%**	**55%**	41%	**89%**
	Disagreeing	40%	45%	**59%**	11%
Gradient	Agreeing	51%	52%	43%	**94%**
	Disagreeing	49%	48%	**57%**	6%
Inter-annotator AT 75% N=742, 21%					
Mean	Agreeing	**58%**	52%	39%	49%
	Disagreeing	42%	48%	**61%**	51%
Amplitude	Agreeing	**65%**	48%	36%	**92%**
	Disagreeing	35%	52%	**64%**	8%
Gradient	Agreeing	58%	47%	49%	**97%**
	Disagreeing	42%	53%	51%	3%

Regarding the mean affect measure, illumination color displays the highest impact, aligning with the results from the previous method on the feature importance of the Random Forest classifier (see Section 5.2.4), and suggesting higher arousal for warmer illumination. The discrepancy for Pleasure suggests a preference for colder illumination. Additionally, the mean measure displays a significant increase in Arousal agreement for curvature at the 75% threshold, similar to the finding displayed for the expert sample in the pilot study (see Section 5.1.2.1), and a mild increase in the Pleasure study at the 66% threshold. Lastly, for room height, Pleasure ratings did not yield significant findings but did display a mild disagreement in mean arousal and considerable disagreement increases for amplitude and gradient measures. This could be attributed to the awe effect (Heath et al., 2000) that raters might have experienced when leaving a small room to enter the dome structure. What needs to be addressed here is that agreements may show minor and major increases in some cases, but this comes at the cost of the dataset

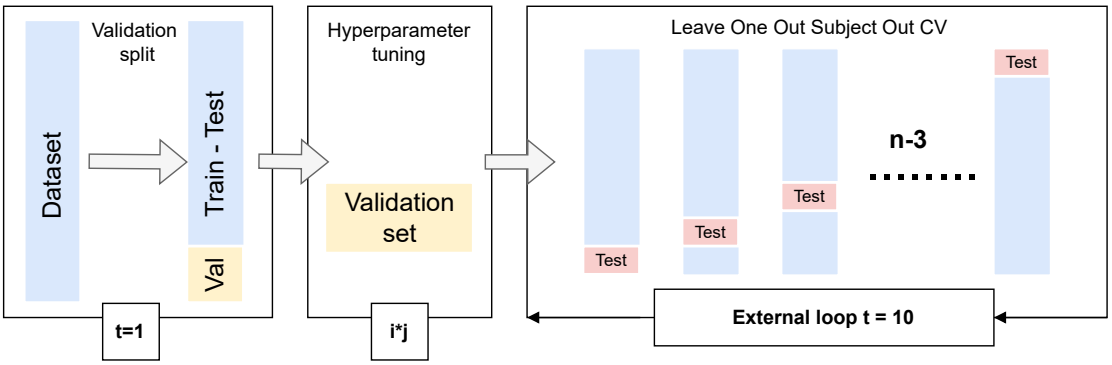

Figure 5.4: The Leave-One-Subject-Out (LOSOCV) cross validation protocol for the Random Forest classifier. Each comprised step indicates its assigned number of iterations bellow.

size, leaving room for uncertainty due to limited data size.

5.2.4 | Modeling Task

In Chapter 3 we described the processing framework and its comprised components. This extensive framework produces datasets of varying sizes by adjusting the Uncertainty Threshold (ϵ), Memory (m), and Inter-rater Agreement Tolerance (a_t) parameters; see Section 3.2 and Section 3.3. It is expected that the least nuanced data cleanup i.e. $a_t = 0\%$, $\epsilon = 0$ and $m = \infty$, would result in the worst performing models. These parameter tests where carried out with four different classifiers: Random Forests, linear and non-linear Support Vector Machines and Naive Bayes. To ensure a 50% random guess baseline for the predictive models, balanced datasets are created for all conditions of the processing framework. The process of creating balanced datasets within this framework is described in Section 3.2.1.

5.2.4.1 | Train-Test Protocol

To evaluate the performance of each model, *Leave-One-Subject-Out cross-validation* (LOSOCV) is performed as illustrated in Figure 5.4. Assuming that n participants contribute with their annotations, 3 random participants are isolated from the dataset for hyper-parameter tuning for each of the aforementioned classification methods, forming the external validation set, while the remaining $n - 3$ participants undergo leave-one-participant-out cross-validation for train and test. The first step of determining a validation set occurs just once for 3 specific raters chosen at random. The hyper-parameters tuned by the validation set are as follows: a) for Random Forests: the number of trees,

tree depth, minimum number of samples per leaf node and the minimum number of samples required to split a tree node, b) for SVMs *C* and *Gamma* and c) for Naive Bayes the smoothing parameters *alpha*. All hyper-parameters are tuned via exhaustive search targeted at maximizing *accuracy* on the validation set. The number of iterations required during hyper-parameter tuning depends on the number of hyper-parameters considered i and the number of their corresponding settings j. Through this process, the best performing hyper-parameters for each model are estimated. All model implementations use the *scikit-learn* Python package (Pedregosa et al., 2011). During testing, data originate from a single participant each time, therefore the *agreement tolerance threshold* parameter (a_t) is ignored for the test set and all data entries are used for that test participant; while memory and uncertainty threshold parameters are applied as normal. Additionally, considering the stochasticity in Random Forests, experiments with leave-one-subject-out were repeated 10 times, as can be illustrated in Figure 5.4.

5.2.4.2 | Parameter Tuning for Affect Prediction

The first thing to notice from Tables 5.6 and 5.7 is that higher agreement thresholds result 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 consecutive 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 for the specified affect metric.

Binomial testing (Cramer, 2003) is used to establish statistical significance at p-value < 0.05, with the hypothesis that the observed prediction is significantly different from chance. Tests show that test accuracies in Tables 5.6 and 5.7 are significantly above the 50% baseline (since the dataset is balanced) 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, 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

Table 5.6: Test accuracies (%) for arousal modeling, bold highlights single highest scores per affect treatment. Accuracies (and training dataset sizes in parentheses) are averaged from 36 leave-one-participant-out experiments.

		Memory (m)			
a_t	ϵ	1	3	5	∞
Arousal Mean					
0%	0	55% (4K)	54% (16.3K)	55% (35K)	54% (61K)
	0.05	56% (3.5K)	55% (13.8K)	55% (28K)	55% (52K)
	0.1	55% (2.5K)	56% (11K)	54% (23K)	55% (43K)
50%	0	55% (723)	54% (2K)	55% (3K)	49% (4.3K)
	0.05	55% (736)	55% (1.9K)	55% (3K)	53% (4.2K)
	0.1	56% (702)	55% (1.9K)	54% (3K)	55% (4.1K)
66%	0	54% (338)	54% (1K)	53% (1.6K)	53% (2.2K)
	0.05	53% (335)	52% (1K)	55% (1.6K)	54% (2K)
	0.1	54% (297)	55% (1K)	55% (1.1K)	55% (2K)
75%	0	53% (150)	55% (513)	54% (1K)	53% (1.3K)
	0.05	55% (130)	53% (511)	54% (1K)	52% (1.1K)
	0.1	53% (101)	55% (476)	55% (1K)	55% (1.1K)
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)
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)

Table 5.7: Test accuracies (%) for pleasure modeling, bold highlights single highest scores per affect treatment. Accuracies (and training dataset sizes in parentheses) are averaged from 34 leave-one-participant-out experiments.

		Memory (m)			
a_t	ϵ	1	3	5	∞
Pleasure Mean					
0%	0	57% (4K)	56% (15.4K)	55% (32K)	54% (58K)
	0.05	57% (3.5K)	54% (13.3K)	56% (27.7K)	55% (50K)
	0.1	58% (2K)	54% (10K)	56% (21.6K)	57% (41K)
50%	0	56% (772)	57% (2K)	56% (3.3K)	54% (4.6K)
	0.05	58% (706)	57% (2K)	55% (3.1K)	55% (4.4K)
	0.1	58% (682)	58% (1.9K)	58% (3.1K)	55% (4.3K)
66%	0	55% (347)	57% (1.1K)	54% (2K)	53% (2.7K)
	0.05	58% (340)	58% (1.1K)	57% (2K)	56% (2.7K)
	0.1	58% (301)	58% (1K)	58% (2K)	57% (2.7K)
75%	0	57% (182)	56% (600)	56% (1K)	56% (1.7k)
	0.05	57% (175)	58% (600)	57% (1K)	57% (1.8K)
	0.1	58% (120)	58% (588)	58% (1K)	58% (1.8K)
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)
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)

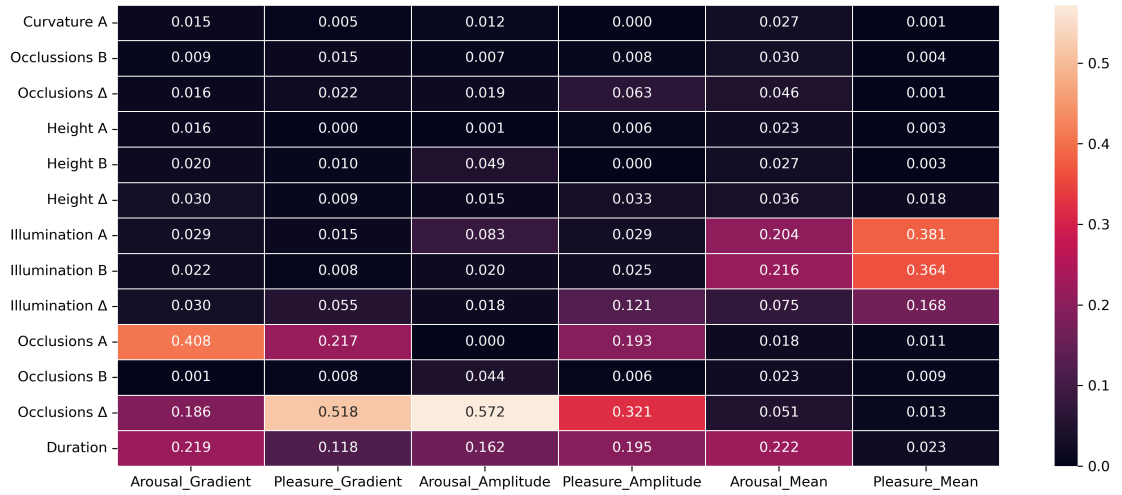

Figure 5.5: Spatial features and their importance of RF predictors for all 3 affect measures of Arousal and Pleasure, calculated based on mean decrease in impurity.

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 $a_t = 50\%$ (263) than for $a_t = 66\%$ (57), the former is preferred and treated 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 $a_t = 50\%$ with the same m and ϵ (65.9% accuracy), but on a far larger dataset (263 data points versus 57 for the best model). Lastly, the room mean affect measure for both affect labels yields the poorest accuracy scores, centered around 54% for arousal and 57% for pleasure. This finding is in line with literature evidence on the relative treatments of affect measures.

As a next step, the Feature Importance analysis reveals which design features contributed the most to this prediction task for both affect labels and measures. Additionally, since the proposed processing method is expected to lead to more robust ground truth data, for the sake of a larger dataset, the parameters with $\epsilon = 0.1$, $m = 1$ and $a_t = 50\%$ are used for both pleasure and arousal as "best" for the following analysis on comparing Random Forests to SVMs and Naive Bayes classifiers.

5.2.4.3 | Feature Importance

Figure 5.5 depicts importance in input features for the Random Forest (based on mean decrease in impurity within each tree) for the chosen processing parameters of $\epsilon = 0.1$, $m = 1$ and $a_t = 50\%$ for all affect metrics and labels. The first indication is that occlu-

sion plays a major role for both affect labels and metrics. Considering that the gradient accounts for how often a user changes their arousal or pleasure annotation within the same room and amplitude represents 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 in gradient and amplitude, but for the room mean measure, it displays the highest importance, which aligns with existing theories regarding the impact of illumination color in digital games (Niedenthal, 2009). Height as a predictor contributes mildly to both Arousal models, while curvature seems to contribute more to both Pleasure models. Both findings align with theories regarding the impact of scale and arousal or awe (Heath et al., 2000) and curvature or non-linearity on pleasure (Ruta et al., 2019; Ruta et al., 2023). Finally, the duration of the navigation in both rooms is an important feature for both models, indicating that factors such as recorded viewing behavior and navigation pace play an essential role in the manner the surrounding environment is perceived and annotated by the participants.

5.2.4.4 | Comparison between Predictive Models

Through the extensive validation process of Section 5.2.4, it was shown that affect gradient and amplitude as measures are more easily predicted than the mean measures when looking at the impact of spatial features of architectural spaces. This section documents an additional analysis on the performance of different classifiers using the datasets with the highest accuracy and sufficient dataset size from Tables 5.6 and 5.7, i.e. with $a_t = 50\%$, $\epsilon = 0.1$ and $m = 1$.

Table 5.8 compares the performance of the different models in predicting arousal and pleasure gradients, amplitudes, and mean changes between rooms. Beyond accuracy, precision (true positives versus all positives), recall (true positive results versus all samples that should have been identified as positive), and F1 score (harmonic mean of precision and recall) are included as established measures for classification tasks. Initially, it can be observed that for the best-performing models, gradient accuracies range around 68% to 70%, for amplitude from 60% to 63%, and mean from 56% to 59%. Upon closer examination of the results, the non-linear SVM does not perform as well as the other three methods in any of the metrics (except perhaps precision), indicating that it tends to predict more false negatives. Linear SVMs seem to perform on par with RFs and NB, performing better for pleasure gradient than RFs but worse than NB and having better recall for both affect gradients. NB models have the highest accuracy on

Table 5.8: Classification performance for the best affect datasets (pleasure gradient and arousal gradient) for RFs and SVMs. Results are averaged from leave-one-subject-out runs and 95% confidence intervals are included. The tuned hyperparameters for these models are shown for each model.

Model	Accuracy	F1 Score	Precision	Recall	Hyperparams
Arousal gradient					
RF	69%±2.6	68%±2.8	67%±2.6	70%±3.9	{100, 3, 6, 6, T}
Lin. SVM	68%±3.0	68%±3.0	**65%±2.5**	**72%±4.3**	{1}
RBF SVM	59%±3.3	39%±6.0	66%±6.0	29%±5.0	{100, 1}
NB	68%±3.3	66%±3.0	67%±3.0	**66%±5.0**	{89}
Pleasure gradient					
RF	66%±5.5	65%±6.0	67%±6.0	63%±6.0	{50, 5, 10, 10, T}
Lin. SVM	69%±4.5	66%±6.0	65%±6.0	**67%±6.0**	{9}
RBF SVM	**60%±4.1**	40%±6.0	67%±8.0	29%±5.0	{1, 1}
NB	70%±4.6	64%±6.0	**71%±6.0**	**59%±6.0**	{1}
Arousal amplitude					
RF	61%±2.5	55%±3.0	61%±4.0	52%±4.0	{50, 3, 2, 2, T}
Lin. SVM	62%±2.5	**51%±3.0**	**67%±4.0**	**43%±4.0**	{3}
RBF SVM	**54%±2.0**	29%±4.0	56%±6.0	21%±3.0	{1, 10}
NB	63%±2.3	**50%±3.2**	69%±4.2	40%±3.2	{56}
Pleasure amplitude					
RF	60%±5.1	50%±6.0	64%±7.0	41%±5.0	{50, 7, 2, 2, T}
Lin. SVM	58%±3.9	**57%±4.0**	**57%±5.0**	58%±5.0	{5}
RBF SVM	**55%±3.7**	29%±5.0	**53%±9.0**	21%±4.0	{1, 0.1}
NB	63%±4.3	49%±6.0	66%±7.0	**40%±5.0**	{56}
Arousal mean					
RF	56%±.05	56%±5.0	56%±5.0	56%±5.0	{100, 5, 6, 1, T}
Lin. SVM	54%±.03	54%±3.0	54%±3.0	54%±3.0	{1}
RBF SVM	**51%±.03**	**51%±3.0**	**51%±3.0**	**51%±3.0**	{100, 10}
NB	54%±.03	54%±3.0	54%±3.0	54%±3.0	{1}
Pleasure mean					
RF	58%±5.0	58%±5.0	59%±5.0	58%±6.0	{100, 3, 10, 1, T}
Lin. SVM	54%±5.0	53%±4.0	53%±4.0	54%±4.0	{2}
RBF SVM	**54%±4.0**	54%±4.0	**54%±4.0**	54%±4.0	{1, 0.1}
NB	**53%±4.0**	**53%±4.0**	**53%±4.0**	53%±4.0	{1}

pleasure gradient and amplitude but under-performs in terms of the remaining classification metrics compared to RFs and linear SVM overall. Another point to note here is that Mean as an affect measure still under-performs when compared to Gradient and Amplitude measures for all the classification methods.

The three suggested modeling techniques have been tested to see if they display

significant ranking differences between their corresponding folds with the proposed Random Forest technique, regardless of direction (greater or less). For this, a Wilcoxon-signed rank test was carried out with a p-value set at .05. Notable effect sizes at the p-value of .05 are highlighted in bold in Table 5.8. The test results highlight the differences present within the sample ranks themselves. Additionally, the Wilcoxon-signed rank test shows that even though some modeling results may display subpar performances, the underlying ranking of the annotator folds displays noticeable differences with the chosen RF classifier.

5.3 | Summary

Comparing the results from non-experts through the crowd-sourcing experiment (see Table 5.4) with the ones from the expert study (see Tables 5.2 and 5.3) for the Arousal label, similar trends are displayed throughout the chosen features. Looking at the mean metric, occluded settings seemed to have a greater impact on the experts than the non-experts, where a significant difference here is displayed in concordant trends, while Curvature and Color warmth follow pretty similar concordant trends; 55% to 69% for non-experts and 75% to 87% for experts in Curved rooms and 65% to 70% for non-experts and 59% for experts in larger rooms. Height did not display significant trends among experts but displayed a mild discrepancy among non-experts when Agreement tolerance is set agreement 66%, but not persistent when (a_t) is set at 75%, suggesting a minor increase in Arousal for smaller rooms. Additionally, relative measures of affect (Amplitude and Gradient) displayed stronger agreements trends than the absolute measure of affect mean, suggesting that under these conditions of continuous affect annotations these measures might be more appropriate. Finally, temporal biases could be minimized with the memory setting, as this is displayed with the consistent higher accuracy results of the RF classifier for shorter memory settings.

The first two studies of this dissertation explored spatial walk-through videos as a stimulus to simulate the experience of navigating continuously changing environments. Two types of raters were employed—experts and non-experts—to capture continuous affect ratings, focusing on the longitudinal effects of space and the influence of spatial parameters on these ratings. The analysis utilized two methodologies: a linear approach for the pilot study with expert raters and a non-linear approach, involving training classifiers with methods across studies in AC, to process the non-expert data. A key finding was the importance of relative analysis in understanding emotional states, as both approaches underscored the subjective nature of emotional assessments. Three types of

signal measures were used for a comprehensive comparison looking at their ability to predict and inform on the impact of spatial parameters. The studies culminated in the creation of two affect datasets, providing valuable resources for future research in the intersection of continuous affect ratings and spatial experience.

Moving forward, the next chapter builds upon this foundation, introducing the results and methods of a study that took place in a lab setting and employed an interactive virtual environment as a new stimulus. Ratings of affect were collected during exploration studying the cognitive impact of "real-time first person annotation." Additionally, the study assessed how varying levels of immersion can influence cognitive load and, consequently, affect the nature of the ratings collected. This shift in stimulus aims to deepen the understanding of how interactive and immersive experiences shape emotional responses.

6

Annotation within Virtual Environments

In Section 4.4 we describe the aim and data collection process of the *AffroomsMR* study. This chapter presents the results of this study where moment-to-moment affect annotations are gathered while exploring and visually interacting with virtual environments. We define this methodology as *"First-person annotation of active stimuli"*. The aim of this study is the collection of pleasure ratings during the process of exploration and interacting with a VE. The pleasure dimension is used as it is a crucial topic in architectural design as emphasized by Hildebrand, 1999 and Lomas et al., 2022. More specifically the parameters of space that deliver positive experiences of comfort and well-being. The spatial parameters we study follow the same feature settings for Contour shape, Scale and Occlusions as in our previous studies (see chapter 5) while Ambient Illumination is explored with two color settings of warm and cold, thus resulting in 16 different synthetic rooms. The final appearance of these rooms are depicted in Figure 6.1.

Additionally, two different methods of display are evaluated and compared with the following hypothesis: *"There are significant differences between media of display (desktop or VR) in terms of usability, distraction and presence, during the task of collecting affect annotations"*. These aspects of user experience are addressed via post session surveys and tested for significant differences attributed to each mode of display.

6.1 | Post-session Evaluation

All items forming the post-session questionnaires use a 5-point Likert scale and anchored at their end points from "strongly agree" to "strongly disagree", with some exceptions where the end points were altered to fit the study. To eliminate scale bias re-

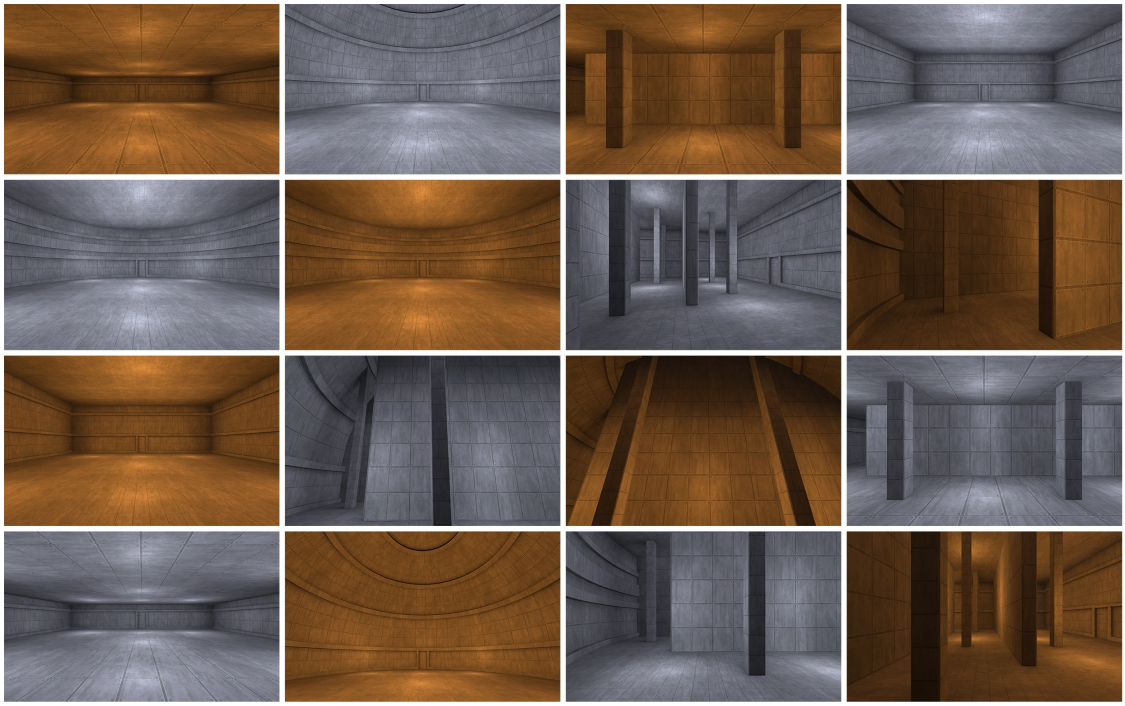

Figure 6.1: Views of the 16 rooms, showcasing different design parameters (light color, occlusions, contour curvature and room height)

garding these endpoints, 3 items have been flipped (i.e. disagreement being a positive response in terms of the measured experience construct). All items on the questionnaire were taken from 3 different studies looking into important areas of designing Immersive Virtual worlds, as described in Section 2.3. Usability of the annotation system is assessed with 7 items (Q1 to Q7) adopted from the Post-Study System Usability Questionnaire (Lewis, 1995) and were adjusted to fit the contents of the system. From the Presence Questionnaire (Witmer and Singer, 1998) we use 3 items addressing distraction and interaction factors with the medium (Q8 to Q10). Lastly, presence is measured with 5 items (Q11 to Q15) from the SUS Questionnaire from Slater et al., 1999.

6.2 | Participants

The experiments for the study took place in two different labs, the Institute of Digital Games (University of Malta) and the Microsoft Computer Games and Emerging Technologies Research Lab (GET Lab) at the Department of Multimedia and Graphic Arts (Cyprus University of Technology), during the period of March and April 2022 (see Fig-

Figure 6.2: Participant setup during VR and Desktop sessions

ure 6.2). Thirty-five (35) participants expressed their interest to be involved in the experiment (15 from University of Malta and 20 from Cyprus University of Technology). Out of the 35 participants, 2 were used as pilot tests and their participation contributed to preparing the study for the remaining subjects. Of the remaining 33, 2 more participants were excluded due to incomplete data and inconsistencies detected during their sessions. The results we report are from the remaining 31 participants (10 female and 21 male). Participant ages ranged from 18 to 45 years old, with the majority between 18 and 25 years old. Almost all participants were involved with either institution: 21 were students and 7 were lecturers or professors. The selected sample self-reported high interaction frequency and familiarity with technology and desktop environments (mean 4 out of a 5-point Likert scale). Familiarity with VR displays was ranked low (mean 1.8 out of 5), with only 2 out of 31 rating themselves as high (5 out of 5) in familiarity with VR displays, and 15 out of 31 as low (1 out of 5). Regarding the order of shown media, 14 participants were shown VR first, desktop second, while 17 were shown the opposite order. This was done to further test if there is impact that could be attributed to medium order.

6.3 | Survey Results

As our general processing approach is the treatment of affect in an ordinal manner, we follow the same approach for the Likert scales of our survey items. Yannakakis et al., 2018 and Yannakakis and Martinez, 2015 highlight the advantages of retaining the ordinal nature of Likert scales in achieving more reliable results. As we follow the same approach we analyze all questionnaire items in both an ordinal approach using

Table 6.1: Survey Response counts, Significant differences based on positive-negative counts between VR and Desktop are marked with (*). Preference score marks the ordinal relationship between the two media and Significant scores are highlighted in bold.

	Question	Pos. VR	Neg. VR	Neu. VR	Pos. DES	Neg. DES	Neu. DES	P_score
PSSUQ Usability	Q1. Overall, I am satisfied with how easy it was to annotate while in the environment.(*)	27	1	3	23	2	6	0.2
	Q2. It was simple to use the annotation system.	28	0	3	29	1	1	0.0
	Q3. I could effectively complete the tasks and scenarios with the present environment.	26	1	4	24	1	6	-0.1
	Q4. I was able to complete the tasks and scenarios quickly using the present environment.	27	0	4	26	0	5	-0.1
	Q5. I was able to efficiently complete the tasks and scenarios using the environment.	28	1	2	28	0	3	0.1
	Q6. I felt comfortable using this environment.	23	3	5	25	3	3	-0.1
	Q7. It was easy to learn to use the annotation system.	29	0	2	27	0	4	0.0
W&S Distraction	Q8. To what extent did the visual display quality interfere or distract you from performing assigned tasks or required activities?(flipped)(*)	14	13	4	22	4	5	**0.3**
	Q9. To what extent did the control devices interfere with the performance of assigned tasks or with other activities?(flipped)(*)	19	9	3	21	3	7	0.3
	Q10. How well could you concentrate on the assigned tasks or required activities rather than on the mechanisms used to perform those tasks or activities?	23	3	5	19	5	7	0.3
SUS Presence	Q11. I had a sense of being there during the experience.(*)	27	0	4	15	11	5	**0.9**
	Q12. There were times during the experience when the virtual world was the reality for me.(*)	18	6	7	10	17	4	**0.8**
	Q13. Thinking back at the experience you just had, do you think that the buildings were images you just saw or more somewhere that you visited?(*)	16	3	12	8	16	7	**0.8**
	Q14. During the time of the experience, which was strongest overall, your sense of being in the virtual environment, or of being in the real world of the laboratory? (flipped)(*)	20	6	5	10	16	5	**0.6**
	Q15. During the time of the experience, did you often think to yourself that you were just sitting in a laboratory or did the experience overwhelm you?(*)	20	5	6	9	15	7	**0.8**

a *preference score*, adopted from Lopes et al., 2017a; and a nominal approach whereby all items are categorized as positive (above Neutral response 3) and negative (below Neutral response 3). The survey results are depicted in Table 6.1. The preference score formula for each item is calculated by *Eq.* 6.1 with Pi as the item's preference score given the participant's responses. For each participant i, z_i is equal to +1 if the item was rated higher in the VR post session questionnaire and -1 if the that item was rated higher for the Desktop version. Participant item ratings that are the same for both versions are counted as 0 and do not contribute to the preference score, thus N represents the number of participants with either VR or Desktop being favored. The preference score P_i ranges are -1.0,1.0 with -1.0 representing complete item preference for desktop and 1.0 complete VR item preference for all participants. For flipped items, the focus on

positive and negative responses in terms of the construct itself, i.e., the reverse of the item's responses and thus Likert scores of 1 or 2 are labeled positive and 4 and 5 as negative. To assess significant differences between VR and desktop we compare the two media via *Wilcoxon signed rank test* with significance set at $p<0.05$.

$$P_i = \left(\sum_{i}^{N} z_i \right) / N \tag{6.1}$$

For the PSSUQ Usability Questionnaire (Q1-Q7), all 7 items display similarly positive tendencies for both setups, with 23 to 29 positive ratings out of 31 total responses. Q7 received the highest Likert score, suggesting that the annotation system was generally perceived as easy to understand and use in both media. There is a minor difference between media, with Q1 showing the only significantly higher scores in favor of VR settings on overall user satisfaction. For the W&S items (Q8-Q10) measuring distraction, 2 of 3 items (Q8, Q9) show significantly higher positive responses for desktop sessions compared to VR, meaning that VR introduces more interference from the medium in the annotation task. These responses could be attributed to participants' lack of familiarity with HMDs or the use of a mouse for the annotations in VR. For Q10, rating the ease of concentrating on the annotation task, VR was rated slightly higher than desktop, but no significant differences were recorded here. This slight favor for VR could be explained by its potential to encourage spatial exploration via the (HMD), suggesting a sensorimotor dependency. For the SUS presence questionnaire (Q11-Q15), VR sessions were rated significantly higher than the desktop on all items. As expected, VR offers a better experience of presence, with 27 of 31 participants agreeing that they had a sense of being there during the experience (Q11).

A subsequent test was carried out to determine if there is an impact on responses based on the order of presentation of each medium. The Mann-Whitney U test was implemented with the hypothesis that there is a significant difference between the two groups. Tests revealed that a few items were significantly different between the VR first and the desktop first group ($p<0.05$): for VR 3 out of 15 items (Q1, Q7 and Q9), and 2 for desktop (Q1 and Q7). Since only a few items showed significant differences, it cannot be assumed with certainty that medium order had an impact on raters feedback.

6.4 | Agreement Analysis

Based on the participant's acquired data (i.e., pleasure annotations and viewing behavior), an analysis is carried out to study how these map to the design parameters of the

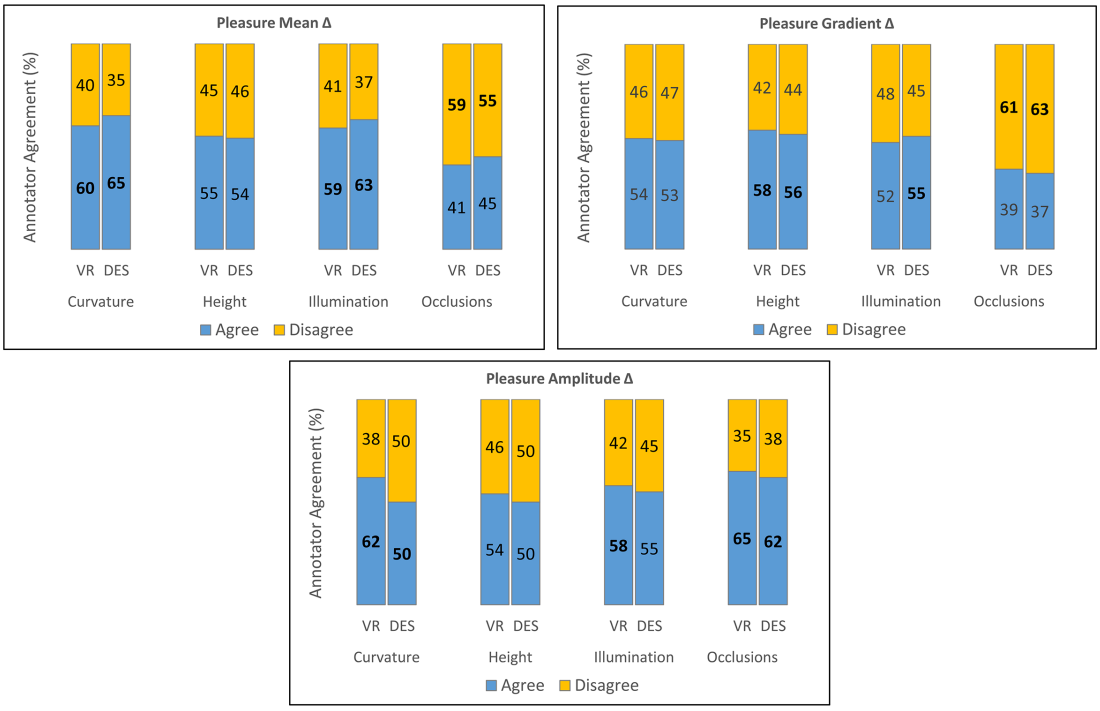

Figure 6.3: Pleasure Mean, Amplitude and Gradient agreements with design parameters for VR & Desktop, bold values denote statistical significance.

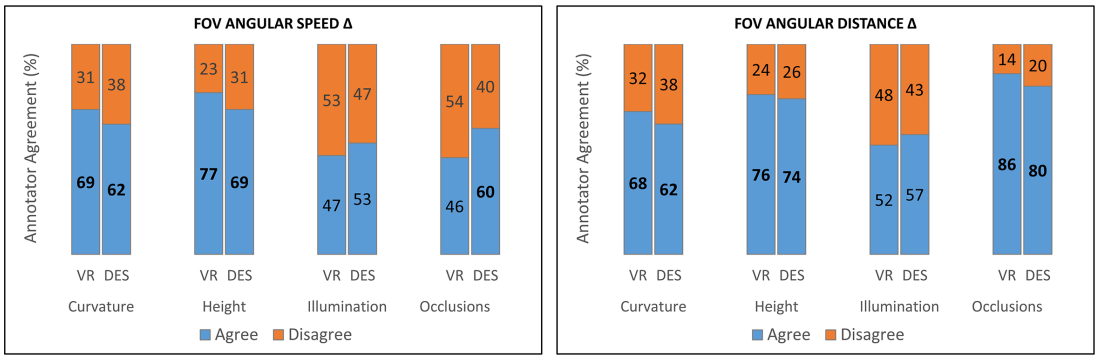

Figure 6.4: Field of View Angular Distance and Speed agreements with Design parameters for VR & Desktop, bold values denote statistical significance.

individual rooms that the rater is in. As described in Chapter 4, the continuous pleasure ratings are acquired by embedding the *RankTrace* annotation protocol within the VE. Following the same processing approach as in Affrooms24 and Affrooms12 studies, the continuous pleasure signal is split into room time windows. Figure 6.5 depicts participant sessions of compartmentalized signals based on the timestamp when the user enters and exits a room for both VR and desktop media displays. Pleasure ratings and FOV angular metrics (pitch and yaw angles) are normalized via [0-1] MinMax Normalization for each rater. For each room in a sequence, we extract the three metrics of mean, gradient and amplitude on the affect signal, while for the viewing behavior, the FOV's mean angular speed and FOV's mean angular distance is computed similarly. As all metrics are processed relatively, we compare the *changes* in affect metrics (and the FOV metrics) between consecutive rooms with the respective design feature changes in these rooms. Consecutively, all aforementioned metrics are processed as differences (or *deltas*), since these convey the metric's comparison between a *previous* and *next* state. Results are summarized in Figure 6.3 and Figure 6.4 for both media, across all four design parameter categories. Significance is based on the binomial distribution of all affect and FOV changes, when the spatial feature changes, assuming a 50% probability that the changes may be in agreement. Significance is established at 95% confidence.

In Figure 6.3, the mean pleasure difference results show significant agreements for both VR and desktop in contour curvature and ambient illumination. This captured tendency for the selected sample indicates a mild preference towards curved contours and colder illumination scenarios, aligning with results of previous studies (Gómez-Puerto et al., 2016; Gómez-Puerto et al., 2018). This finding is mainly displayed for the mean metric and amplitude metrics but not for gradient. Regarding the presence of occluding elements of walls and columns, results show a decrease of pleasure ratings, slightly emphasized in VR for the mean metric, which could be explained by discomfort of enclosed spaces, particularly noticeable in VR settings due to higher presence effect. Additionally both gradient and amplitude metrics display significance here in pleasure rate of decrease and variance. Height on the other hand did not display any significant effect for pleasure ratings in mean and amplitude metrics but mild tendencies were uncovered for the gradient metric (increased pleasure). This could be attributed to the relative nature of the gradient metric, as pointed out by Camilleri et al., 2017b, being able to capture changes at the frame level, in contrast to the absolute measure of *mean*.

Results regarding participants' viewing behavior (see Figure 6.4) show consistent agreements for both media. Especially for occlusion and height, participants move their heads more and with greater speed in high-ceiling rooms or in rooms with occlusion. Thus, users tend to explore visually more in such rooms compared to empty rooms

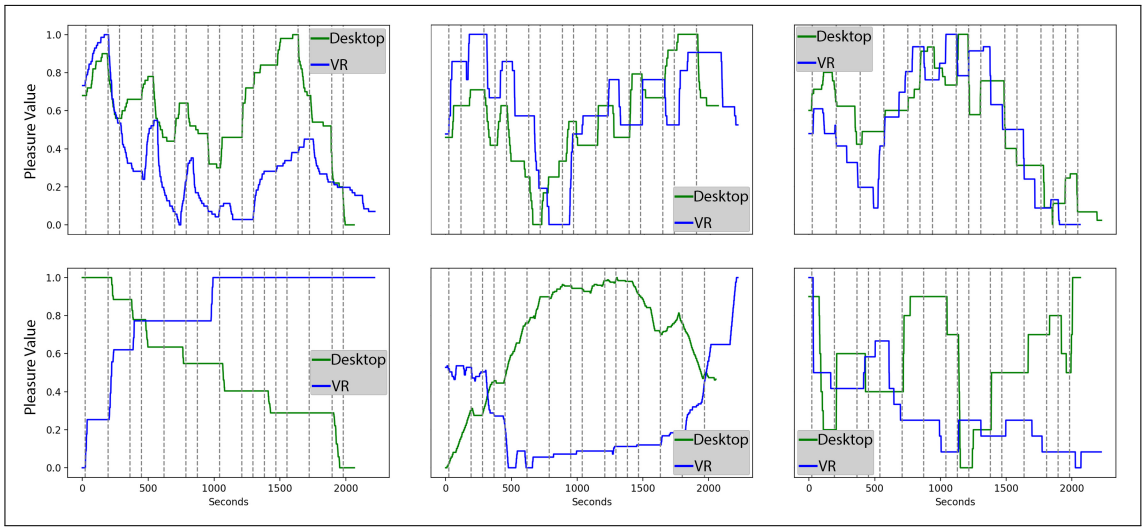

Figure 6.5: VR and Desktop Pleasure annotations for 6 participant sessions. SDA rankings of ordinal similarity between traces, Top row: best 3 sessions, Bottom row: 3 worst sessions. Dashed lines represent the room's limits.

or rooms with lower ceilings. This viewing behavior is understandable for both features considering a transition from a lower ceiling room to a high ceiling (Meyers-Levy and Zhu, 2007) or from an empty room to one that has occluding elements. Moreover, rooms with curved contours displayed the same tendency of increased FOV distance and speed. Lastly, ambient illumination settings did not yield any significant effect on participant's viewing behavior.

6.5 | Agreement across Displays

An in-depth descriptive analysis on the annotated pleasure traces and viewing behavior (via the FOV's angular distance measure) of both media display, was carried out to estimate the intra-rater agreement across these two media. The primary objective was to estimate the intra-rater agreement, irrespective of the medium of representation. This analysis would help in comprehending the impact of spatial conditions in virtual walkthroughs.

Studying the results of the survey in Section 6.3, there are noticeable differences between the two media (Desktop and VR), in aspects of presence, immersion, and the embedded annotation mechanisms. This analysis explores the rater's agreement between acquired signals of affect ratings and viewing behavior across the two displays. For this we adopt a reliability method seeking similarities through ordinal comparisons.

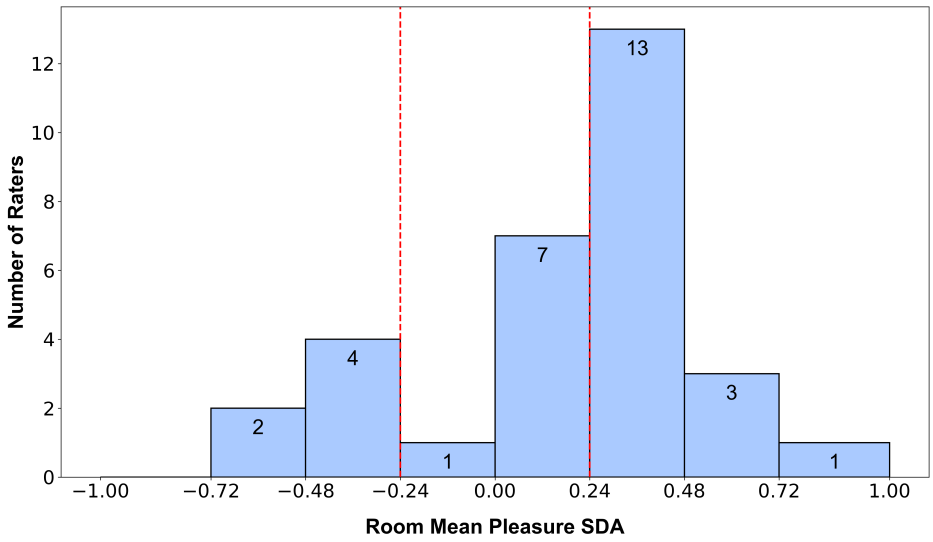

Figure 6.6: Ranking Raters using Signed Differential Agreement between their respective Pleasure traces across VR and Desktop sessions. Vertical red lines indicate binomial significance threshold.

The SDA measure we described in Section 3.3 is a method to quantify rater agreement between continuous annotations and a known ground truth. We use this method to compare the acquired signals from the two media display. This is achieved by looking at state x_t and previous state x_{t-1} from signal a (VR display) and y_t and its corresponding previous state y_{t-1} for signal b (Desktop display) and compare if the direction of change for the chosen affect measure is the same. The SDA measure (Equation 3.4) aggregates these delta comparisons, assigning equal weight to each pair of annotation comparisons. assigning the same weight for each annotation comparisons and its output ranges between [-1,1]. As such 1 indicates complete intra-rater agreement across the two media, -1 denotes complete disagreement and 0 represents no relationship, indicating that the two annotation series are uncorrelated.

Figure 6.6 displays the distribution of raters according to their SDA score for the Room Mean Pleasure measure. For the two different media of display, the selected sample had 4 annotators displaying strong agreement between their ratings of affect (above 0.6), 6 raters displayed mild to moderate disagreement (above 0.24) and 20 displayed weak to moderate agreement, while 8 raters showed weak to no agreement (absolute SDA below 0.24). Lastly, 7 raters displayed no relationship between their rating of affect for these two display methods. Significance was calculated using binomial testing on

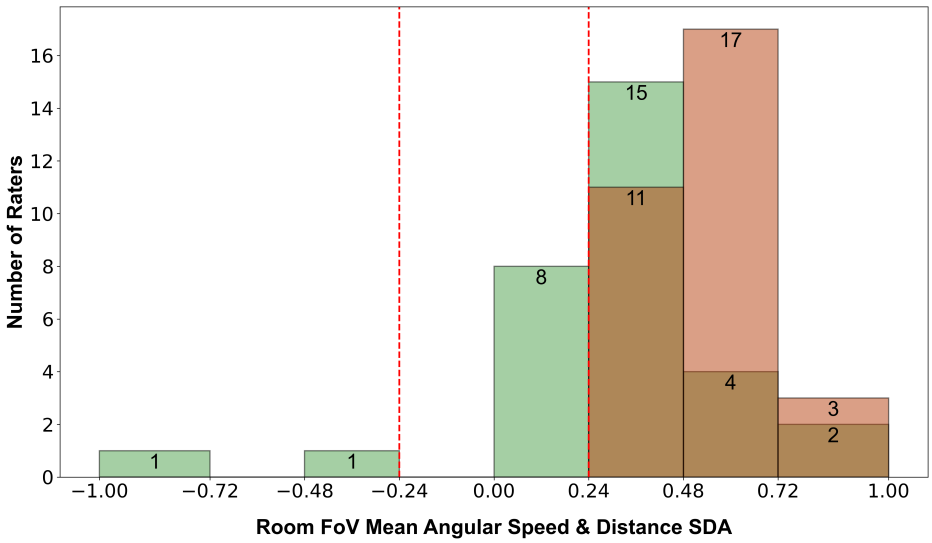

Figure 6.7: Ranking Raters using Signed Differential Agreement between their respective room mean FoV angular speed and distance across VR and Desktop sessions. Vertical red lines indicate binomial significance threshold. Green bins mark FOV distance, red bins mark FOV speed.

each annotator looking at agreement and disagreement counts for p-value of 0.05.

Figure 6.7 illustrates the distribution of raters SDA scores for room mean Field of view angular distance measure. In comparison with the Room mean pleasure measure, viewing behavior yields much higher agreement between the two media, with 21 annotators placed above moderate agreement (0.25 and above) and only 2 out of the 31 annotators revealing disagreements (1 annotator showing almost complete disagreement with -0.94); and 8 annotators revealing mild to no relationships (below 0.23). Additionally for this analysis the room mean FOV speed is included showing complete agreement among raters with everyone above the significance threshold.

This descriptive analysis reveals that regardless of the medium (HMD or Desktop display), there is a considerate amount of raters that are consistent with both their pleasure annotations and viewing behavior. Around 55% of the raters (17 out of 31) display above mild agreement within their pleasure ratings and 68% for FOV change behavior. These results highlight an importance of the parameters of space that is worth considering. Future works should consider the above findings when capturing continuous measurements regarding various spatial conditions across different viewing displays.

Table 6.2: Test accuracies (%) for pleasure modeling, underline highlights single highest scores per affect treatment. Test baseline is 50%, Accuracies are averaged from leave-one-subject-out experiments. Accuracies with * did not pass the binomial significance testing at p-level=0.05.

		VR Display			
M	a_t	**1**	**3**	**5**	∞
(μ)	0	60% (819)	59% (2.3K)	55% (3.6K)	53% (6.7K)
	.05	60% (616)	58% (1.9K)	55% (3.1K)	54% (6.1K)
	.10	**67%** (450)	63% (1.5K)	54% (2.6K)	54% (5.4K)
(A)	0	57% (752)	59% (2.1K)	55% (3.3K)	58% (6.1K)
	.05	57% (680)	59% (1.9K)	59% (2.9K)	**60%** (5.5K)
	.10	56% (597)	**60%** (1.7K)	59% (2.6K)	59% (5.0K)
(∇)	0	64% (792)	61% (2.2K)	63% (3.4K)	62% (6.4K)
	.05	63% (726)	**67%** (2.1K)	62% (3.2K)	63% (6.0K)
	.10	**66%** (686)	64% (1.9K)	63% (3.0K)	**64%** (5.6K)
		Desktop Display			
M	a_t	**1**	**3**	**5**	∞
(μ)	0	61% (831)	59% (2.4K)	58% (3.7K)	52% (6.9K)*
	.05	60% (601)	57% (1.9K)	59% (3.1K)	53% (6.1K)*
	.10	59% (438)	56% (1.5K)	59% (2.6K)	54% (5.2K)
(A)	0	57% (738)	54% (2.1K)	53% (3.2K)*	54% (6K)
	.05	56% (690)	55% (1.9K)	55% (2.9K)	54% (5.5K)
	.10	57% (609)	56% (1.7K)	55% (2.6K)	56% (5.0K)
(∇)	0	61% (788)	61% (2.2K)	61% (3.4K)	60% (6.4K)
	.05	58% (734)	62% (2.0K)	60% (3.2K)	60% (5.9K)
	.10	**64%** (688)	62% (1.9K)	61% (3.0K)	60% (5.6K)

6.6 | Modeling Task

Following the same process as in Chapter 5, we train here a Random Forest classifier and develop models of pleasure based on the annotations gathered for the VR and Desktop display sessions. We follow the same Leave-One-Subject-Out cross validation (LOSOCV) as described in Section 5.2.4.1. For this sample of 31 participants, in each fold we keep one participant for Hyper-parameter tuning while the remaining 30 undergo a Train-Test, keeping one participant as a test sample at a time. This process is repeated until all participant data have been used in the Hyper-parameter tuning step. All the above are repeated $n = 3$ times to minimize the classifier's stochastic nature.

The classification performance of the Random Forest model was evaluated using accuracy as the evaluation metric across the various conditions. The conditions were defined by the three affect measures of mean (μ), amplitude (A), and gradient (∇), three

ambiguity thresholds (a_t): 0%, 5%, and 10%, and four memory settings: 1, 3, 5, and ∞ (all-vs-all). The results are shown in Table 6.2. As we discussed in Chapter 3, in this study we do not include an inter-rater agreement threshold as in the Affrooms12 study, as participants did not experience the same order of rooms with each other.

What we initially observe is that overall the VR display sessions consistently show slightly higher accuracies across most conditions compared to the desktop sessions. This trend is most pronounced for memory settings of 1 and 3, suggesting that annotators in VR might produce more consistent ratings for temporally close comparisons. However, for larger memory settings (5 and 11), the accuracies for VR and desktop converge. Memory 1 generally yielded the highest accuracies, supporting the hypothesis that pairwise comparisons between temporally adjacent rooms result in more reliable labels. For example, in the VR display condition under the mean (μ) measure and $a_t = 0\%$, the accuracy at memory 1 was 0.60, dropping to 0.55 at memory 5 and 0.53 at memory 11. This decline indicates that increasing the temporal distance between compared rooms introduces more variability, potentially due to reduced rater recall or increased complexity in their affective judgments.

Introducing ambiguity thresholds of 5% and 10% did not yield a clear improvement in classification accuracy. For example, under the mean (μ) measure in VR, the accuracy for $a_t = 0\%$ was comparable to $a_t = 0.05$ (e.g., memory 1: 0.60 vs. 0.60). This suggests that noise reduction using ambiguity thresholds did not significantly impact model performance, potentially indicating that the dataset's inherent noise was minimal or not strongly influenced by ambiguous annotations. Finally, looking at the three measures of affect, the gradient (∇) generally outperformed the mean (μ) and amplitude (A), particularly at lower memory settings. For instance, in the VR display condition with $a_t = 0\%$ and memory 1, the gradient yielded an accuracy of 0.64, compared to 0.60 for the mean and 0.57 for the amplitude. This suggests that the gradient, which captures dynamic changes in affect, may be more informative for modeling the temporal progression of affective responses. Accuracies of 0.52 and 0.53 were not statistically significant according to binomial significance testing, indicating that the classifier's performance under these conditions was comparable to chance. These results were primarily observed under larger memory settings, where the temporal bias and dataset size likely diluted the quality of pairwise labels.

6.6.1 | Feature Importance

Figure 6.8 shows feature importance across 24 Random Forest models and is based on the mean of the accumulation of the impurity decrease within each tree for each model.

Figure 6.8: Random Forest feature importance for VR and Desktop sessions for all three affect measures. Figure shows results for Uncertainty Thresholds 5% and 10% and short Memory settings of 1 and 3.

Values are categorized by three affect measures (Mean, Amplitude, and Gradient), two uncertainty threshold settings (5% and 10%), and two short memory settings (1 and 3). These models are further divided based on data from VR and Desktop sessions, with 12 models each. The feature importance values highlight the contribution of each of the 12 features to the classification performance.

Occlusions Δ consistently demonstrate higher feature importance compared to other features, particularly in models using the Gradient affect measure across both VR and Desktop sessions. Illumination Δ also shows relatively higher importance, especially for models with Mean and Amplitude affect measures in VR sessions. Other Δ features (e.g. Height Δ, Curvature Δ) contribute moderately but are less impactful than Occlusions Δ and Illumination Δ. Gradient affect measure appears to be more sensitive to the Δ features (e.g. Occlusions Δ, Height Δ, Illumination Δ), showing higher feature importance in these models. Mean and Amplitude measures exhibit lower feature importance values overall, with more even contributions from features, except for occasional peaks like the Illumination Δ feature. Additionally,the overall trends in feature importance appear consistent between the VR and Desktop datasets, but VR models often show slightly higher feature importance values for the 'DELTA' features (e.g. Occlusions Δ, Illumination Δ). Desktop sessions have a more even distribution of feature importance, suggesting less reliance on specific features compared to the VR sessions. As for the

Uncertainty thresholds of 5% and 10% no noticeable differences between the two were detected for this feature importance analysis.

6.7 | Summary

This chapter introduced an approach for exploring a sequence of interior spaces with real-time annotation, emphasizing the temporal nature of environment assessment. The task of spatial exploration occurred by navigating through predetermined paths with 3 degrees-of-freedom (Field Of View control solely). Data from thirty-one (31) participants was used to study annotated pleasure and viewing behavior using two media of display: Desktop vs. Head-mounted Display.

Initially we studied the effect of medium display medium on the task of experiencing synthetic interior forms while continuously reporting affect changes. A post-session survey containing items on usability, distraction and presence revealed the following: (1) significant indications of presence for VR sessions, (2) higher distraction tendencies for VR sessions and (3) positive feedback with no significant differences between the two media in terms of usability. The interference in VR could be attributed to the limited familiarity that participants had with VR displays, the use of a computer mouse instead of a VR-ready controller or potentially device-specific parameters (device weight, comfort, etc.); note that for both experiments we used *HTC Vive Pro 2*[1].

Beyond usability, ease of annotation and pleasure ratings, we analyzed the acquired data from within the Virtual Environment, aimed at comparing participants' viewing behavior against the design parameters. Results between design features and viewing behavior or affect annotation were similar in both media, with users moving their field of view more in high ceiling rooms and rooms with occlusions, while reporting higher valence when moving from a rectilinear to a curved room and from warm illumination to cold illumination. A subsequent analysis of these two measures was performed irrespectively of the medium, suggesting mild to moderate agreements among most raters, and opening discussions on the impact of space and its evaluation across different media display.

Finally, we build models of affect based on the pleasure annotations collected under these two display conditions. The results emphasize the importance of temporal proximity in pairwise comparisons for preference learning tasks in continuous affect annotation. The superior performance at memory 1 suggests that annotators' immediate affective judgments are more reliable than those involving temporally distant com-

[1]https://www.vive.com/us/product/vive-pro2/overview/

parisons. Interestingly, the gradient affect measure produced unexpected results performing better than anticipated. This contrasts with findings from previous models of affect for video stimuli (see Section 5.2.4) where Amplitude measures consistently outperformed Gradient measures. These results suggest that the display modality (passive stimulus versus active stimulus) play a critical role in the annotation task, a difference that should be further investigated. Furthermore, slightly higher accuracies were observed in the VR models compared to the Desktop models. This suggests that immersive environments may enhance annotators' sensitivity to affective changes. Existing literature suggests the effectiveness of immersive displays in inducing intense affective experiences (Chirico et al., 2017; Marín-Morales et al., 2018). However, the specific effects of immersive environments on annotation behavior remain unclear and require further study. Additional data need to be collected to strengthen these findings and provide more significant insights into the impact of display conditions on continuous affect annotation.

In this chapter we outlined the process and outcomes of capturing and modeling affect collected within interactive VEs. In the next chapter we document the final study of the dissertation where we study expressed rather than reported affect, in relation to environmental and game-related parameters, focusing on level design within video games. Interacting with a virtual world designed to extract affect reports is very different from playing a video game designed for entertainment. The following study is aimed at exploring video games as a potential informant for affect in space, and exploit pre-trained models to capture affect states.

7

Annotation of Gameplay

In the previous chapters we focus in methods that follow a forced-response approach, requiring participants to actively annotate and report their level of affect. However, this thesis has yet to explore a significant aspect: the capturing of manifested rather than reported affect within synthetic spaces. In this work we study manifested affect by exploring video games as a potential case of investigation, as we aim to contribute to the existing body of knowledge on the interplay between emotion and space. More specifically, We use an approach of collecting affect responses from videos of gameplay streamed online (i.e. *Let's Play videos*). "Let's play" videos are popular forms of crowd-sourced content where a streamer plays the game while narrating their play-through and interacting in real-time with an audience. We treat such "let's play" videos by popular YouTube streamers as *in-the-wild* affect data, using machine-generated affect annotations, a similar approach to the one introduced by Kollias and Zafeiriou, 2018, and process them in terms of three moment-to-moment multi-modal affect expression modalities: *facial* expression affect labels, *utterance* affect labels (what has been said) and *para-linguistic* affect ratings (how it has been said).

Following our initial affect annotation and stimuli relationship analysis (see Chapter 2) this method is categorized as *3rd-person annotation of active stimuli*, where affect annotations are produced by three pre-trained affect recognition models. The manual annotations and processed affect streams compose the *Outlast Asylum Affect corpus* that features 16 popular YouTube streamers playing the first map of the *Outlast* (Red Barrels, 2013) horror game. As detailed in Chapter 4, we collected all streamer videos from pre-recorded sessions of gameplay that were streamed online via the *YouTube* video streaming platform. Our goal is to identify relationships between game design features and affect manifestations. We do this via linear relationships as agreements between affect changes and spatial or gameplay changes during room transitions in the game map, and

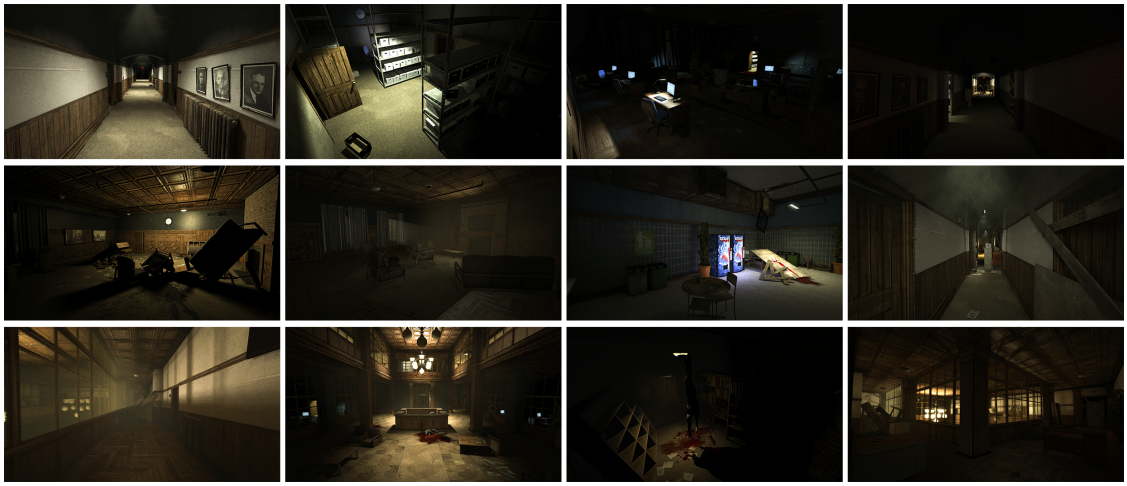

Figure 7.1: Views of 12 rooms of the Outlast Asylum Affect dataset, showcasing different parameters (room height, illumination brightness, interior complexity)

via a Random Forest model trained at predicting affect changes. Results indicate that certain aspects of gameplay define important factors for emotion manifestation. This study seeks to establish a more nuanced understanding of how game spaces can evoke specific emotional experiences, thereby contributing to our understanding of the design of spaces, experiences and games more broadly. A rooms sample of the *OutlastAFF* corpus is illustrated in Figure 7.1.

7.1 | Data Processing

Following the same methodology as with the other studies of the dissertation, we treat the duration where a player is in a specific room of the game level as a single time window and measure the affect manifestations within it. We view affect in each room in terms of its overall *mean value*, its *amplitude* (i.e. the maximum value of this affect signal within the room, minus its minimum value) and its *mean gradient* (i.e. the moment to moment direction of change within each room). As described in Chapter 3, mean affect is an *absolute* measure while amplitude and gradient measures of affect are a *relative* measure which is expected to be less prone to biases and, in our case, averaging artifacts due to different duration per room visit (Lopes et al., 2017b).

We produce these affect measures on the chosen five affect signals of the different manifestations: fear and surprise of facial expressions (F_f and F_s respectively), arousal of voice (V_a), fear and surprise of utterances (U_f and U_s). We wish to match these affect measures with game-room properties. Thus, we track the game and spatial properties

Table 7.1: Outlast Asylum Affect Corpus affect changes for each metric and affect dimension.

Affect Signal	Mean changes	Amplitude changes	Gradient changes
Surprise of Face (F_s)	1214	1292	696
Fear of Face (F_f)	950	802	448
Surprise of Utterance (U_s)	1092	1138	622
Fear of Utterance (U_f)	774	862	430
Arousal of Voice (V_a)	1358	1382	958

of the room at the moment the player entered it. While there are some properties that may change when the player enters the room for the first time (e.g. a cut-scene may play, or the player may pick up the note) and in their second visit these properties have already been triggered (the cut-scene has been played, or the player has already picked up the note), these properties are labeled accordingly.

With the above processing steps, we have each room's conditions and an affect measure for each affect modality. Pairwise transformation is achieved here using a single memory processing (see Section 3.2.1), where all rooms are compared with their adjacent in a single play-through. We classify each room transition as an *increase* (e.g. the mean affect in one room increases compared to the mean affect in the previous room) or as a *decrease*; we discard transitions where there is no discernible change in the affect metric. We only consider changes in the affect metric (increasing or decreasing) if their absolute difference (between rooms) is above a threshold ϵ, similarly done throughout the dissertation. Based on best practices in the literature (Makantasis et al., 2023; Pinitas et al., 2023), we use a threshold $\epsilon = 5\%$ for all affect measures and each signal. With this threshold, the number of changes per signal are somewhat similar for both mean and amplitude between ≈700 for F_f and U_f and ≈1300 for F_s and V_a signals, but for gradient we notice slightly less data indicating that $\epsilon = 5\%$ might be considered somewhat strict for the specific metric (see Table 7.1).

Finally, we aim to match the changing characteristics of the two rooms with increases or decreases in the affect metric. Since all properties in Table 4.2 are scalar, we note whether each property increases or decreases during a room transition. As an example, if a player moves from a brightly lit room with a note to a dark room with an event, the properties "light levels" and "note present" decrease while "event" increases during the room transition. We match these changes in rooms' properties in a linear fashion or as inputs for machine learning to predict affect changes in Section 7.2.

7.2 | Results

As described above, we have collected a dataset of 16 play-throughs of the *Asylum* level from the *Outlast* game. The goal of this study is to assess how room transitions impact the manifestations of emotion captured in para-linguistic, facial, and utterance data (see Chapter 4). We follow two methods to address this goal: one assuming a linear relationship between each property of the room (in Section 7.2.1) and one assuming that all properties combined can predict changes in affect by training a machine learning model (in Section 7.2.2).

7.2.1 | Agreement Analysis

As in the previous studies, we look into linear relationships between a room property changing and an affect metric changing, and note whether these changes *coincide* and are *in agreement*. To do so, we observe cases where both the affect metric increases and a room property has a positive change, and mark this as an agreement; if the affect metric increases and the room property decreases (or vice versa) we mark it as a disagreement. In cases where only one of the two (affect or property) changes, we ignore this room transition altogether. We measure and report agreement ratio as the number of agreements between affect metric and room property aggregated across all 16 play-throughs, divided by the number of agreements and disagreements (i.e. when there was a definitive shift in both gameplay property and affect manifestation). Significance (at $p < 0.05$) is calculated with binomial testing (Cramer, 2003) on this agreement ratio, assuming a 50% chance of agreement or disagreement. Table 7.2 shows the agreement ratio for changes in affect mean, amplitude and gradient, aggregated from all 16 play-throughs in the dataset.

What we can initially observe is that changes in affect amplitude between rooms more often match changes in room features when transitioning from one room to the next. Specifically, when measuring changes in affect mean there are in total 10 significant agreements and 5 significant disagreements; while for changes in affect amplitude there are 44 significant agreements and 17 significant disagreements; as for gradient we observe the least amount of significant changes with 6 agreements and 5 disagreements. This deviation is surprising given the fact that the number of changes for mean and amplitude of the same signal are approximately the same. For mean affect, we observe that the presence of an event in a room coincides with increased arousal in the streamer's voice and fear in the streamer's utterances. This is not surprising, as these events are often jump scares. Interestingly, cut-scenes (which are often also designed to be, arguably,

more elaborate jump scares) have a strong effect on mean affect changes but coincide with a drop in mean fear, likely due to the long duration of cut-scene segments. Regarding spatial features and mean affect changes, we observe that uneven illumination settings (higher light contrast) coincides with increased fear in streamers' utterances, but the reverse is true as lights get colder (increased color temperature). This could partly be explained by the fact that rooms with warmer or 'yellowish' illuminations exhibit higher contrast and uneven lighting conditions, with pronounced shadows and dark corners that developers sought to exploit by incorporating various triggers and jump scares. Conversely, environments with white or 'colder' illumination settings (increased in color temperature), characterized by even light distribution (decreased light contrast), are reserved for other player interactions such as reading notes or using key items acquired throughout the map.

Changes in affect amplitude, as noted earlier, coincide more often with changes between adjacent rooms' features. This is in part due to the way amplitude is computed. Looking at the distance between lowest and highest state within a room, a jump scare would result in a high amplitude as the streamer would momentarily cry out and promptly return to a more neutral state; if traversing to the next room takes more than a few seconds, this temporary increase would not be very pronounced for mean affect (e.g. for V_a) but would be evident in affect amplitude. With this in mind, it is not surprising that both events (usually jump scares) and cut-scenes result in increased amplitude for all affect signals. A more interesting finding is that the presence of notes or batteries also triggers increased amplitude across all signals. This is also not surprising considering the "let's play" live commentary medium format. When a note is found, it is immediately read out loud by the streamer: the text of the note is often creepy, which is then captured in the streamers' utterances as they read it, but also on their expressions. Moreover, reading a note offers the streamer a pause from the game (as usually no-one is chasing them during those times) and allows the streamer to engage with the audience and discuss the game, leading to more cadence in voice and expressions. Furthermore, batteries and hiding places, which are essential to the player's in-game survival, are often placed near hostile NPCs. Therefore, finding a battery or a hiding place may signal to the streamer that danger is nearby, increasing anticipation registered via affect manifestations.

Far more spatial features seem to have an impact on affect amplitude than on affect mean and gradient, most notably the presence of an empty room or a blocked path and increased light levels (leading to decreased affect amplitude for all signals, except F_f for blocked path) and an increase in light color temperature (from warm to cold lights) which leads to increased amplitude on all affect signals. While these are interesting ob-

Table 7.2: Agreement ratio between spatial (top features) and in-game properties (bottom features) of the game level and affect mean, amplitude and gradient, for different affect manifestations. We mark agreements as △ and disagreements as ▽ when statistically significant (above chance).

	Feature	F_f	F_s	V_a	U_f	U_s
Changes in Affect Mean	Area size	50%	56%△	50%	54%	51%
	Ceiling height	53%	54%	50%	51%	54%
	Light contrast	54%	50%	50%	62%△	50%
	Light levels	56%△	46%	49%	49%	52%
	Light temperature	50%	56%	46%	44%▽	52%
	Blocked path	41%▽	52%	50%	52%	48%
	Empty room	52%	58%△	48%	50%	47%
	Interior arrangement	53%	55%	44%▽	59%△	50%
	Hiding place	51%	55%	55%△	52%	47%
	Triggers present	47%	51%	54%	45%	50%
	Battery present	51%	53%	50%	51%	50%
	Note present	49%	56%	56%△	47%	54%
	Cutscene	43%▽	55%△	49%	45%▽	50%
	Event	52%	51%	56%△	61%△	51%
Changes in Affect Ampltitude	Area size	54%△	51%	52%	52%	53%△
	Ceiling height	53%	57%	57%	54%	58%△
	Light contrast	54%△	52%	55%	50%	54%
	Light levels	40%▽	40%▽	39%▽	38%▽	36%▽
	Light temperature	69%△	77%△	72%△	69%△	71%△
	Blocked path	48%	45%▽	43%▽	43%▽	45%▽
	Empty room	34%▽	28%▽	33%▽	37%▽	29%▽
	Interior arrangement	45%	42%▽	49%	50%	43%▽
	Hiding place	74%△	80%△	74%△	72%△	73%△
	Triggers present	47%	50%	44%▽	49%	47%
	Battery present	64%△	70%△	67%△	64%△	68%△
	Note present	54%△	57%△	60%△	59%△	60%△
	Cutscene	95%△	91%△	77%△	90%△	87%△
	Event	74%△	79%△	69%△	81%△	73%△
Changes in Affect Gradient	Area size	44%	44%▽	50%	48%	47%
	Ceiling height	51%	47%	58%	38%▽	50%
	Light contrast	52%	49%	53%	44%	56%
	Light levels	61%△	55%	52%	54%	45%
	Light temperature	42%	48%	48%	33%▽	57%
	Blocked path	45%	48%	48%	54%	45%▽
	Empty room	54%	56%	48%	55%	47%
	Interior arrangement	56%	46%	47%	53%	48%
	Hiding place	45%	48%	45%	41%▽	58%△
	Triggers present	42%	57%△	50%	51%	51%
	Battery present	47%	45%	46%	51%	56%
	Note present	61%△	50%	49%	49%	47%
	Cutscene	55%	52%	46%	52%	48%
	Event	56%	58%△	58%△	55%	51%

servations, and in general match expectations from the literature regarding e.g. room illumination (Joosten et al., 2012; Graja et al., 2020), it is difficult to estimate why such spatial features have such profound impact without considering other factors such as the co-occurrence of in-game events: the non-linear models of Section 7.2.2 which combine all features can perhaps address this limitation.

7.2.2 | Modeling Task

To assess how the combination of changes in room properties impact affect state transition as a whole, we train RF classifiers with all property changes (or degree of change) as input and the affect measure change (increase or decrease) as output. Given the initial small dataset of affect changes (between 679 and 1377 data points), we leverage RFs for their more robust performance on small datasets and on similar experiments for predicting affect change (Melhart et al., 2021b). The input is the difference between properties of the final room versus those in the previous room for the 15 features (see Table 4.2), while the desired output is whether there is an increase or a decrease in the affect metric in the final room compared to the previous room. We only use data for which there is a clear increase or decrease in the affect metric (with the threshold of $\epsilon = 5\%$), and ignore any room transitions where there is no change in the affect metric. Moreover, to balance the data we mirror the ordinal relationships: i.e. for each room transition, we produce two data points, one with the difference of the first room minus the second room and one with the difference of the second room minus the first room. We thus attain a baseline accuracy for every fold (training, testing and validation) of 50%; see details in Chapter 3.

A leave-one-subject out cross validation protocol is followed for training and testing the RF classifier. Since we have 16 play-throughs of the Asylum level by 16 streamers, we reserve a single streamer's data for *hyperprameter (HP) tuning* and the remaining 15 streamers are used for training and testing. Out of these 15 streamers, we perform a *Leave-one-Subject-out Cross-Validation (LOSOCV)*, where 14 streamers are used for the training set while the remaining streamer is used for testing, doing this repeatedly choosing a new streamer for testing (15 repetitions). This ensures that the data of this streamer (both the play-through stimulus and the emotion manifestations) are unseen by the trained models. We repeat the LOSOCV process 16 times, selecting a new streamer for HP tuning. As RFs are stochastic, we repeat training per fold 5 times and the average RF statistics on the test set from all 1200 trials (16 validation sets $\times$ 15 test sets $\times$ 5 repetitions). Results for these tests are shown in Table 7.3. The tuned HPs for this task are: the number of trees used, maximum tree depth, minimum number

Table 7.3: Random Forest statistics on the test set and dataset sizes, averaged from 1200 trials. Single best accuracy scores per affect measure are underlined. Test baseline is 50%. Under-performing and non-significant Accuracy scores denoted with *.

Changes in Affect Mean						
Affect	Accuracy	F1 score	Precision	Recall	Train (n)	Test (n)
Fear of face (F_f)	<u>63%</u>	63%	63%	63%	831	59
Surprise of face (F_s)	60%	60%	60%	60%	1062	76
Arousal of voice (V_a)	55%	55%	55%	55%	1188	85
Fear of utterances (U_f)	55%	55%	55%	55%	677	48
Surprise of utterances (U_s)	57%	56%	57%	57%	955	68
Changes in Affect Amplitude						
Affect	Accuracy	F1 score	Precision	Recall	Train (n)	Test (n)
Fear of face (F_f)	<u>78%</u>	78%	78%	78%	701	50
Surprise of face (F_s)	73%	73%	73%	73%	1130	81
Arousal of voice (V_a)	70%	70%	70%	70%	1209	86
Fear of utterances (U_f)	70%	70%	70%	70%	754	50
Surprise of utterances (U_s)	69%	69%	69%	69%	995	71
Changes in Affect Gradient						
Affect	Accuracy	F1 score	Precision	Recall	Train (n)	Test (n)
Fear of face (F_f)	52%*	51%	52%	52%	392	28
Surprise of face (F_s)	<u>57%</u>	56%	57%	57%	609	44
Arousal of voice (V_a)	48%*	48%	%%	48%	838	60
Fear of utterances (U_f)	52%	51%	52%	52%	376	27
Surprise of utterances (U_s)	52%*	51%	52%	52%	544	39

of samples per leaf node and the minimum number of samples required to split a tree node.

Table 7.3 shows the accuracy results of the trained RFs on the test set (the unseen play-through of an unseen streamer), averaged from 80 trials (16 folds with 5 repetitions). As expected from observations in Section 7.2.1, predicting increase or decrease in affect mean is more challenging (with best accuracies at 63% for F_f) compared to predicting changes in affect amplitude. This is largely due to the way the metric is computed, and the game genre which relies on short bursts of emotion via jump scares. Moreover, the affect gradient measure yielded the lowest test accuracies among the three measures, with the best performance reaching only 57% accuracy for F_s. We note that overall it is easier to predict emotion in facial expressions for the chosen spatial and gameplay inputs. Best accuracies are achieved when predicting this modality, whether considering mean affect gradient, or amplitude. The most challenging modality to predict are the emotions (at least fear and surprise) of utterances. Overall, however, we observe that

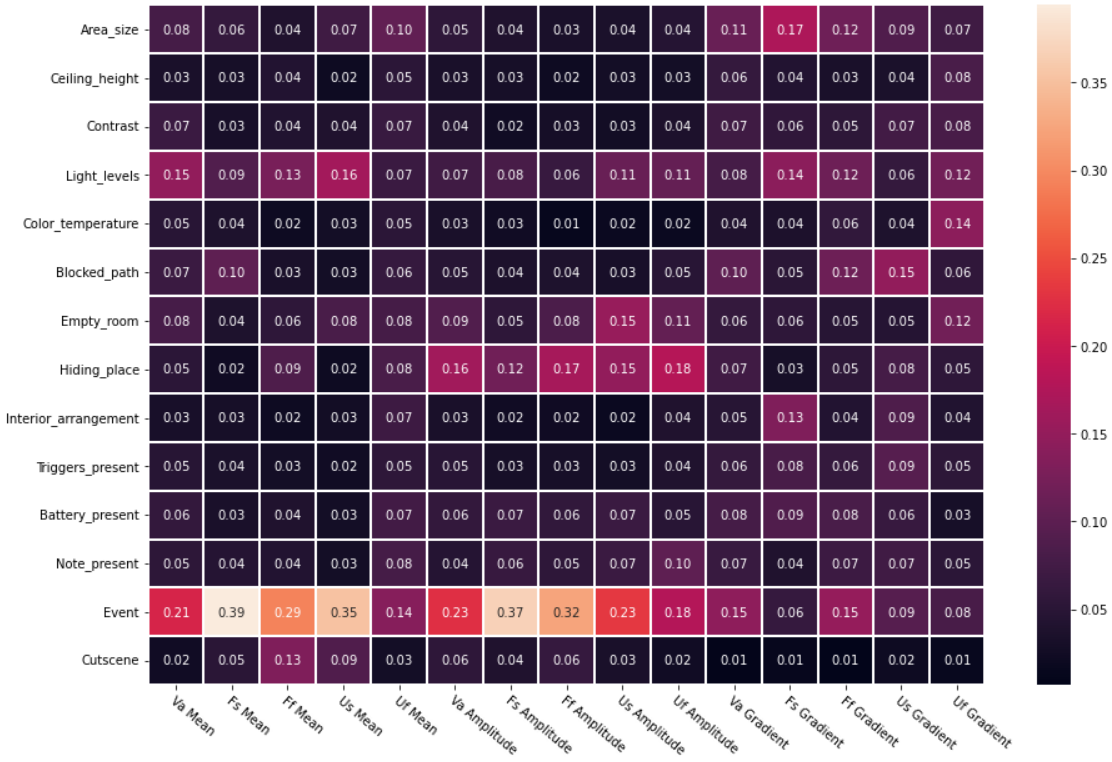

Figure 7.2: Most impactful properties for RF predictions, per affect modality.

even with the simple room properties labeled by experts (which do not include, for example, visual decor or audio information) we can reach accuracies over 70% for several affect signals when considering their highs and lows (as amplitude) instead of, for example, the mean values throughout a room traversal. Another thing to note is the final dataset sizes for the training set.

Both for amplitude and mean all test accuracies are significantly ($p < 0.05$) higher than the baseline 50% considering a binomial distribution.

7.2.2.1 | Feature importance

While test accuracies paint a promising picture, they are not very informative for the impact of level design features. To illustrate which spatial and gameplay features are most impactful when predicting affect change, we use the impurity-based feature importance metrics calculated by RFs. In Figure 7.2 we report each affect metric per signal and its importance on that prediction task. We include all models in Figure 7.2, even if they do not reach high accuracies. What can be observed is that regardless of affect metric (mean, amplitude or gradient) or signal, affect changes between rooms are mostly

predicted based on the presence of an "event". This matches observations of linear relationships in Table 7.2; however, we see that presence of cut-scenes is far less impactful for the RF models (only being important for changes in mean of F_f and U_s). Observing changes in amplitude (which result in far more accurate RF models), the presence of a hiding place is consistently important, as is the presence of empty rooms and batteries. Illumination seems to play a role as well, although it is not consistent: in some cases it is changes in color temperature and in some cases it is the light level (dark or light). Since similar linear relationships were found for affect amplitude in Table 7.2, we expect that changes in light color temperature and levels coincides with other in-game elements such as the presence of cut-scenes and jump-scare events under specific conditions. The compound impact of space, gameplay, and other aspects not captured in features of this study (such as audio) is difficult to tease apart into concise linear relationships. We revisit this limitation in the following section.

7.3 | Summary

This study focused on a real-world case of emotional activities (gameplay) based on a popular commercial horror game and the affect manifestations of professional YouTube streamers. Results indicate that, when viewing the traversal of the game's level architecture (and embedded narrative events) there is a relationship between each room's properties and the players' emotion manifestations. Comparing the mean arousal, fear, or surprise between two consequent rooms makes for a challenging affect modeling task. On the other hand, the genre of horror games leads to many short bursts of emotion and thus comparing those bursts (via affect amplitude) between rooms leads to more accurate models with average test accuracies as high as 78% for fear in facial expressions (see Table 7.3). In terms of the insights such models offer, the choice of using amplitude as the metric expectedly offers limited insights: for the most part, jump scares (tagged as "events") are the most efficient triggers for increased affect amplitude.

While results are promising given the challenging task of in-the-wild multi-modal affect analysis, it is worth considering the complex nature of the stimulus and the limitations of our approach for parsing it. Leveraging expert labels of custom design features (see Table 4.2) allows us to track properties that are deemed important in the literature, but also specific to the game (e.g. the presence of batteries). However, the onerous task of annotating almost 9 hours of gameplay videos cannot scale to any size of dataset. Possible annotation errors, especially when features change multiple times during a play-through (e.g. due to the player re-spawning), may have added noise to the input

data for the models of Section 7.2. In addition, some important features cannot be easily labeled manually: visual effects, background music, audio cues, and creepy iconography may impact emotion but are difficult to capture in timed labels. For the most part, audio cues coincided with jump scares and are included in the all-encompassing "event" tag. However, the nature or context of such audio cues is not fully captured. Future work could explore expanding the manual labeling with outputs of pre-trained models for visuals, e.g. a vision transformer (Dosovitskiy et al., 2020), or audio, e.g. the BEATS model (Chen et al., 2023). However, such inputs would reduce the explainability of the model, which was already problematic due to the composite nature of the stimulus (Liapis et al., 2018).

We also note that this study viewed only a subset of possible affect manifestations. While observing face cameras, voice, and utterances provides a holistic view of the streamer's affect state, we only processed fear, surprise and arousal predictions of pre-trained models in these modalities. Observing other affect signals (e.g. sadness or joy) would lead to more experiments and would likely dilute the findings of this study. Similarly, treating the affect signals in additional ways beyond changes in mean and changes in amplitude—such as the gradient of affect within a time window (Lopes et al., 2017b)—did not result in very accurate models and offered limited insights on the impact of design features in the first place. Instead, future work could seek to aggregate the affect derived from these different modalities into more concise metrics, such as fusing them into a singular affect construct (e.g. transforming categorical labels to dimensional affect data) rather than predict each signal separately. However, this would require a ground truth via third-person affect annotations which would likely also add reporting biases due to complex stimuli and an extensive corpus.

This study proposes a way of mapping game design properties to affect manifestations of professional players and YouTube streamers. The benefit of this approach is that this data exists in-the-wild, and is easy to acquire. Even though the streamer community is not necessarily the most diverse, it is also easy to acquire a balanced corpus with sufficient search. The emotion manifestations are also of fairly high quality, given the fact that streamers learn to talk through their gameplay and emote (especially in horror games). There may be noise within such corpora, as some streamers over-emote or trigger specific events in order to increase audience engagement and viewership; however, such noise is expected and perhaps even more evident in other in-the-wild affect datasets (Kollias and Zafeiriou, 2018). Future work should further explore the potential of such high-quality stimuli (commercial games) and real-time emotions from players. On the one hand, a validation study regarding the output of pre-trained models compared to expert annotations of affect would assess the validity of our approach. On

the other hand, testing the method proposed in this study on other game genres beyond horror could also gauge its generalizability. Other games may trigger less evident emotion manifestations, or those manifestations could be due to cognitive processing of the game state versus the visceral reactions of *Outlast*. We consider the corpus of "let's play" videos, which combine gameplay footage and multi-modal affect manifestations, a fertile ground for research on the impact of visuals, audio, and—in this study—level design (Liapis et al., 2014) on players' emotions.

As described above, we characterize this methodology as "third-person annotation" (the pre-trained models log affect responses) "on active stimuli" (affect data are collected during the interaction with the VE). In the following chapter we compare all four studies and their methodologies and discuss their impact on the collected affect responses.

8

Discussion and Conclusions

This dissertation aims at informing on the affective impact of space and building models of affect using different stimuli and spatial representations, while studying different affect gathering methods using free and forced response and first-person and third-person annotation approaches. All these are done in accordance to the problem statement formulated in Section 1.1 and the research questions in Section 1.3.

Two main study areas contribute within this work, *Architectural Design* and *Affective Computing*, one feeding into the process of space synthesis by identifying the key elements of space, while the other contributes with the theory and methods of capturing, processing and modeling affect. Two research questions are formulated and investigated throughout this work in our attempt to address the problem statement, one focusing on the key elements of space and determining their affective impact across different stimuli. The second research question focuses on the modeling aspect of this relationship. In the remainder of this section we address how the findings of this dissertation address each one of these two questions.

The first question is formulated as follows: *how can we reliably estimate the impact of key elements in architecture?*, and is composed of three further sub-questions one focusing on parameter impact across stimuli; and the other on parameter impact across affect dimensions.

- **RQ1.1 Can we observe similarities across different types of virtual stimuli.** In Chapters 5, 6 and 7 we conducted a linear approach to determine the affective impact based on the chosen elements of Contour Geometry (Curvature), Height, Occlusions and Ambient Illumination. This analysis followed an inter-rater agreement analysis determining the level of agreement between annotators based on short-term memory feature changes on consecutive rooms in each sequence. The

results for all three types of stimuli have attributed the highest impact on the Occlusions parameter, followed by Curvature and Ambient Illumination setting, while Height displayed the weakest agreement trends out of the four. For our *OutlastAFF* corpus (Chapter 7), Curvature was not investigated as all rooms of the Asylum map followed a rectilinear form. As for the remaining features, Occlusions and Illumination settings displayed the highest impact in affect changes, with the Height feature not contributing to a significant degree here as well. Additionally in Chapter 6 we investigate parameter impact across immersive and non-immersive displays. A noticeable difference between the two media displays is that pleasure annotations for Occlusions produced higher agreements between participant for the VR sessions when compared to the desktop display sessions. For Curvature and Illumination parameters, desktop sessions displayed slightly higher agreements than the VR sessions. As for the Height feature, pleasure mean ratings displayed almost similar effect between the two media.
In summary, we were able to observe similarities across different types of virtual stimuli and different types of immersive displays, regarding the impact of spatial parameters. Additionally, findings reveal similarities even between different contexts, comparing the OutlastAFF horror-and-play study and the minimal setup of the AffroomsMR study.

- **RQ1.2a How do key elements in architectural design affect perceived levels of arousal and pleasure (valence)?** The *Affrooms12* corpus in Chapter 5 demonstrates a comparison between the two affect dimensions on the same stimulus but on different set of annotators, as half of the sample was assigned on the Arousal sessions and the other half on the Pleasure sessions. The feature impact based on the inter-rater agreement trends shows several differences between the two affect labels, with Arousal sessions displaying higher agreement trends than Pleasure sessions, across the different affect measures and inter-annotator thresholds. Findings revealed that Occluding elements of walls and columns display in both affect labels the highest agreements, following that Illumination color displayed the highest agreement trends in the arousal sample and contour curvature in the pleasure sample. Regarding room height only arousal sessions could display noticeable trends with pleasure sessions displaying only mild effects.
  In brief, we were able to notice some differences between the two dimensions of arousal pleasure across measures and different features. These findings highlight how specific spatial parameters influence different aspects of affect.

- **RQ1.2b How do key elements in architectural design contribute to affect emo-**

tion manifestations during gameplay? Affect expressions during gameplay are investigated in Chapter 7. Here we study the Horror game Outlast as a potential informant of the impact of Level design and gameplay and the test sample of streamers reacting during their gameplay sessions. Five affect dimensions were explored; fear from face, surprise from face, vocal arousal, fear from utterances and surprise from utterances; across three expression modalities that were generated using pre-trained emotions recognition models.
The key takeaways from this work reveal the importance of gameplay events such as cut-scenes and jump-scares, as well as gameplay elements such as hideouts, supplies and notes. These elements play a crucial role in advancing the plot and shaping the player's emotional experience. As for the impact of spatial attributes, lighting levels, color, interior configuration and occluding elements displayed the highest impact, but parameters like scale (area size and ceiling height) did not display the same effect.
In summary, the findings revealed an expected outcome of the significance of gameplay features and the distinction of the most important spatial features. Including additional levels and gameplay sessions would enrich the dataset and could improve the current state of these results.

- **RQ1.3 Are absolute or relative measures of affect more appropriate measures of the impact?** This question is investigated in Chapters 5 to 7. All four studies used the same measures of affect, one absolute measure; the mean affect level per room and two relative measures; amplitude and gradient of affect per room. This work displays an expected consistency (with a few exceptions) of relative measures showing better agreement trends indications than the mean affect as our absolute measure. Mean affect generally displayed higher trends than the adjacent relative measures in the expert annotation sessions (see Chapter 5) and in the *first-person annotation of active stimuli* sessions for the AffroomsMR study (see Chapter 6), for the spatial features of curvature and illumination color.
The reliability of relative measures over absolute has been the study of several works in AC (Camilleri et al., 2017a; Lopes et al., 2017b). Here we were able to display this with the inter-annotator agreement analysis of all four user studies.

The second research question, *How effectively can we capture and model affective responses (arousal & pleasure) based on environmental key elements across different stimuli?* focuses on the modeling aspect of this work. The following sub-questions were composed and investigated:

- **RQ2.1 Can we minimize temporal biases through effective data processing?** We answer this question in Chapter 5, where we compare 4 different memory settings from a single memory to 3, 5 and all-vs-all. The *Affrooms12* study displays consistently better results when looking at consecutive rooms rather than rooms that are further apart in a sequence. This is expected as annotators display short-term behavior that may not hold throughout the entire session, making single and short-term memory settings more appropriate for studying on the modeling and feature impact tasks. Considering the aforementioned, the memory setting parameter integrated within our framework, is an integral component during the processing of continuous annotations, especially across longer sessions. For the remaining studies (Affrooms24, AffroomsMR and OutlastAFF), only the short-term effects are investigated using solely consecutive rooms for pairwise comparisons.

- **RQ2.2 To what extent can we model elicited arousal and pleasure of virtual spaces?** This question is answered with the constructed models of all four studies in Chapters 5, 6 and 7. In the *OutlastAFF* study (Chapter 7), where horror games are investigated as a potential elicitor, modeling the Amplitude of changes in affect expressions showed the best accuracy results in this work with a score of 78% and a 65% for mean affect. Comparing these affect models with the models we constructed for the *Affrooms12* study, where we use videos as the type of elicitor and study affect annotations, the best performing models were constructed using the gradient of affect with 69% for arousal and 68% for pleasure, 62% for both affect dimensions for the amplitude measure and predictions in affect mean showed 58% for pleasure and 54% for arousal. Additionally, in the *AffroomsMR* study, pleasure modeling for the VR sessions reached accuracies up to 67% and 64% for the Desktop sessions. One important thing to note here is that the stimuli for *Affrooms12* and *AffroomsMR* compared with *OutlastAFF* differ greatly considering that the former consists of pre-recorded traversal of rooms or environments with minimal interaction (3-degrees-of-freedom) with limited feature set and the latter an intense and interactive experience comprising of additional game-based features and richer space feature set. Results from both conditions demonstrate the feasibility of building models of affect across two different affect dimensions, with different virtual settings and different context (horror game and minimal VEs).

- **RQ2.3a How can we capture first-person annotations during the act of virtual exploration and what is the impact of display type during this experience?** Chapter 6 aimed at answering this question by implementing continuous affect measurements within interactive VR and Desktop display experiences. The study

investigated a real-time approach of exploring spaces while "forcing" annotators to indicate their perceived pleasure changes as they traverse through a series of rooms. The proposed environment generates traversals of randomized sequences of different rooms on predetermined paths at a fixed movement speed. Annotators were solely allowed the control of the camera (3-degrees of freedom). This limitation in navigation ensured firstly a consistent duration for all annotators within each room and secondly, fixed room stimuli configuration for the same sequence. The present study demonstrated the feasibility of collecting first-person annotations during the act of exploration using an active stimulus (interactive virtual experience).

Additionally, the primary goal of the AffroomsMR study (see Chapter 6), was to evaluate the usability of an affect annotation tool integrated within the virtual environment for two display types. A total of 31 took part in the study and contributed with their feedback, experiencing both setups and assessing each with a post-experience survey. The survey assessed the key areas of usability, distraction and presence. The findings reported in Chapter 6 reveal that even though VR environments are generally preferred in terms of usability and presence, the simultaneous task of affect annotation can pose challenges if not carefully designed to align with the immersive nature of the experience. Participants noted medium distractions, that could in turn affect the quality of the gathered annotations. Moreover, participants' familiarity with the medium was rated as low, which may have influenced their responses to the distraction-related questions. This suggests that prior experience with VR could play a role in how users perceive and interact with affect annotation tools within virtual environments.

8.1 | Contributions

This section highlights the different contributions of the dissertation regarding the areas of affective computing and architectural design exploring various methods of spatial representations. This work focuses mainly in two types of stimulus; *passive stimulus* in the form of pre-recorded videos of room traversals and *active stimulus*; in the form of interactive VEs and video game environments. The main contributions of this work are detailed as follows:

- **A continuous affect capturing methodology of non-static spatial stimuli using the PAGAN platform and Ranktrace annotation method:** As highlighted in the problem statement (see Section 1.1), most recent works that look into space and

affect either propose methods of collecting biomarkers or subject feedback with discrete affect labels. One of the aims of the present dissertation was to acquire subject continuous annotations of affect in relation to non-static spatial stimuli. This dissertation proposed an approach of gathering these type of affect data using *forced-response first-person annotations* on videos and immersive VEs using the *RankTrace* annotation method. This methodology allows for a more fine-grained analysis of spatial stimuli and an ordinal approach in collecting affect annotations. Future work could use these studies as a pilot for continuous annotations in relation to spatial stimuli and explore other methods such as BTrace (Melhart et al., 2019), *AffectRank* Cowie et al., 2013) or GTrace (Cowie et al., 2013) (Yannakakis and Martinez, 2015).

- **An extended processing framework for continuous affect annotations of non-static spatial stimuli**: Based on the works of Cowie and McKeown, 2010, this study introduced an extended framework for processing continuous data of affect that were captured under real-time conditions while experiencing synthetic stimuli both interactive and and passive in nature. This framework and all of its components were used in the four datasets that were the result of the four different studies of this dissertation. The aim for this framework was to produce more reliable affect labels based on the principles of Preference Learning and Ordinal treatment of affect data, while minimizing temporal biases caused by the nature of the chosen stimuli.

- **The Affrooms24 Corpus**: The first study using expert annotators introduced the *Affrooms24* dataset which contributed to understanding how transitions between spaces and arousal levels can be analyzed in a time-continuous manner. The dataset includes 24 different room configurations based on different settings of the studied features of Curvature, Height, Lighting color and Occlusions. Twenty random sequences of rooms were generated and navigated, resulting in 20 playthrough videos that capture a diverse set of spatial transitions. The average duration of each video is 186 seconds (ranging from 164 to 240 seconds) and total duration of all videos is 62 minutes of footage of 3D spatial navigation and encompasses a variety of design features of architectural form and light color. The three annotators that were employed for the specific study, contributed with their arousal annotations resulting around 186 minutes of raw arousal annotations.

- **The Affrooms12 Corpus**: The second study leveraged crowd-sourcing platforms to create a robust dataset with over 200 annotated videos and data from 76 annota-

tors for both **arousal** and **pleasure**, which significantly expands on previous work of the *Affrooms24* that relied on a small group of expert annotators. The dataset re-uses rooms designed in the work for the *Affrooms24* dataset. To lower cognitive load and time requirements for annotating such long videos, the *AffRooms12* dataset contains 55 short videos of spatial navigation between 12 different rooms, chosen randomly from the 24 possible room configurations. The total dataset duration is on average 17.2 hours with a an average duration of 141 seconds per video. After the cleanup process, the dataset contains 224 annotated videos of arousal from 39 participants and 215 annotated videos of pleasure from 37 participants.

To the author's knowledge, both *Affrooms24* and *AffRooms12* are the only existing corpora of synthesized spaces that are paired with continuous dimensional affect annotations, featuring contributions from both expert and non-expert annotators.

- **A data collection and affect extraction pipeline for capturing elicited affect of readily available streamed content in the wild.** The OutlastAFF study – documented in Chapters 4 and 7– proposed a way of mapping game design properties to affect manifestations of professional players and YouTube streamers. The benefit of this approach is that this data exists in the wild, and is easy to acquire. Even though the streamer community is not necessarily the most diverse, it is also easy to acquire a balanced corpus with sufficient search. The emotion manifestations are also of fairly high quality, given the fact that streamers learn to talk through their gameplay and emote (especially in horror games). There may be noise within such corpora, as some streamers over-emote or trigger specific events in order to increase audience engagement and viewership. However, such noise is expected and perhaps even more evident in other in-the-wild affect datasets (Kollias and Zafeiriou, 2018). Additionally, testing the method proposed in this work on other game genres beyond horror could also gauge its generalizability.
  Future works could further explore the potential of using streamed content on other themes beyond videogames. Other types of streamed content, such as reaction videos or live streaming sessions focusing on non-gaming topics, could offer similarly rich opportunities for research. For example, *reaction videos* could provide insights into shared emotional experiences while analyzing viewership dynamics during live streams and could reveal patterns of attention, engagement, and real-time feedback, similarly done in Melhart et al. (2020a).

- **The Outlast Asylum Affect Corpus (OutlastAFF)**: The last study of the disser-

tation introduces the OutlastAFF corpus containing approximately 8.5 hours of YouTube streams of different play-throughs on the same level (*Asylum*) of the horror game *Outlast*. The collected videos, were in the form of 16 complete runs form 16 popular YouTube streamers. All processed data have a full-screen view of the game (containing its own audiovisual content) and a face camera overlay of the streamer along their own vocal narration of the play-through. Complete play-throughs in the dataset lasted between 22 and 48 minutes. All videos in the dataset include manual annotations of game events and spatial properties of all the 41 rooms of the Asylum level. Additionally, multi-modal labels of affect were generated through pre-trained models, using the streamer's facial expressions, vocal features and utterances during gameplay.

8.2 | Limitations

This section presents the limitations of each study throughout this dissertation. The general limitation of the dissertation is that the focus of affect capturing methods is on forced-affect annotations, with the exemption of OutlastAFF study, that explored manifested affect capturing techniques. As mentioned in Chapter 1, most of the architectural design and affect research relies on intrusive methods that require participants to be physically present during sessions. A significant portion of this dissertation was conducted during the COVID-19 pandemic, which required the first half of the work to focus on collecting affective data through remote methods. As a result, an emphasis in gathering affect through remote means is also explored with this work.

Additionally, no platform, tool or end solution is proposed, such as *Sentient Sketchbook* (Liapis, 2015) and PAGAN (Melhart et al., 2019), but rather affect data pipelines for capturing, processing and treating continuous affect, with the aim to produce spatial feature impact and train models of affect. However, the main outputs of this work could be considered as components for a design-aiding tool, meant to inform designers during consecutive room synthesis on the potential affect changes observers could experience.

Additional limitations regarding the collected corpora are as follows:

- **Affrooms12 & Affrooms24:** The study identified several key limitations. The Affrooms12 and Affrooms24 datasets followed a bottom-up approach, using minimal room configurations for the synthesized spaces, and the relatively weak stimuli in the first two studies, which focused on videos, may have influenced the

results. Since the spaces were often similar to each other, there was no goal or obstacle to the player's navigation, and the annotators were simply observing another player navigate, it is possible that the annotations were based on cognitive rather than affective evaluations. To address these issues, the subsequent studies focused on interactive virtual environments (Chapter 7) and the Outlast Horror game (Chapter 6). Additionally, several assumptions were made in constructing the ground truth for affect, particularly in linking it to characteristics of the built environment. Methods like majority rule for determining "universal" annotation behaviors may have overlooked smaller groups of annotators with valid but less common perspectives. Moreover, crowd-sourcing allowed access to a larger pool of annotators, but this also introduced variability and potential biases in individual interpretations of affect, which could affect the dataset's quality and the analyses. Furthermore, the limited scope of spatial stimuli used in the study may not have been sufficient to elicit strong affective reactions, potentially adding noise and bias to the ground truth data, as a lack of diversity in environmental features may have restricted the range of emotional responses captured in the annotations.

- **AffroomsMR Corpus:** Limitations regarding the *AffroomsMR* corpus were that the study used randomly generated sequences of rooms per participant, for a limited amount of participants. This allowed us to more robustly check how users perceive the annotation task in many scenarios, but could not allow us to perform inter-annotator agreement tests in terms of their reported valence or their FOV behavior. Secondly, the traversal of the rooms was done on a fixed path (on "rails") at a fixed move speed. This ensured consistent task durations and easier data analysis regarding the varying design parameters and their order of appearance. However, it does not capture how users move around the space and explore it. Lastly, immersive media such as VR call for domain-specific affect annotation mechanisms. Our current methodology leveraged a successful affect annotation tool for desktop and implemented it in VR while occluding part of the user's view. It is important to define ad-hoc goals and criteria as in Xue et al., 2021 and Toet and van Erp, 2019 and design real-time continuous affect annotation interfaces explicitly for immersive media and mixed realities.

- **Outlast Asylum Affect Corpus:** Several limitations impacted the study and results of the Outlast Asylum Affect Corpus. One issue was the subjective annotation of design features. The game's level design features, such as room architecture, illumination, and gameplay affordances, were annotated by a single individual. While some aspects, like the presence of batteries, were relatively objective,

this approach introduced the risk of annotation errors, especially given the nearly nine hours of gameplay videos. Features that frequently changed during the playthrough, such as player re-spawns or object interactions, were particularly prone to these errors. Another limitation was the limited labeling of complex stimuli. Key design elements like visual effects, background music, and audio cues, which strongly influence emotional responses, were difficult to label manually. For instance, while audio cues coincided with jump scares, they were not detailed further in terms of sound nature or context, reducing the precision of emotional impact analysis. The study also focused on limited affect signals, specifically arousal, fear, and surprise, derived from facial expressions, voice, and utterances, while other emotions, such as joy or sadness, were not analyzed. This narrowed scope, along with the use of simpler metrics like mean or amplitude change, may have limited the depth of insights provided by the models. Further challenges included the potential noise introduced by streamers, as streamers might exaggerate emotional expressions to engage viewers or retain their audience. This behavior introduced noise into the dataset, possibly affecting the accuracy of emotional models, especially for subtler or less intense emotional reactions. The study also relied on pre-trained models to predict emotions, but these models may not fully capture the complexities of emotions in horror games. Without validating these models against expert-annotated affect signals, the accuracy of the results could be compromised. Furthermore, the study employed limited input modalities, focusing on facial expressions, voice, and utterances. However, the complex nature of gameplay experiences may have benefited from a more holistic approach, such as combining these signals into a singular metric. This would require more advanced ground-truth data and could introduce new biases. Finally, there was variability in feature importance. Similar to the Affrooms12 dataset, the analysis revealed variability in how different design parameters affected affective predictions, but the study did not explore these relationships in depth. Further investigation into how specific features interact and their combined effects on affective responses is necessary.

- **Affect model Designer validation:** One additional limitation of this dissertation is the absence of user or designer validation within the proposed affect model. While the research successfully introduces an approach for gathering continuous affect annotations in response to spatial stimuli using virtual spaces, videos, and gaming environments, it primarily focuses on three key aspects: synthesizing appropriate stimuli grounded in architectural design principles, proposing and implementing

data collection strategies for continuous affect annotation, and developing data processing pipeline to extract affective labels and features to evaluate the impact of space on affect and build affect models. However, the step of closing the loop in the affective loop model—where the derived affect model is utilized and validated by an external entity such as a designer, user, or other stakeholders—remains beyond the scope of this thesis. This limitation arose primarily due to the complexity and scope of integrating this iterative validation phase, which would require additional methodologies, extended collaboration with external stakeholders, and further iterations of the model. Future research should aim to address this limitation by implementing this crucial validation phase, thereby completing the affective loop and enabling practical applications of the model for spatial design and related fields.

.

8.3 | Extensibility

The contributions and limitations of this work have been described above but what are the next steps for future research? This section contributes to further steps in extending the constructed datasets, directions for future implementations extending spatial knowledge-base and additional affect capturing modalities.

8.3.1 | Extending Feature Representations

The four conducted user studies use spatial features of the environment were classified into four high-level categories: geometry, scale, lighting, and interior arrangement. These categories were treated with representations either on an interval scale (low, mid, high) or as binary presence indicators (0, 1). While this approach provides a broad understanding of how these elements impact affect, more granular and dynamic representations can extend the analysis. One method is the use of pixel-based approaches, leveraging computer vision models like Convolutional Neural Networks (CNNs) or Transformer architectures (Vaswani, 2017). These methods could automatically extract features by analyzing the visual properties of the environment at the pixel level. For instance, CNNs can capture textures, edges, and object details, while transformers offer attention mechanisms to contextualize spatial relationships across the scene. This could provide a more nuanced understanding of how specific visual stimuli, such as shading, object placement, or fine details, correlate with affective responses.

In addition, spatial representation can be extended by incorporating Isovists (Koutsolampros et al., 2019; Benedikt, 1979) and Space Syntax (Hillier et al., 1976). Isovists describe the area visible from a specific point within a space, providing insights into how the visibility of architectural features, room layout, and occlusions influence perception. Space Syntax, on the other hand, examines spatial relationships through the connectivity and accessibility of different areas, revealing patterns of movement and interaction with the built environment. By employing these methods, one can model how the navigability and visibility of spaces, rather than just their static properties, affect human emotions. For example, complex spaces with hidden corners might generate anxiety, while open and highly connected spaces could foster comfort or relaxation.

Another important theoretical framework is *Processing Fluency Theory* (Reber et al., 2004), which can be applied to spatial feature analysis. This theory posits that stimuli that are easier to process are generally perceived more positively, and this can be translated into spatial design by measuring the cognitive load required to comprehend different spatial features. For instance, simple, familiar geometries might evoke positive affect due to their high processing fluency, while more complex or irregular forms might require greater cognitive effort, leading to neutral or negative responses. Methods that quantify this fluency, such as gaze tracking or reaction time studies, can provide additional layers of data for predicting affective responses to spatial stimuli. Additionally, libraries like like *Imagefluency* (Mayer, 2021) have made it easy for researchers to extend their video or images databases with the theorized parameters of *Fluency*.

Additional spatial feature representations could also include the use of point clouds and 3D scanning technologies to capture volumetric data of environments, offering a highly accurate representation of scale and layout. These data can be analyzed to assess volumetric complexity, surface texture, and object density, further extending the feature set for predicting affect. Another promising approach is the use of environmental acoustics as a spatial feature—capturing how sound behaves in different spaces (echoes, reverberations) can also contribute to the affective perception of space.

8.3.2 | Immersive Media Affect Annotation

In Chapter 6, we instigated the use of immersive environments and HMDs as potential stimuli for affect annotation. We refer to this method as *Real-time Affect Annotation* method or *First-person annotation of Active stimulus*, as users annotate their perceived architectural experience while interacting with a VE. In this specific study, we implemented a simplified version of *RankTrace* within the VE, using a desktop mouse to record one-dimensional affect changes. This approach allowed us to create a scenario compara-

ble to desktop experiences. However, as we shift towards more immersive experiences, it becomes necessary to adopt a more specialized affect annotation mechanism suited to VR.

A more advanced interaction system in a VR setting would support multiple input modes, including *controller-based interactions*. These controllers –now enhanced with haptic feedback– can provide tactile responses and seamlessly register different types of input. Continuous affect annotation tools like *RankTrace* (Lopes et al., 2017b) or *BTrace* (Melhart et al., 2019) could be adapted to the corresponding controller based on the chosen HMD. For instance, *haptic feedback* could notify users of intense affect changes, enabling them to adjust their annotations throughout an experience. Another growing trend is *gesture-based input*, where hand and finger movements are tracked without controllers. Recently introduced HMDs like *Meta Quest 3*[1] and *Vive Focus 3*[2] use depth sensors to allow natural interaction with virtual objects, which could enhance the fluidity of affect annotations. Additionally, many headsets now support *voice input*, allowing for *think-aloud* (El-Nasr et al. (2016)) methods where users can verbally register affect in a semi-structured manner, providing more nuanced data. Lastly, the integration of *eye-tracking* in HMDs offers the possibility of collecting behavioral data alongside subjective affect annotations, enabling a more comprehensive analysis of users' experiences in virtual environments.

Advancements in mixed reality, such as those highlighted here, have the potential to significantly advance the work in this dissertation and contribute to the broader study of architectural design and its relationship with affect.

8.3.3 | Real-World Affect Annotation

The dissertation has successfully investigated synthetic stimuli to draw inferences regarding the impact of spaces and proposed a framework in building predictive models of affect in room to room transitions. Inspired by the works that were detailed in Section 2.3 particularly the studies utilizing Extended Reality (XR) applications (Xue et al., 2021; Xue et al., 2020b), one extension is proposed to build upon the framework of this dissertation. Our framework, which collects and processes continuous affective data in relation to spatial parameters, could be adapted for use in real-world scenarios and physical buildings. Within this scope, a proposed solution for XR-based continuous affect annotation tools is introduced. A potential mobile-based solution that leverages the sensor-rich environment of mobile devices (such as tablets or smartphones) would

[1] `https://www.meta.com/quest/quest-3/`
[2] `https://www.vive.com/eu/product/vive-focus3/overview/`

Figure 8.1: Example of a continuous XR-based Affect annotation tool for Architecture.

enable continuous affect annotation in a real-world context. This solution integrates the device's camera, gyroscope, and GPS location sensors to gather spatial data while users explore physical environments. As users move through these spaces, they can continuously assign affect ratings, using dimensional (e.g., valence, arousal), ordinal, discrete labels, or verbal descriptions to reflect their moment-to-moment experiences. The affect annotations are directly synchronized with captured video data, creating a dynamic, multi-modal dataset that represents both the visual environment and the corresponding emotional states.

The spatial data captured via GPS also allows for the mapping of emotions onto real-world locations, resulting in a "map of emotions" that can provide a more holistic view of affective responses tied to specific spatial contexts. Solutions such as this, could be particularly useful in capturing affect during interactions with urban environments, architecture, or specific social spaces, offering a flexible and scalable tool for in-situ affective research. Such a system could be employed across a variety of disciplines, including psychology, architecture, and urban studies, enabling researchers to study how real-world conditions influence emotional experiences at a granular level. An example for an XR-based affect annotation tool is illustrated in Figure 8.1.

8.4 | Summary

This Chapter concludes the dissertation by outlining the key contributions of this research in *Section 8.1.* The main advancements include the development of an extended processing framework for continuous affect annotations of non-static spatial stimuli, the

creation of affect datasets from four distinct user studies across three different types of stimuli, and the formulation of two models of affect utilizing video content and video games. *In Section 8.2*, we address the general limitations of this work as well as the specific limitations encountered in each user study, focusing primarily on the availability of spatial descriptors, affect annotation methods, and annotator biases related to each stimulus. Finally, we propose future directions that build on the identified limitations and leverage advancements in fields such as XR technologies, affect modeling methods, and AI algorithms, aiming to further enhance the contributions of this dissertation.

References

Nouf Abukhodair, Meehae Song, Serkan Pekçetin, and Steve DiPaola. Designing a wheel-based assessment tool to measure visual aesthetic emotions. *Cognitive Systems Research*, 84:101196, 2024.

Artur Aiguzhinov, Carlos Soares, and Ana Paula Serra. A similarity-based adaptation of naive bayes for label ranking: Application to the metalearning problem of algorithm recommendation. In *International Conference on Discovery Science*, pages 16–26. Springer, 2010.

Nouf Alajmi, Eiman Kanjo, Nour El Mawass, and Alan Chamberlain. Shopmobia: An emotion-based shop rating system. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 745–750. IEEE, 2013.

Christopher Alexander. *A pattern language: towns, buildings, construction*. Oxford university press, 1977.

João Almeida, Luís Vilaça, Inês N Teixeira, and Paula Viana. Emotion identification in movies through facial expression recognition. *Applied Sciences*, 11(15):6827, 2021.

Sergio Altomonte and Stefano Schiavon. Occupant satisfaction in leed and non-leed certified buildings. *Building and Environment*, 68:66–76, 2013.

Anjok07. GitHub - Anjok07/ultimatevocalremovergui: GUI for a Vocal Remover that uses Deep Neural Networks. — github.com. `https://github.com/Anjok07/ultimatevocalremovergui`, 2023. [Accessed 29-Jun-2023].

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, 2020.

Maryam Banaei, Ali Ahmadi, and Abbas Yazdanfar. Application of AI methods in the clustering of architecture interior forms. *Frontiers of Architectural Research*, 6(3):360–373, 2017a.

Maryam Banaei, Javad Hatami, Abbas Yazdanfar, and Klaus Gramann. Walking through architectural spaces: the impact of interior forms on human brain dynamics. *Frontiers in human neuroscience*, 11:477, 2017b.

Maryam Banaei, Ali Ahmadi, Klaus Gramann, and Javad Hatami. Emotional evaluation of architectural interior forms based on personality differences using virtual reality. *Frontiers of Architectural Research*, 9 (1):138–147, 2020.

Matthew Barthet, Chintan Trivedi, Kosmas Pinitas, Emmanouil Xylakis, Konstantinos Makantasis, Antonios Liapis, and Georgios N Yannakakis. Knowing your annotator: Rapidly testing the reliability of affect annotation. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8. IEEE, 2023.

Michael L Benedikt. To take hold of space: isovists and isovist fields. *Environment and Planning B: Planning and design*, 6(1):47–65, 1979.

Marco Bertamini and Michele Sinico. A study of objects with smooth or sharp features created as line drawings by individuals trained in design. *Empirical Studies of the Arts*, 39(1):61–77, 2021.

Fabio Bianconi, M Filippucci, G Magrini, and M Seccaroni. Designing with emotional awareness. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 46:55–62, 2021.

Casper S Boonen and Daniel Mieritz. Paralysing fear: Player agency parameters in horror games. In *Proceedings of 2018 international DiGRA Nordic conference*, 2018.

Brandon M Booth and Shrikanth S Narayanan. Fifty shades of green: Towards a robust measure of inter-annotator agreement for continuous signals. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 204–212, 2020.

Isabella S Bower, Gillian M Clark, Richard Tucker, Aron T Hill, Jarrad AG Lum, Michael A Mortimer, and Peter G Enticott. Enlarged interior built environment scale modulates high-frequency eeg oscillations. *Eneuro*, 9(5), 2022.

Margaret M Bradley and Peter J Lang. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, 25(1):49–59, 1994.

Elizabeth Camilleri, Georgios N Yannakakis, and Alexiei Dingli. Platformer level design for player believability. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2016.

Elizabeth Camilleri, Georgios N. Yannakakis, and Antonios Liapis. Towards general models of player affect. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2017a.

Elizabeth Camilleri, Georgios N Yannakakis, and Antonios Liapis. Towards general models of player affect. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 333–339. IEEE, 2017b.

Kynthia Chamilothori. Effects of façade and daylight pattern geometry on subjective and physiological responses: findings from experiments in immersive virtual reality. In *Kongsberg Vision Meeting 2019: Immersive technologies for eye care and lighting design*, 2019.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Giorgia Chinazzo, Kynthia Chamilothori, Jan Wienold, and Marilyne Andersen. Temperature–color interaction: subjective indoor environmental perception and physiological responses in virtual reality. *Human factors*, 63(3):474–502, 2021.

Francis DK Ching. *Architecture: Form, space, and order*. John Wiley & Sons, 2023.

Alice Chirico, Pietro Cipresso, David B Yaden, Federica Biassoni, Giuseppe Riva, and Andrea Gaggioli. Effectiveness of immersive videos in inducing awe: an experimental study. *Scientific reports*, 7(1):1218, 2017.

Maria Christofi, Despina Michael-Grigoriou, and Christos Kyrlitsias. A virtual reality simulation of drug users' everyday life: the effect of supported sensorimotor contingencies on empathy. *Frontiers in psychology*, 11:1242, 2020.

Andrea Clerico, Cindy Chamberland, Mark Parent, Pierre-Emmanuel Michon, Sebastien Tremblay, Tiago H Falk, Jean-Christophe Gagnon, and Philip Jackson. Biometrics and classifier fusion to predict the fun-factor in video gaming. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 1–8. IEEE, 2016.

Alexander Coburn, Omid Kardan, Hiroki Kotabe, Jason Steinberg, Michael C Hout, Arryn Robbins, Justin MacDonald, Gregor Hayn-Leichsenring, and Marc G Berman. Psychological responses to natural patterns in architecture. *Journal of Environmental Psychology*, 62:133–145, 2019.

Büşra Coşgun, Kemal Yldrm, and Mehmet Lutfi Hidayetoglu. Effect of wall covering materials on the perception of cafe environments. *Facilities*, 40(3-4):214–232, 2021.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20 (1):37–46, 1960.

Büşra Coşgun, Kemal Yıldırım, and Mehmet Lutfi Hidayetoglu. Effect of wall covering materials on the perception of cafe environments. *Facilities*, 40(3/4):214–232, 2022.

Roddy Cowie and Gary McKeown. Statistical analysis of data from initial labelled database and recommendations for an economical coding scheme. *SEMAINE Report D6b*, 2010.

Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 'feeltrace': An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

Roddy Cowie, Martin Sawey, Cian Doherty, Javier Jaimovich, Cavan Fyans, and Paul Stapleton. Gtrace: General trace program compatible with emotionml. In *2013 humaine association conference on affective computing and intelligent interaction*, pages 709–710. IEEE, 2013.

Duncan Cramer. *Fundamental statistics for social research: step-by-step calculations and computer techniques using SPSS for Windows*. Routledge, 2003.

Gaia Crippa, Valentina Rognoli, Marinella Levi, et al. Materials and emotions, a study on the relations between materials and emotions in industrial products. In *ut of Control: Proceedings of 8th International Design and Emotion Conference, Design and Emotion Society*, pages 1–9. Central Saint Martins College of Arts & Design, London, England., 2012.

Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951.

Sibel Seda Dazkir. Emotional effect of curvilinear vs. rectilinear forms of furniture in interior settings. Master's thesis, Oregon State University, 2009.

Marco De Gemmis, Leo Iaquinta, Pasquale Lops, Cataldo Musto, Fedelucio Narducci, Giovanni Semeraro, et al. Preference learning in recommender systems. *Preference Learning*, 41(41-55):48, 2009.

Cláudio Rebelo de Sá, Carlos Soares, Arno Knobbe, and Paulo Cortez. Label ranking forests. *Expert systems*, 34(1):e12166, 2017.

Pieter MA Desmet, Martijn H Vastenburg, and Natalia Romero. Mood measurement with pick-a-mood: review of current methods and design of a pictorial self-report scale. *Journal of Design Research*, 14(3): 241–279, 2016.

Pierre Destrée. Aristotle's aesthetics. 2021.

Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2106–2112. IEEE, 2011.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

P. Ekman. Argument for basic emotions. *Cognition and Emotion*, 6:169–200, 1992.

Paul Ekman. Emotions revealed. *Bmj*, 328(Suppl S5), 2004.

Magy Seif El-Nasr, Athanasios Vasilakos, Chinmay Rao, and Joseph Zupko. Dynamic intelligent lighting for directing visual attention in interactive 3-d scenes. *IEEE Transactions on Computational Intelligence and AI in Games*, 1(2):145–153, 2009.

Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa. *Game analytics*. Springer, 2016.

Semiha Ergan, Zhuoya Shi, and Xinran Yu. Towards quantifying human experience in the built environment: A crowdsourcing based experiment to identify influential architectural design features. *Journal of Building Engineering*, 20:51–59, 2018.

Kirill Fayn, Steven Willemsen, R Muralikrishnan, Bilquis Castaño Manias, Winfried Menninghaus, and Wolff Schlotz. Full throttle: Demonstrating the speed, accuracy, and validity of a new method for continuous two-dimensional self-report and annotation. *Behavior Research Methods*, pages 1–15, 2022.

Marta Ferreira, Ana Pinha, Micaela Fonseca, and Phil Lopes. Behind the door: Exploring horror vr game interaction and its influence on anxiety. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, pages 1–11, 2023.

Anton Flyvholm, Sumit Sen, Emmanouil Xylakis, Stine Maria Louring Nielsen, Georgios Triantafyllidis, Linda Andresen, and Mette Merete Pedersen. A personalised and adaptive intelligent system to adjust circadian lighting for elderly housing. In *15th International Symposium on Ambient Intelligence and Embedded Systems*, 2016.

Bárbara Formiga, Francisco Rebelo, Jorge Cruz Pinto, and Emerson Gomes. How architectural forms can influence emotional reactions: An exploratory study. In *International Conference on Human-Computer Interaction*, pages 37–55. Springer, 2022.

Gerald Franz, Markus Von Der Heyde, and Heinrich H B¨ulthoff. An empirical approach to the experience of architectural space in virtual reality—exploring relations between features and affective appraisals of rectangular indoor spaces. *Automation in construction*, 14(2):165–172, 2005.

Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine learning*, 73:133–153, 2008.

Ervin Garip and Beren Seymen. Research for evaluating perception of concrete material by using visual research methods in learning environments. *A | Z ITU JOURNAL OF THE FACULTY OF ARCHITECTURE*, 18(1):17–28, 2021.

Gerardo Gómez-Puerto, Enric Munar, and Marcos Nadal. Preference for curvature: A historical and conceptual framework. *Frontiers in human neuroscience*, page 712, 2016.

Gerardo Gómez-Puerto, Jaume Rosselló, Guido Corradi, Cristina Acedo-Carmona, Enric Munar, and Marcos Nadal. Preference for curved contours across cultures. *Psychology of Aesthetics, Creativity, and the Arts*, 12(4):432, 2018.

Hugo C Gomez-Tone, Jorge Martin-Gutierrez, John Bustamante-Escapa, Paola Bustamante-Escapa, and Betty K Valencia-Anci. Perceived sensations in architectural spaces through immersive virtual reality. *VITRUVIO-International Journal of Architectural Technology and Sustainability*, 6(2):70–81, 2021.

Google. MediaPipe | Google for Developers — developers.google.com. `https://developers.google.com/mediapipe`, 2023. [Accessed 29-Jun-2023].

Sarra Graja, Phil Lopes, and Guillaume Chanel. Impact of visual and sound orchestration on physiological arousal and tension in a horror game. *IEEE Transactions on Games*, 2020.

Timothy Greer, Benjamin Ma, Matthew Sachs, Assal Habibi, and Shrikanth Narayanan. A multimodal view into music's effect on human neural, physiological, and emotional experience. In *Proceedings of the 27th ACM international conference on multimedia*, pages 167–175, 2019.

Timothy Greer, Karel Mundnich, Matthew Sachs, and Shrikanth Narayanan. The role of annotation fusion methods in the study of human-reported emotion experience during music listening. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE, 2020.

Lara Gregorians, Pablo Fernández Velasco, Fiona Zisch, and Hugo 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*, 82:101841, 2022.

Emanuela Guglielmi, Simone Scalabrino, Gabriele Bavota, and Rocco Oliveto. Towards using gameplay videos for detecting issues in video games. *arXiv preprint arXiv:2204.04182*, 2022.

Ellen Kathrine Hansen, Thomas Bjørner, Emmanouil Xylakis, and Mihkel Pajuste. An experiment of double dynamic lighting in an office responding to sky and daylight: Perceived effects on comfort, atmosphere and work engagement. *Indoor and Built Environment*, 31(2):355–374, 2022a.

Ellen Kathrine Hansen, Mihkel Pajuste, and Emmanouil Xylakis. Flow of light: balancing directionality and cct in the office environment. *Leukos*, 18(1):30–51, 2022b.

Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell. On the importance of both dimensional and discrete models of emotion. *Behavioral sciences*, 7(4):66, 2017.

Jochen Hartmann. Emotion english distilroberta-base. `https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/`, 2022.

Tom Heath, Sandy G Smith, and Bill Lim. Tall buildings and the urban skyline: The effect of visual complexity on preferences. *Environment and behavior*, 32(4):541–556, 2000.

Harry Helson. Current trends and issues in adaptation-level theory. *American psychologist*, 19(1):26, 1964.

Grant Hildebrand. *Origins of architectural pleasure*. Univ of California Press, 1999.

Bill Hillier, Adrian Leaman, Paul Stansall, and Michael Bedford. Space syntax. *Environment and Planning B: Planning and design*, 3(2):147–185, 1976.

ISO 16817:2017. Building environment design — Indoor environment — Design process for the visual environment. Standard, International Organization for Standardization, Geneva, CH, 2017.

Ju Yeun Jang, Eunsoo Baek, So-Yeon Yoon, Ho Jung Choo, et al. Store design: Visual complexity and consumer responses. *International Journal of Design*, 12(2):105–118, 2018.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142, 2002.

Evi Joosten, Giel Van Lankveld, and Pieter Spronck. Influencing player emotions using colors. *Journal of Intelligent Computing*, 3(2):76–86, 2012.

Caroline Karmann, Bahar Aydemir, Kynthia Chamilothori, Seungryong Kim, Sabine Süsstrunk, and Marilyne Andersen. Saliency prediction in 360° architectural scenes: Performance and impact of daylight variations. *Journal of Environmental Psychology*, 92:102110, 2023.

Boa Kim, Emmanouil Xylakis, Andrei-Ducu Predescu, Georgios Triantafyllidis, Ellen Kathrine Hansen, and Michael Mullins. Designing a lighting installation through virtual reality technology-the brighter brunnshög case study. In *Interactivity, Game Creation, Design, Learning, and Innovation: 6th International Conference, ArtsIT 2017, and Second International Conference, DLI 2017, Heraklion, Crete, Greece, October 30–31, 2017, Proceedings 6*, pages 43–53. Springer, 2018a.

Boa Kim, Emmanouil Xylakis, and Georgios Triantafyllidis. Interactive lighting art installation in virtual environments as a stimulus for public ownership in urban development–brighter brunnshög. In *SHS Web of Conferences*, volume 43, page 01003. EDP Sciences, 2018b.

Jeongmin Kim and Nayeon Kim. Quantifying emotions in architectural environments using biometrics. *Applied Sciences*, 12(19):9998, 2022.

Michael Kipp. Anvil-a generic annotation tool for multimodal dialogue. In *Seventh European conference on speech communication and technology*. Citeseer, 2001.

Igor Knez. Effects of indoor lighting on mood and cognition. *Journal of environmental psychology*, 15(1): 39–51, 1995.

Dimitrios Kollias and Stefanos Zafeiriou. Aff-wild2: Extending the aff-wild database for affect recognition. *arXiv preprint arXiv:1811.07770*, 2018.

Iason Konstantzos, Seyed Amir Sadeghi, Michael Kim, Jie Xiong, and Athanasios Tzempelikos. The effect of lighting environment on task performance in buildings–a review. *Energy and Buildings*, 226:110394, 2020.

Petros Koutsolampros, Kerstin Sailer, Tasos Varoudis, and Rosie Haslem. Dissecting visibility graph analysis: The metrics and their role in understanding workplace human behaviour. In *Proceedings of the 12th International Space Syntax Symposium*, volume 12. International Space Syntax Symposium, 2019.

Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

Christian Krüger, Tanja Kojić, Luis Meier, Sebastian Möller, and Jan-Niklas Voigt-Antons. Development and validation of pictographic scales for rapid assessment of affective states in virtual reality. In *2020 twelfth international conference on quality of multimedia experience (QOMEX)*, pages 1–6. IEEE, 2020.

Patri Lankoski, Staffan Björk, et al. Game research methods: An overview. 2015.

Michael Lekan-Kehinde and ABIMBOLA Asojo. Impact of lighting on children's learning environment: a literature review. *WIT Trans. Ecol. Environ*, 253:371–380, 2021.

James R Lewis. Ibm computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction*, 7(1):57–78, 1995.

Antonios Liapis. Searching for sentient design tools for game development. 2015.

Antonios Liapis, Georgios N Yannakakis, and Julian Togelius. Computational game creativity. Proceedings of the International Conference on Computational Creativity, 2014.

Antonios Liapis, Georgios N Yannakakis, Mark J Nelson, Mike Preuss, and Rafael Bidarra. Orchestrating game generation. *IEEE Transactions on Games*, 11(1):48–68, 2018.

Dayi Lin, Cor-Paul Bezemer, and Ahmed E Hassan. Identifying gameplay videos that exhibit bugs in computer games. *Empirical Software Engineering*, 24:4006–4033, 2019.

Ruby Lipson-Smith, Julie Bernhardt, Edoardo Zamuner, Leonid Churilov, Nick Busietta, and Damian Moratti. Exploring colour in context using virtual reality: Does a room change how you feel? *Virtual Reality*, 25:631–645, 2021.

Tim Lomas, Meike Bartels, Margot Van De Weijer, Michael Pluess, Jeffrey Hanson, and Tyler J VanderWeele. The architecture of happiness. *Emotion Review*, 14(4):288–309, 2022.

Phil Lopes and Ronan Boulic. Towards designing games for experimental protocols investigating human-based phenomena. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, pages 1–11, 2020.

Phil Lopes, Antonios Liapis, and Georgios N. Yannakakis. Targeting horror via level and soundscape generation. In *Proceedings of the AAAI Artificial Intelligence for Interactive Digital Entertainment Conference*, 2015.

Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. Framing tension for game generation. In *Proceedings of the Seventh International Conference on Computational Creativity, June 2016*. ICCC, 2016.

Phil Lopes, Antonios Liapis, and Georgios N Yannakakis. Modelling affect for horror soundscapes. *IEEE Transactions on Affective Computing*, 10(2):209–222, 2017a.

Phil Lopes, Georgios N. Yannakakis, and Antonios Liapis. Ranktrace: Relative and unbounded affect annotation. In *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2017b.

Konstantinos Makantasis, David Melhart, Antonios Liapis, and Georgios N Yannakakis. Privileged information for modeling affect in the wild. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2021.

Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. The pixels and sounds of emotion: General-purpose representations of arousal in games. *IEEE Transactions on Affective Computing*, 14(1), 2023.

Javier Marín-Morales, Juan Luis Higuera-Trujillo, Alberto Greco, Jaime Guixeres, Carmen Llinares, Enzo Pasquale Scilingo, Mariano Alcañiz, and Gaetano Valenza. Affective computing in virtual reality: emotion recognition from brain and heartbeat dynamics using wearable sensors. *Scientific reports*, 8 (1):13657, 2018.

Daryl Marples, Duke Gledhill, and Pelham Carter. The effect of lighting, landmarks and auditory cues on human performance in navigating a virtual maze. In *Proceedings of the Symposium on Interactive 3D Graphics and Games*, 2020.

Alfred J Marrow. *The practical theorist: The life and work of Kurt Lewin.* Teachers College Press, 1977.

Hector P Martinez, Georgios N Yannakakis, and John Hallam. Don't classify ratings of affect; rank them! *IEEE transactions on affective computing*, 5(3):314–326, 2014.

Samantha Matuke. Mathematical beauty in renaissance architecture. 2016.

Paris Mavromoustakos-Blom, David Melhart, Antonios Liapis, Georgios N Yannakakis, Sander Bakkes, and Pieter Spronck. Multiplayer tension in the wild: A hearthstone case. In *Proceedings of the 18th International Conference on the Foundations of Digital Games*, pages 1–9, 2023.

Stefan Mayer. Imagefluency: Image statistics based on processing fluency. *R package version 0.2*, 3, 2021.

Cade McCall, Guy Schofield, Darel Halgarth, Georgina Blyth, Aaron Laycock, and Daniela J Palombo. The underwood project: A virtual environment for eliciting ambiguous threat. *Behavior research methods*, pages 1–16, 2022.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17, 2011.

Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

David Melhart. The anatomy of gameplay: general affect prediction across games and genres. *University of Malta library*, 2021.

David Melhart, Antonios Liapis, and Georgios N Yannakakis. Pagan: Video affect annotation made easy. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 130–136. IEEE, 2019.

David Melhart, Daniele Gravina, and Georgios N Yannakakis. Moment-to-moment engagement prediction through the eyes of the observer: Pubg streaming on twitch. In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, pages 1–10, 2020a.

David Melhart, Konstantinos Sfikas, Giorgos Giannakakis, and Georgios Yannakakis Antonios Liapis. A study on affect model validity: Nominal vs ordinal labels. In *Proceedings of IJCAI 2018 2nd Workshop on Artificial Intelligence in Affective Computing*, pages 27–34, 2020b.

David Melhart, Antonios Liapis, and Georgios N Yannakakis. The affect game annotation (again) dataset. *arXiv preprint arXiv:2104.02643*, 2021a.

David Melhart, Antonios Liapis, and Georgios N. Yannakakis. Towards general models of player experience: A study within genres. In *Proceedings of the IEEE Conference on Games*, 2021b.

Joan Meyers-Levy and Rui Zhu. The influence of ceiling height: The effect of priming on the type of processing that people use. *Journal of consumer research*, 34(2):174–186, 2007.

Talya Miron-Shatz, Arthur Stone, and Daniel Kahneman. Memories of yesterday's emotions: Does the valence of experience affect the memory-experience gap? *Emotion*, 9(6):885, 2009.

Ali Mollahosseini, Behzad Hasani, and Mohammad H Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

Frederik Nagel, Reinhard Kopiez, Oliver Grewe, and Eckart Altenmüller. Emujoy: Software for continuous measurement of perceived emotions in music. *Behavior Research Methods*, 39(2):283–290, 2007.

Simon Niedenthal. Patterns of obscurity: Gothic setting and light in resident evil 4 and silent hill 2. *Horror video games: Essays on the fusion of fear and play*, pages 168–180, 2009.

Kaushika Pal and Biraj V Patel. Data classification with k-fold cross validation and holdout accuracy estimation methods with 5 different machine learning techniques. In *2020 fourth international conference on computing methodologies and communication (ICCMC)*, pages 83–87. IEEE, 2020.

Juhani Pallasmaa. *The eyes of the skin: Architecture and the senses*. John Wiley & Sons, 2012.

Srinivas Parthasarathy, Roddy Cowie, and Carlos Busso. Using agreement on direction of change to build rank-based emotion classifiers. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 24(11):2108–2121, 2016.

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*, 12:2825–2830, 2011.

Rosalind W Picard. *Affective computing*. MIT press, 2000.

Andres Pinilla, Jaime Garcia, William Raffe, Jan-Niklas Voigt-Antons, Robert Spang, and Sebastian Möller. Affective visualization in virtual reality: An integrative review. *arXiv preprint arXiv:2012.08849*, 2020.

Kosmas Pinitas, David Renaudie, Mike Thomsen, Matthew Barthet, Konstantinos Makantasis, Antonios Liapis, and Georgios N. Yannakakis. Predicting player engagement in tom clancy's the division 2: A multimodal approach via pixels and gamepad actions. In *Proceedings of the 25th ACM International Conference on Multimodal Interaction*, 2023.

Paolo Presti, Davide Ruzzon, Pietro Avanzini, Fausto Caruana, Giacomo Rizzolatti, and Giovanni Vecchiato. Measuring arousal and valence generated by the dynamic experience of architectural forms in virtual environments. *Scientific reports*, 12(1):13376, 2022.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL `https://arxiv.org/abs/2212.04356`.

Mark S Rea, MJ Ouellette, and ME Kennedy. Lighting and task parameters affecting posture, performance and subjective ratings. *Journal of the Illuminating Engineering Society*, 15(1):231–238, 1985.

Mark S Rea, Mariana G Figueiro, Andrew Bierman, and John D Bullough. Circadian light. *Journal of circadian rhythms*, 8(1):2, 2010.

Rolf Reber, Norbert Schwarz, and Piotr Winkielman. Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? *Personality and social psychology review*, 8(4):364–382, 2004.

Red Barrels. Outlast, 2013. URL `https://redbarrelsgames.com/games/outlast/`.

Harrison Ridley, Stuart Cunningham, John Darby, John Henry, and Richard Stocker. The affective audio dataset (aad) for non-musical, non-vocalized, audio emotion research. *IEEE Transactions on Affective Computing*, 2024.

Siobhan Rockcastle and Marilyne Andersen. Measuring the dynamics of contrast & daylight variability in architecture: A proof-of-concept methodology. *Building and Environment*, 81:320–333, 2014.

Siobhan Francois Rockcastle, Kynthia Chamilothori, and Marilyne Andersen. An experiment in virtual reality to measure daylight-driven interest in rendered architectural scenes. In *Proceedings of Building Simulation*, 2017.

Shaghayegh Roohi, Elisa D Mekler, Mikke Tavast, Tatu Blomqvist, and Perttu Hämäläinen. Recognizing emotional expression in game streams. In *Proceedings of the annual symposium on computer-human interaction in play*, pages 301–311, 2019.

David Rozado, Ruth Hughes, and Jamin Halberstadt. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. *Plos one*, 17(10): e0276367, 2022.

James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

Nicole Ruta, Stefano Mastandrea, Olivier Penacchio, Stefania Lamaddalena, and Giuseppe Bove. A comparison between preference judgments of curvature and sharpness in architectural façades. *Architectural Science Review*, 62(2):171–181, 2019.

Nicole Ruta, Javier Vañó, Robert Pepperell, Guido B Corradi, Erick G Chuquichambi, Carlos Rey, and Enric Munar. Preference for paintings is also affected by curvature. *Psychology of Aesthetics, Creativity, and the Arts*, 17(3):307, 2023.

Vera Sacharin, Katja Schlegel, and Klaus R Scherer. Geneva emotion wheel rating study. *Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences. Aalborg University, Aalborg*, 2012.

Elizabeth B-N Sanders, Eva Brandt, and Thomas Binder. A framework for organizing the tools and techniques of participatory design. In *Proceedings of the 11th biennial participatory design conference*, pages 195–198, 2010.

Sven Schneider, Saskia Kuliga, René Weiser, Olaf Kammler, and Ekaterina Fuchkina. Vreval-a bim-based framework for user-centered evaluation of complex buildings in virtual environments. *VR, AR & VISUALISATION| Explorations*, 2:833–842, 2018.

Sumit Sen, Anton Flyvholm, Emmanouil Xylakis, Stine Maria Louring Nielsen, Ellen Kathrine Hansen, Michael Finbarr Mullins, and Georgios Triantafyllidis. Towards assessing the impact of circadian lighting in elderly housing from a holistic perspective. In *ARCH17-The 3rd International Conference on Architecture, Research, Care and Health*, pages 227–240. Polyteknisk Boghandel og Forlag, 2017.

Vidhyasaharan Sethu, Emily Mower Provost, Julien Epps, Carlos Busso, Nicholas Cummins, and Shrikanth Narayanan. The ambiguous world of emotion representation. *arXiv preprint arXiv:1909.00360*, 2019.

Noor Shaker, Georgios N Yannakakis, and Julian Togelius. Crowdsourcing the aesthetics of platform games. *IEEE Transactions on Computational Intelligence and AI in Games*, 5(3):276–290, 2012.

Noor Shaker, Julian Togelius, and Mark J Nelson. Procedural content generation in games. 2016.

Avishag Shemesh, Ronen Talmon, Ofer Karp, Idazdan Amir, Moshe Bar, and Yasha Jacob Grobman. Affective response to architecture–investigating human reaction to spaces with different geometry. *Architectural Science Review*, 60(2):116–125, 2017.

Mel Slater and Maria V Sanchez-Vives. Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3:74, 2016.

Mel Slater et al. Measuring presence: A response to the witmer and singer presence questionnaire. *Presence: teleoperators and virtual environments*, 8(5):560–565, 1999.

Hyeonho Song, Kunwoo Park, and Meeyoung Cha. Finding epic moments in live content through deep learning on collective decisions. *EPJ Data Science*, 10(1):43, 2021.

Nataša Šprah and Mitja Košir. Daylight provision requirements according to en 17037 as a restriction for sustainable urban planning of residential developments. *Sustainability*, 12(1):315, 2019.

Alina Striner, Andrew M Webb, Jessica Hammer, and Amy Cook. Mapping design spaces for audience participation in game live streaming. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

Ana Tajadura-Jiménez, Pontus Larsson, Aleksander Väljamäe, Daniel Västfjäll, and Mendel Kleiner. When room size matters: acoustic influences on emotional responses to sounds. *Emotion*, 10(3):416, 2010.

Naoya Takahashi and Yuki Mitsufuji. Multi-scale multi-band densenets for audio source separation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 21–25, 2017.

Alexander Toet and Jan BF van Erp. The emojigrid as a tool to assess experienced and perceived emotions. *Psych*, 1(1):469–481, 2019.

Christopher W. Totten. *An Architectural Approach to level Design*. CRC Press, 2014.

Christopher W Totten. *Architectural Approach to Level Design*. CRC Press, 2019.

Oshin Vartanian, Gorka Navarrete, Anjan Chatterjee, Lars Brorson Fich, Helmut Leder, Cristián Modroño, Marcos Nadal, Nicolai Rostrup, and Martin Skov. Impact of contour on aesthetic judgments and approach-avoidance decisions in architecture. *Proceedings of the National Academy of Sciences*, 110 (Supplement 2):10446–10453, 2013.

Oshin Vartanian, Gorka Navarrete, Anjan Chatterjee, Lars Brorson Fich, Jose Luis Gonzalez-Mora, Helmut Leder, Cristián Modroño, Marcos Nadal, Nicolai Rostrup, and Martin Skov. Architectural design and the brain: Effects of ceiling height and perceived enclosure on beauty judgments and approach-avoidance decisions. *Journal of environmental psychology*, 41:10–18, 2015.

Oshin Vartanian, Gorka Navarrete, Anjan Chatterjee, Lars Brorson Fich, Helmut Leder, Cristián Modroño, Nicolai Rostrup, Martin Skov, Guido Corradi, and Marcos Nadal. Preference for curvilinear contour in interior architectural spaces: Evidence from experts and nonexperts. *Psychology of Aesthetics, Creativity, and the Arts*, 13(1):110, 2019.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Jan-Niklas Voigt-Antons, Eero Lehtonen, Andres Pinilla Palacios, Danish Ali, Tanja Kojic, and Sebastian Möller. Comparing emotional states induced by 360 videos via head-mounted display and computer screen. In *2020 twelfth international conference on quality of multimedia experience (QOMEX)*, pages 1–6. IEEE, 2020.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *arXiv preprint arXiv:2203.07378*, 2022.

WI Well. Well building standard. *International Well Building Institute: New York, NY, USA*, 2014.

Bob G Witmer and Michael J Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3):225–240, 1998.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. Elan: A professional framework for multimodality research. In *5th international conference on language resources and evaluation (LREC 2006)*, pages 1556–1559, 2006.

Tong Xue, Surjya Ghosh, Gangyi Ding, Abdallah El Ali, and Pablo Cesar. Designing real-time, continuous emotion annotation techniques for 360° vr videos. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI EA '20, 2020a. doi: 10.1145/3334480.3382895.

Tong Xue, Surjya Ghosh, Gangyi Ding, Abdallah El Ali, and Pablo Cesar. Designing real-time, continuous emotion annotation techniques for 360 vr videos. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020b.

Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar. Rcea-360vr: Real-time, continuous emotion annotation in 360 vr videos for collecting precise viewport-dependent ground truth labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

Qi Yang and Saleh Kalantari. Real-time continuous uncertainty annotation (rcua) for spatial navigation studies. *arXiv preprint arXiv:2207.14651*, 2022.

Qi Yang and Saleh Kalantari. Real-time continuous perceived uncertainty annotation for spatial navigation studies in buildings. *Journal of Building Engineering*, 82:108250, 2024.

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*, 2017.

Georgios N Yannakakis and Hector P Martinez. Grounding truth via ordinal annotation. In *2015 international conference on affective computing and intelligent interaction (ACII)*, pages 574–580. IEEE, 2015.

Georgios N Yannakakis and Héctor P Martínez. Ratings are overrated! *Frontiers in ICT*, 2:13, 2015.

Georgios N Yannakakis and Julian Togelius. Experience-driven procedural content generation. *IEEE Transactions on Affective Computing*, 2(3):147–161, 2011.

Georgios N Yannakakis and Julian Togelius. *Artificial intelligence and games*, volume 2. Springer, 2018.

Georgios N Yannakakis, Roddy Cowie, and Carlos Busso. The ordinal nature of emotions: An emerging approach. *IEEE Transactions on Affective Computing*, 2018.

Robert B Zajonc. Attitudinal effects of mere exposure. *Journal of personality and social psychology*, 9(2p2):1, 1968.

Tianyi Zhang, Abdallah El Ali, Chen Wang, Alan Hanjalic, and Pablo Cesar. Rcea: Real-time, continuous emotion annotation for collecting precise mobile video ground truth labels. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2020.

Zhihui Zhang, Josep M Fort, Lluis Giménez Mateu, and Yuwei Chi. Uncovering the connection between ceiling height and emotional reactions in art galleries with editable 360-degree vr panoramic scenes. *Frontiers in psychology*, 14:1284556, 2023.

Zengqun Zhao, Qingshan Liu, and Feng Zhou. Robust lightweight facial expression recognition network with label distribution training. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3510–3519, 2021.

Yangming Zhou and Guoping Qiu. Random forest for label ranking. *Expert systems with applications*, 112: 99–109, 2018.