Master of Science
Human-Computer Interaction

The Significance of Spatiotemporal Image Complexity on Gaze Dynamics in VR-Based 360º Video Interactions: An Integrated Oculesics and Computer Vision Approach

Rik Hazekamp
5936403

Utrecht University
Department of Information and Computing Sciences
Graduate School of Natural Sciences
Utrecht, The Netherlands

August 18th, 2023

This is a Master of Science thesis.

In accordance with the guidelines set forth by the university, this thesis is submitted as part of the requirements for the degree of Master of Science in Human-Computer Interaction. The degree of Master of Science comprises 120 ECTS credits (2 years of full-time studies). The HCI thesis project accounts for 40 ECTS credits.

The research presented in this thesis was carried out under the supervision of Dr. Wolfgang Hürst as the primary supervisor and Dr. Christof van Nimwegen as the second examiner.

Utrecht University
Department of Beta Sciences
Heidelberglaan 8
3508 TC Utrecht
The Netherlands

ABSTRACT

The immersive experience of 360-degree video in virtual reality headsets expands the limits of innovative human-computer interactions, necessitating a comprehensive understanding of the intricate interaction process for the optimisation of 360-degree content. Despite its significance to the user, current research has remained predominantly focused on the technical challenges, neglecting the impact of content-specific aspects on user behaviour. This thesis employs a content-aware approach to discern the independent role of spatial- and temporal properties of a 360-degree video sequence in shaping gaze behaviour. By employing spatiotemporal image complexity, this work reveals the intricate dynamics between content-specific attributes, gaze behaviour, cognitive perceptions and usability context. Instrumental to this thesis was the introduction and formulation of the quadrifactorial exploration index N_ψ, a novel metric to quantify complex gaze patterns. By integrating computer vision techniques and eye-tracking data, the index measures gaze patterns based on extent, intensity, variability and randomness. Moreover, an initial framework, utilising another unique metric δ, was devised to examine the confounding influence of diegetic artefacts in 360-degree videos on user gaze. Utilising a mixed-methods design, 52 participants were observed viewing six 360-degree videos with varying spatiotemporal image complexities via a head-mounted display, while seated on either fixed or rotating chairs. The interaction was captured utilising eye-tracking technology and subjective user evaluations. The results demonstrated a negative correlation between temporal complexity – the rate of visual change over time in consecutive frames – and the extent of user gaze. In contrast, spatial complexity – the level of visual richness in each frame – did not significantly impact the user gaze, seemingly attributed to underlying cognitive factors. An observed dichotomy between objective gaze metrics and subjective user experiences further emphasises the role of cognition in the observed spatial- and temporal effects on gaze behaviour. In addition, it was found that these effects were significantly moderated by the different seating types. The employment of oculesics, computer vision and user-centric evaluations revealed the autonomous significance of 360-degree video content within the multifaceted interaction model, while taking into account cognitive and usability influences. The insights provide a theoretical understanding of gaze dynamics during 360-degree video interaction, as well as carry significant implications for the development of more engaging 360-degree content and immersive VR environments.

Keywords: 360-degree video technology, eye-tracking, virtual reality, usability, human-computer interactions, spatiotemporal complexity, computer vision, gaze behaviour analysis, image segmentation, cognitive perceptions.

PREFACE

The writing of this thesis has been an incredibly valuable process of discovery, challenge and growth, both academically and personally. The study of complex theoretical constructs, exploration of the various research domains, and mastering of new techniques and methodologies to conduct insightful research have not only stimulated me on an intellectual level, but also significantly elevated my academic aspirations. This work merges my personal intrigue for innovative technology and the artistry of film with my academic ambitions in human-computer interaction, information science and computer sciences.

I would like to express my sincerest gratitude to my primary supervisor, Dr. Wolfgang Hürst, for his devoted supervision, invaluable insights and academic encouragement. His expertise, both as a thesis supervisor and as a professor in two of my graduate courses, has been fundamental to my theoretical understanding of the research domain. My appreciation is extended to Dr. Christof van Nimwegen for his dedication in both time and resources, notably the accessibility to the eye-tracking laboratory, and for his involvement in the project as second examiner. A special note of appreciation goes to Kiara Heide, Global Director CSM from iMotions A/S, for her remote assistance and commitment to the seamless integration and operation of the VR eye-tracking software modules.

As this thesis signifies the culmination of my academic efforts thus far, I would like to express my gratitude to the professors, students, and colleagues of the Graduate School of Natural Sciences at Utrecht University, with whom I have had the sincerest pleasure of working.

Lastly, I'm beyond grateful to my mother, father and brother, along with my family, friends and loved ones for their continued support and encouragement throughout.

Contents

List of Acronyms

AB	Actor Behaviour
ACR	Absolute Category Rating
AOI	Area of Interest
CDF	Cumulative Distribution Function
DAS	Diegetic Artefact Score
DCR	Degradation Category Rating
DIME	Distributed Interactive Multimedia Environments
DOF	Degrees of Freedom
DV	Dependent Variable
ECG	Electrocardiographic
EDA	Electrodermal Activity
EEG	Electroencephalographic
EMG	Electromyographic
EAC	Equi-Angular Cubemap
ERP	Equirectangular Projection
FOV	Field of Vision
FOMC	Fear of Missed Content
FR-IQA	Full-Reference Image Quality Assessment
GMT	Game and Media Technology
HCI	Human-Computer Interaction
HVS	Human Visual System
IQA	Image Quality Assessment
IVA	Intended Viewing Area
IV	Independent Variable
LMM	Linear Mixed-Effects Model
LSD	Latin Square Design
LTS	Long-Term Support
M-ACR	Modified Absolute Category Rating
MOS	Mean Opinion Score
MS-SSIM	Multi-Scale Structural Similarity Index Measure

OCR	Optical Character Recognition
OV	Omnidirectional Video
PI	Place Illusion
POV	Point of View
PCA	Principal Component Analysis
PSI	Plausibility Illusion
QOE	Quality of Experience
ROI	Region of Interest
RS-EEG	Resting-State Electroencephalogram
SC	Sensorimotor Contingency
SI	Spatial Information
SSI	Semi-Structured Interview
SSIM	Structural Similarity Index Measure
SSQ	Simulator Sickness Questionnaire
TI	Temporal Information
UEQ	User Evaluation Questionnaire
UX	User Experience
VIF	Variance Inflation Factors
VOD	Video On Demand
VP	Viewer Position
VR	Virtual Reality

List of Equations

List of Figures

List of Tables

Introduction

The emergence of virtual reality technology has introduced an entirely new digital landscape, evolving the way users interact with technology and establishing transformative advances in the way users engage with virtual environments. New forms of interactive media arise and manifest within the digital landscape of VR, signifying a paradigm shift in content consumption and immersive experiences. Enabled by the increasing accessibility to VR technology and distribution of omnidirectional content, the rising popularity of omnidirectional video interaction emerges as a key contributor to the field of human-computer interaction [2]. Unlike conventional two-dimensional video, 360-degree video enables a three-dimensional view in all possible directions, significantly immersing users in a multi-dimensional experience [64, 375]. The relatively novel 360-degree format poses a prominent area of research, as the complex interplay of multi-modal interaction, usability and cognition, coupled with the unique format and expansive boundaries of VR technology, require a continued and comprehensive understanding of the intricate interaction dynamics.

This new interaction paradigm governs the 360-degree medium, especially through the use of head-mounted displays [64, 275, 315]. The inherently complex interaction model, which integrates factors of technological intricacies, cognitive perceptions and content characteristics, elicits a range of unique challenges within the field of VR research [54, 356] – specifically, since the user experience of two-dimensional video interaction becomes multifaceted in the domain of 360-degree video interaction. The vital role of attention guidance in immersive experiences becomes even more complex, due to the unrestricted control and navigation of the user within the virtual 360-degree environment. As such, the user simultaneously adapts both roles of participant and observer, rendering traditional cinematographic techniques and attention guiding mechanisms ineffective and necessitating a greater level of understanding of the user interaction process in such immersive virtual environments [193, 223, 262, 306].

The novel nature of the 360-degree format presents challenges both in terms of technology and in gaining a comprehensive understanding of the intricate user experiences. However, despite its increased relevance, research on the 360-degree video interaction remains predominantly focused on the technical aspects, rather than on the adoption of more nuanced content-aware and user-centric approaches [375]. While technical challenges such as bandwidth, storage, and encoding requirements remain inherently important to the development of the format, the continued oversight of content-specific influence often leads to content-agnostic methodologies, neglecting the independent role of 360-degree content within the interaction process.

A comprehensive understanding of the dynamic influences of content-specific attributes, cognitive perceptions, and the usability context in which the medium is consumed is essential for the development of immersive and engaging virtual 360-degree environments, as the intricate interplay between the 360-degree video format and the user's underlying cognitive processes can carry behavioural consequences to the user's gaze and overall interaction experience [21, 290, 356]. Imperative to the understanding and development of immersive virtual 360-degree content is the examination of the user's gaze behaviour throughout the 360-degree video interaction process.

This thesis aims to bridge the current gap within the research domain, by approaching the multi-dimensional interaction model from a content-aware perspective, emphasising the autonomous and independent role of a 360-degree video sequence in the VR interaction process. Motivated by related literary works and advanced methodologies such as eye-tracking and computer vision techniques, this work primarily focuses on the complex dynamics between user interaction, cog-

nitive perceptions and usability context. As such, this work aims to provide a comprehensive understanding on how users navigate and engage with varying 360-degree video content in the virtual 360-degree environments.

Central to this thesis is the integration and holistic exploration of related fields of research. This work explores the fields of human-computer interaction, computer science, cognitive science, and film studies to approach the main research objective holistically. Primarily, this thesis focuses on the integration of oculesics and computer vision to elucidate the user's gaze behaviour during 360-degree video interactions. In particular, providing a systematic framework to analyse behavioural patterns by transforming the unique spatial- and temporal properties of 360-degree video sequences into quantifiable dimensions and integrating quantitative and qualitative methodologies to examine the dynamics of user gaze behaviour, cognition, cinematography and usability contexts [105, 356, 371].

In conclusion, by integrating oculesics, usability methodologies and computer vision techniques, this thesis aims to bridge the existing research gaps by considering the 360-degree video sequence as an autonomous and independent component of the multifaceted interaction model. This content-aware approach provides a nuanced understanding of how 360-degree content, in terms of space and time, influences the user interaction in VR. As such, this work not only aims to advance the theoretical discourse but also provide actionable insight for 360-degree content development.

This thesis adheres to the following structure. Chapter 1 explores the related work by examining the current existing body of literature from the fields of human-computer interaction, cognitive science, film studies and computer vision. The chapter presents the foundational scope of the thesis and explores relevant theoretical frameworks, such as omnidirectional video characteristics, 360-degree user interaction models, oculesics and gaze in VR, quantitative imagery analysis in computer vision and attention guidance in VR. Importantly, the chapter utilises the insights gathered from the related work to further refine, formulate and detail the main research objective and corresponding sub-questions. The research methodology of this thesis is presented in Chapter 2, detailing the comprehensive approach applied in this study. It presents an overview of the employed research design and integrated objective, subjective and physiological approaches. Furthermore, the chapter details the use of required materials, sampled participants and procedures, as well as the analytical framework and data analysis approaches. Chapter 3 is dedicated to the formulation and introduction of a novel metric to quantify complex gaze behaviour patterns. In this chapter, the integration of both extracted computer vision techniques and oculesics data to formulate a methodology is detailed. Similarly, Chapter 4 presents a novel methodology to explore and examine the relevance and confounding influence of cinematographic principles on gaze behaviour. Both chapters present the utilised principles of computer vision to propose novel metrics, instrumental to this thesis. In Chapter 5, the results of the conducted analyses are presented, adhering to the provided analytical framework. The quantitative and qualitative findings are discussed, of which the statistical analyses, regression models and subjective insights provide a comprehensive and multi-dimensional perspective of the user interaction. The research findings are discussed in Chapter 6. Furthermore, this chapter details the theoretical and practical implications of this work, discusses the limitations and proposes interesting future research directions. Lastly, a conclusion to this work is presented.

Chapter 1

Related Work

This chapter presents the current state of research in the domain of user interaction with 360-degree video in VR. By taking an holistic approach, this literature study explores the common research limitations of user behaviour and interaction with 360-degree video content, as well as includes related fields of oculesics, computer vision, cognition and film theory. The main research objective of this thesis encompasses the independent influence of spatiotemporal image complexity in 360-degree video sequences on user behaviour in VR. The related work, as presented in this chapter, provides an extensive overview of the relevant literature by:

- Introducing omnidirectional video as a format, including current research challenges and limitations occurring within the domain;
- Providing a construct of 360-degree video aspects and the complex user interaction, such as quality of experience, interaction layers and representative metrics;
- Discussing relevant objective and subjective assessment methods, such as physiological methodologies and oculesics to assess user behaviour;
- Presenting relevant mathematical constructs and imagery analysis techniques within the field of computer vision;
- And exploring the cognitive principles, attentional guidance mechanisms, user engagement factors and cinematographic concepts related to the perception of omnidirectional video content.

The literature study is structured as follows: section 1.1 introduces the concept of omnidirectional video. The characteristics, as well as the challenges and limitations of 360-degree video, are discussed in section 1.2. Section 1.3 entails the current literature on the user experience and interaction with 360-degree video content. Furthermore, section 1.4 elaborates on the existing research in eye-tracking as an instrument to research user gaze and analyse behavioural data in 360-degree video interactions. Subsequently, section 1.5 presents relevant quantitative imagery analysis techniques, derived from the field of computer vision research. In addition, section 1.6 encompasses the overall perception of omnidirectional content by discussing the perceptual differences between traditional and omnidirectional content, as well as discusses relevant cinematographic principles, addresses the use of visual attention guidance mechanisms and focuses on user engagement and immersion in VR environments. Lastly, the literature study is concluded in section 1.7, in which the main research objective and related sub-questions are formulated.

1.1 Omnidirectional Video

Omnidirectional content, i.e., 360-degree video, has seen an increase in popularity over the recent years, mainly due to the developments in Virtual Reality (VR) and the arrival of more interactive displaying systems, such as head-mounted displays (HMDs) and omnidirectional capturing systems [64]. The emergence of more wireless HMDs, such as the HTC Vive, Daydream and Oculus Rift, enable more users to interact with this relatively novel form of content. The head-mounted display (HMD) is a device that positions the displays in close proximity to the user's eyes, and through head-tracking technology, the left and right images are adjusted according to the user's movements within the virtual environment. The user experiences stereo vision due to both the left and right images, as well as experiences stereo audio transmitted through earphones. This immersive setup creates the illusion of navigating through a three-dimensional space with stationary and moving objects, possibly including representations of other people, whom can either be actual individuals in remote locations or virtual entities controlled entirely by a computer program. Additionally, users can interact with the virtual environment through hand-tracking technology to manipulate objects [275]. It is estimated that the VR market will entail 275 million users in 2025 [315]. This ongoing development and commercialisation of HMDs has ensured a substantial increase in virtual reality (VR) applications, interest and usage [8, 54, 375], allowing for the 360-degree video format to gain popularity [343]. Furthermore, the availability of commercial-grade omnidirectional cameras, like the Ricoh Theta X and Insta360, have lowered the entry-level

requirements of omnidirectional content creation. This increased interest in 360-degree video is also heavily influenced by the dominance of online video streaming [319, 320]. Video broadcasting platforms, such as YouTubeVR, support the distribution of 360-degree video, enabling users to easily create, consume and distribute omnidirectional content. As a format, 360-degree video can be experiences utilising a variety of interactive displaying systems such as HMDs and mobile viewports, the latter of which are most commonly achieved through a smartphone or tablet. The combination of increased levels of physiological-tracking, stereoscopic visuals and wide FOV displays, achieved through use of HMDs, enable higher levels of immersion among users compared to other delivery systems [73]. Notably, users tend to respond to virtual stimuli in similar ways as real-world settings due to the higher levels of immersion. An example of this was found by Wilcox et al. (2006), which demonstrates that invasion of virtual personal space evokes similar responses as in real-life settings [13, 352].

Contrary to traditional and two-dimensional video, 360-degree video captures a complete 360-degree view from wherever the camera is positioned and internally projects it onto a spherical surface [375]. The user's field of vision (FOV) is determined by the viewport, which is bound to the physical limitations of the displaying device or the user's head orientation. The FOV is expressed in degrees, which represent the amount of visual angle, and is manipulated by the user through mouse / finger controls or head movements in a custom video player. There are multiple devices that act as a viewport, the most common of which are the HMDs, mobile viewports and static 2D displays. The viewport can be seen as a physical or digital "keyhole" from which the user perceives the omnidirectional content. As such, the user's viewport relies on the user's head orientation when using a HMD to view a 360-degree video. To display high-quality 360-degree video, the entire 360-degree frame must support high-quality resolution.

The 360-degree video is projected as a spherical video, captured in all directions from a single point. Recording setups usually consist of 2-6 separate camera's, composited in a cubic formation. Using specialised software, the video files are synchronised and combined into a singular spherical view. As part of the end-to-end 360-degree video streaming network, capturing and mapping plays a vital role. After the mapping process, the file is then encoded, and the entire package is transmitted before it is decoded and played [362]. Traditional video encoders require two-dimensional imagery, and as such, it is vital to convert the spherical video onto a two-dimensional plane to enable encoding and transmission [375]. This process of mapping, projecting a sphere onto a plane, is achieved using a variety of methods [64, 215]. As identified by Yu et al. (2015), the most applied projection methods for 360-degree video encoding are: equirectangular, cubemap, pyramid and dodecahedron projection [366]. Equirectangular projection (ERP) is considered the most adopted projection method of omnidirectional video. During ERP, the spherical image is projected onto a rectangle, during which both poles of the spheres are stretched to fit the rectangular shape [115, 206, 370, 375]. The resulting video is compressed into a preferred format, such as H.264, and uploaded to a server with metadata indicating the 360-degree format and any special processing requirements. The server can then convert the video into a streaming format, such as MPEG-DASH, before making it available for playback.

1.2 360-degree Video Characteristics

Omnidirectional video content can be accessed through a variety of different content providers and streaming platforms. While many 360-degree video content providers are brand-specific, offered through native HMD apps such as the HTC Viveport or Vuze+, there are also providers and services available which are less platform-dependent and that run on multiple platforms through desktop players (i.e., YouTube VR). User-generated content offers easy access to create and consume 360-degree video content and is therefore a very popular method of interacting with 360-degree video content. Despite the existence of professional content providers (e.g. NBC), user-generated content platforms (e.g. YouTube VR and Facebook) remain most accessible and popular among users. In a study by Afzal et al. (2017), the 360-degree video library of YouTube was characterised by utilising the aggregate statistics, bit-rate, resolution and duration [2]. By

analysing the dataset and characteristics of 360-degree video from their study, it is possible to extract distinct categories in which the majority of the 360-degree video library can be categorised into. The dataset encompasses a total of 2285 360-degree videos and, through calculating term frequency, were mapped into one of the following identified categories: animals, cartoon, concert, documentary, driving, horror, movie trailer, roller coaster, scenery, shark, skydiving, space, sports and video game [2]. Analysis of video duration across categories indicate a higher mean duration of 360-degree videos within the categories of concert, driving and documentary, and a lower mean duration within the categories of roller coaster and movie trailer. Across the entire dataset, it was found that 360-degree videos contain a mean duration of 143 seconds, which is comparatively shorter than the non-360-degree videos with a mean duration of 490.5 seconds. Comparing the duration CDF plots shows a shorter tail of the distribution of 360-degree video duration. The relative novelty of the 360-degree video format remains one of its biggest limitations. The current state of 360-degree video development is still quite experimental, resulting in the distribution of shorter clips to determine viewer experience and perceptibility. Furthermore, longer videos increase levels of viewer fatigue and cybersickness, especially when viewing through HMDs. In terms of resolution, it was found that 360-degree videos have higher maximum resolutions (i.e., 8192x8192) when compared to traditional videos. However, both formats share similar minimum resolution of 82x144. This correlates to a higher bit-rate for 360-degree video, yet when compared using effective resolution [2] are quite similar in terms of bit-rate. As before-mentioned, the study by Afzal et al. (2017) also examined the motion characteristics of 360-degree video. Motion in 360-degree is of great importance in terms of bit-rate and file transmission (i.e., bandwidth requirements). Interestingly, their work hypothesised that reduced motion in 360-degree videos lead to lower bit-rate variability of the 360-degree videos compared to traditional videos. This is because the motion in 360-degree video remains an intrinsic part of the scene, rather than being caused by camera panning or rotation. As such, the camera acts as a static element in the scene which captures the view in all directions from a single-point perspective. Consequently, in 360-degree video interactions, the user is able to alter the video angle while watching it, which incites more network responsiveness to adopt the changes in field of view. The lower bit-rate variability can be explained by the notion that traditional video contains video during which the camera moves or the shifts in perspective. The camera motion in traditional video is therefore higher, thus requiring higher variability. In addition, the categories that are characterised as "high motion", according to Afzal et al. (2017), are: driving, skydiving, sports and roller coaster. Other categories such as cartoon, video game, movie trailer and horror vary significantly in level of motion. This effect is usually caused by the content developer, where cinematographic rules and stylistic preferences dictate the viewer's attention and scene direction.

It is important to distinguish specific terminology when it comes to 360-degree video. Omnidirectional content refers to content that can be viewed from multiple directions, often times including panoramic images or 360-degree video. Traditional or conventional video refers to the standard, two-dimensional video content that is perceived as flat or planar. These videos can only be viewed from one direction and the terminology flat, planar, 2D, traditional or conventional all refer to this type of video and will be used interchangeably throughout this thesis. The terms 360-degree video and omnidirectional video are expressed similarly.

Omnidirectional video (OV) provides interesting user experiences and interaction. However, due to the relatively novel nature of 360-degree video consumption, there exist various challenges that are unique to 360-degree degree video format. Many of which are technical limitations, tightly-bound to available delivery systems and quality thereof [375]. Furthermore, streaming 360-degree video in a high resolution requires ultra-high bandwidth and large storage capabilities. However, for 360-degree video to gain more popularity as a standardised form of content consumption, there are more challenges that ought to be studied and addressed [375]. As identified by Zink et al. (2019), most challenges and limitations within 360-degree video can be categorised under the following set:

- Ultra-high bandwidth requirements;
- Ultra-large storage requirements;

- Ultra-low motion-to-photon delay;
- Complex view adaptation;
- Complex rules and metadata for viewing the videos;
- Video Quality of Experience (QoE).

Approaches to solve these technical challenges is a complex area of research, and ongoing studies on the effectivity of tile-based approaches on common scalability issues are promising [111, 366]. However not mutually exclusive, these limitations are of influence to each other. Subjective and objective quality of experience are commonly affected by factors in technical limitations, such as video coding, projection schemes, geometric distortion and 360-degree stream quality, which limit the overall performance of the system [2, 54, 185]. The quality of experience (QoE) in 360-degree video is a complex problem that requires content-aware metrics, and analysis thereof remains relatively unexplored. As identified by Ebrahimi et al. (2009), the following influential factors encompass multimedia QoE [90]:

- System Influence: technical factors (i.e., device, network, or content format)
- Human / User Influence: task application factors (i.e., usability context)
- Context Influence: social and psychological factors (i.e., environment, user expectations and content)

Current research remains predominantly focused on traditional video content, with majority of QoE studies related to omnidirectional video specifically targeting the technical factors and parameters, such as impact of target bit-rate, encoding schemes and video stream quality [109, 132, 151, 174, 282]. Due to the higher levels of interactivity during omnidirectional video interaction, user experiences are less predictable and can be different every time. Furthermore, the user-tracked head movements, when using a HMD, requires for the content to be updated in real-time. Any form of latency severely affects the immersive adoption of users and negatively impacts the exposure to rich media content in an immersive environment [362]. The ability to change viewpoints allows the user to decide and alter their viewing direction whenever they see fit, allowing for high control over their viewing experience. In addition, the context influential factors that impact 360-degree video QoE, such as the effect of content, environment and user expectations, remain underrepresented in current literature, providing a very interesting and promising area of research.

1.3 User Experience and Interaction with 360-Degree Video

Evidently, 360-degree video format offers unique and various forms of interaction. The interactive nature of virtual reality through head-mounted displays allows for the user to interact with and experience a virtual environment in a way that is otherwise difficult to achieve with traditional modalities, such as a smartphone or desktop computer. As a result, this advanced form of interaction composes complex user experiences. As described by Tran et al. (2017), understanding 360-degree video user experience is very complex and proves to be a huge challenge [328]. Traditionally, quality of experience (QoE) is a common method to study and measure user experience, specifically when interacting with video content. Moreover, when viewing planar (2D) video-content, QoE has been proven to be an effective tool to measure subjective viewer experience [54].

1.3.1 Quality of Experience in 360-Degree Video

As a concept, Quality of Experience has acquired interest from different areas of research over the past few decades. Mainly due to the consensus that the objective metrics used in Quality of Service (QoS) analyses, which has been the de facto standard in quality analysis for a long time, were not sufficient enough. That is, Quality of service does not include user-specific preferences and lacks the ability to express overall subjectivity in the assessment. While QoS assessment focuses

on the performance of a system, QoE is more centred around the user's assessment of this system, influenced by user preference and expectations. The gap between human-centric evaluations and system assessment is a common pitfall for systems with a high QoS, leading to failure in terms of user adoption [76]. This is where QoE as a concept has gained relevance, specially within the domain of human-computer interfaces and interactions. However, despite the similarities, QoE is distinct from user experience (UX), which is centred around "studying, designing and evaluating the experiences of a user when using a system" [264]. Quality of experience shares these characteristics, while also extending its analysis to include the content itself as part of the system interaction [42]. As identified by Brunnström et al. (2013), many dedicated studies failed to provide a consistent view on the concept of QoE [153, 212, 255]. Their paper defines QoE as a relevant and applicable concept within most domains. However, this thesis will entail the concept of Quality of Experience, as defined by Wu et al. (2009):

> QoE is a multi-dimensional construct of perceptions and behaviours of a user, which represents his / her emotional, cognitive, and behavioural responses, both subjective and objective, while using a system [356].

One of the major influential factors in QoE is its reliance on context of use, which is often determined by the application domain. Content modalities, such as delivery (broadcasting, streaming) of content (video, audio), educational and medical, or collaborative applications, all require different requirement configurations that take into account the user behaviour in both on- and offline behaviour as well as the levels of interactivity [42]. Application areas such as multimedia learning, sensory experiences, haptic communications and cloud-computing all require different metrics, ranging from unidirectional to bi- / multi-directional services [128, 154, 204, 308, 323]. The overarching QoE features can be categorised, classified and assessed on four distinct levels [42]:

- Level of direct perception: refers to perceptual information created during media consumption, i.e., space, motion, colour, darkness and distortion (video).
- Level of interaction: refers to the human-to-human and human-to-machine interactions i.e., responsiveness, naturalness of interaction and communication effectiveness.
- Level of usage situation: focused on accessibility and stability of the service or application i.e., the physical and social situation.
- Level of service: relating to the usage of a service beyond i.e., joy, usefulness and ease of use.

In terms of 360-degree video QoE, this chapter focuses on the levels of direct perception, interaction and usage situation, dealing with user assessment of QoE features such as involvement, motion, and actions.

1.3.2 QoE Metrics of 360-Degree Video

As a form of multimedia, 360-degree video QoE can be modelled best as a 'multi-dimensional construct of user perceptions and behaviours' [356]. The study by Wu et al. (2009) proposes a theoretical framework for modelling QoE in Distributed Interactive Multimedia Environments (DIMEs), which introduces a QoE model that is practically generalisable for DIME performance assessments. The key characteristic of omnidirectional video content is its interactivity, where users can interact with the virtual 360-degree environment. DIMEs use multi-modality media to connect different users into a joint interactive space for collaboration, characterised by three key roles: executant of tasks, user of technology and participant in group telecommunication. While the latter is not relevant for 360-degree video, the theoretical framework presented in the study by Wu et al. (2009) still presents a significant framework that is applicable to 360-degree video in terms of task execution and technology users. The framework consist of a set of representative dimensions, defined as cognitive perceptions and behavioural consequences, respectively. Cognitive

perceptions can be categorised using three sub-dimensions: flow, perceived technology acceptance and telepresence. This chapter only elaborates on flow and perceived technology acceptance, as telepresence is not relevant to 360-degree video within the scope of this thesis. The perception of flow functions as a main intrinsic motivator that drives people to perform certain tasks, without the promise of a reward [70, 356]. Common metrics to measure flow , i.e., clear goals, feedback, concentration, distorted sense of time and intrinsic enjoyment, [71] are often defined too broadly and lack applicability to multi-modality media. Wu et al. (2009) proposes three metrics that are significantly relevant and applicable to DIMEs: sense of control, concentration, and intrinsic enjoyment. While flow metrics refer to the psychological experience of users, the perceived acceptance of technology metrics aim to measure the user's perception and attitude towards a system [356]. Using the Technology Acceptance Model (TAM) [75] as foundation, the two proposed metrics for measuring acceptance of technology are perceived usefulness and perceived ease of use. As defined in their study, the behavioural consequences are a subsequent result of the cognitive perceptions described above. The identified subdomains that are associated with behavioural consequences are performance gains, technology adoption and exploratory behaviours, each encompassing a variety of both subjective and objective metrics. Common metrics used to asses performance gains are closely related to the user's performance, both subjectively and objectively, which rely on the actual task application and requirements. Examples of these metrics are completion time and ratio of successful attempts [252]. Assessing technology adoption can be achieved by utilising both subjective metrics, such as intention to use [133, 146, 229], and objective metrics, such as actual usage. For the latter of which, a longitudinal study is highly recommended [335]. Lastly, the metrics identified for the assessment of exploratory behaviours are application-specific as well. Common metrics used are both subjective and objective ones and can be defined utilising specific experiential statements regarding the spontaneous exploration of the system [226, 356].

1.3.3 Subjective QoE Assessment of Omnidirectional Video

Identified by the International Telecommunication Union (1997), common methodologies for subjective QoE assessment are the Absolute Category Rating (ACR) and Degradation Category Rating (DCR) scoring methods [144]. ACR is usually applied to tests that are predominantly focused on qualification, where a single stimulus is presented in a way that is very representative of every day usage of the technology at hand. Also known as the Single Stimulus method, ACR is mostly tested in a setting where the tests are presented in sequence and rated independently of each other using a five-level scale. The method presents the stimulus one at a time, usually with a duration of approximately 10s, after which the subject is asked to evaluate the quality of the entire sequence. For this, a voting duration of less or equal to 10s is recommended [144]. Notably,, this method does not test transparency or fidelity usually. For testing the fidelity or transparency, DCR would be more suited, as it focuses more on the objective quality of a system. The DCR method, also known as the Double Stimulus Scale method, requires paired test sequences. Applying these methodologies to omnidirectional video often requires adaptation of some sort, since the methodologies were developed for short, 2D video. The omnidirectional nature of 360-degree video and typical longer-length viewing sessions on a HMD require adaptation of these current methodologies. The Modified Absolute Category Rating (M-ACR) is an effective modified method for QoE assessment, applicable to omnidirectional video [233, 290]. Contrary to ACR, the M-ACR method presents two sequences twice, enriched with an approximate 8s "blank" sequence in between and a voting time of less or equal to 20s. This double presentation of a sequence is due to the priming effect of the subject, users are not adapted to viewing 360-degree video on a regular basis in most cases. The first presented sequence serves as a primer, allowing the subject to be acquainted with omnidirectional video content. By doing so, the rating validity is much higher when the sequence is presented the second time. Another subjective quality evaluation method for omnidirectional video is the Double-Stimulus Impairment Scale (DSIS). However, DSIS is less reliable than M-ACR and is prone to higher levels of cybersickness [144, 293]. As implemented by Singla et al. (2017), the M-ACR method was used to assess how different quantisation and resolution are evaluated by users in terms of their perceptual quality [290].

The approaches for subjective QoE assessment, as defined in current literature, are mostly studied with respect to perception [105, 371], presence [137, 376], cybersickness [291], usability [277, 292] and sensor-based [91, 271] aspects of QoE. This emergence of sensor-based metrics, which utilise physiological data obtained from users, is an intriguing advancement that blurs the dichotomy between subjective and objective measures [91, 271]. Fore example, the study by Egan et al. (2016) utilises a Fitbit heart rate monitor and a PIP biosensor to objectively measure user QoE in terms of heart rate and electrodermal activity (EDA). As such, it is the first work to focus predominantly on the correlation between the objective metrics and user QoE. The findings of their study indicate a correlation between the objective EDA measures and subjective self-reported measures acquired through a post-experience questionnaire. The results from the study by Egan et al. (2016) also demonstrated higher QoE-values when viewers were using HMDs as compared to utilising conventional displaying systems (e.g. mobile viewports), which is in line with the findings from Tran et al. (2017) [328].

Furthermore, the assessment of QoE for 360-degree video can encompass both technical aspects as well as content characteristics. The continued work on QoE evaluation focuses mainly on video [354] and / or audio [260] quality perception. However, the advancements made in multimedia technology, such as VR and omnidirectional video, require more complex approaches. The objective metrics from video and audio quality are under-performing as a valid metric to assess perceptual quality, as the post-test subjective evaluation is heavily influenced by user preferences and external conditions during assessment [9]. Various methodologies have been developed to evaluate visual quality by using physiological measurements based on electroencephalographic (EEG), electrocardiographic (ECG) and electromyographic (EMG) signals, which indicated high levels of correlation with MOS. However, these methods rely on the premise that brain activity signals hold potential as valid metrics for multimedia QoE assessment [10, 92, 280]. Another interesting approach to evaluate QoE through physiological sensors was made through implementing eye-tracking technology [10]. As demonstrated by Egan et al. (2016) and Arndt et al. (2014), utilising physiological sensors to assess subjective QoE metrics and their correlation presents a very promising area of research [10, 91].

As before-mentioned, 360-degree video quality is heavily influenced by a variety of (technical) parameters, such as low latency or bit-rate variability. However, while very complex, these parameters and their impact on the user interaction in VR have been widely represented in current literature. Maintaining a strong immersive and engaging environment for 360-degree video is challenging for a variety of factors, not just because of the technical implementation thereof. The before-mentioned challenges within the field of 360-degree video research by Zink et al. (2019) highlights complex view adaptation alongside QoE and other challenges, as well as common limitations [375]. Similar to, and further accentuating these findings, Yaqoob et al. (2020) proposes a total of 4 additional key challenges, required for maintaining an immersive and engaging environment: 360-degree live streaming, low latency streaming, Quality of Experience (QoE) and viewport prediction [362].

The prior sections focused primarily on the limitations and challenging aspects of QoE measurement in 360-degree video. This section focuses more in-depth on the limitation of complex view adaptation, in particular viewport prediction. Head-mounted displays (HMDs) make use of the device's gyroscopic values to precisely accumulate and respond to changes in the user's head-movement. These detections are used to correctly translate and render the virtual environment to match that of the user's movement. By doing so, the user experiences increased levels of immersion and presence, further emphasising the importance of low latency and rendering quality. It is crucial to the 360-degree video experience that HMDs correctly identify and process sensor-based interaction signals to achieve an accurate visual representation in the user's viewport. As previously defined, a viewport is the portion of the omnidirectional virtual environment which is rendered and displayed through the stereoscopic lenses on the head-mounted displays. Accurate viewport rendering is not only essential for enhancing the sense of immersion, but attaining higher levels of realism also reduces cybersickness, increases presence and elevates the overall user experience. Therefore, viewport prediction is a common method to achieve this. By utilising the HMD's positioning sensors and gyroscopic values, the virtual environments is correctly rendered

based on the viewing behaviour and orientation of the user. Accurate viewport prediction can be achieved using either:

- Content-Agnostic Approaches, by predicting future viewport positions based on previous user behaviour, or
- Content-Aware Approaches, by predicting future viewport positions based on the video content itself.

Common content-agnostic approaches make use of linear regression models, clustering, encoder-decoder architecture and machine learning to optimise viewport prediction [22, 85, 136, 203, 247, 365]. The techniques make use of tile-based systems, which utilises the equirectangular frames and splits them into three distinct regions: viewport, adjacent and outside. The extrapolation of anterior watch-history allows these models to accurately predict future fixation points and salient regions. While some of these approaches are highly accurate, the labile nature of user behaviour requires extensive training for these models to achieve high efficiency.

Content-Aware Approaches As introduced, an alternative approach to optimise viewport prediction is by anticipating specific viewer behaviour when presented with specific virtual content. These approaches make use of the content-specific characteristics (i.e., visual features) to generate viewport predictions. A major contributor to the effectiveness of content-aware approaches is the use of saliency maps. Most current methods generate significant predictions by analysing saliency patterns and positional information, acquired through the HMD sensor features [217, 358, 359]. Common limitations of these saliency-based models are their dependency on predictor models and the exclusion of user behaviour, as evident in the study by Aladagli et al. (2017). Their work in particular, in which the user's viewing behaviour was not considered, probes the importance of understanding the user's unique visual attention [3]. An alternative approach is the integration of motion maps to make estimates on future fixation points. As such, the motion maps can be employed to account for the influence of cinematographic principles present in 360-degree videos, such as diegetic mechanisms that guide the user's visual attentions to objects in motion [97].

1.3.4 Cybersickness and Spatial Presence

There exists a need for specifying QoE as much as possible in regards to the specific application it will be assessing. As described above, the QoE assessment of a system or service contains many influential factors at play. Even though the framework presented by Wu et al. (2009) is relatively dated, it is still applicable to the multimedia modality of virtual reality and therefore, could be considered as serving guidelines rather than an obligatory system. However, within the domain of 360-degree video, there are more influential factors at play. Aspects such as cybersickness [137, 158] and presence [91, 307] ought to be equally considered in the evaluation of QoE in 360-degree video interactions. Cybersickness occurs when users experience dizziness or nausea due to the discrepancy between their own physical movements and the relative motion of the virtual scene, as displayed in the HMD [126, 138, 166]. In addition, the term presence describes the experience of feeling present in a digital world [156, 192]. Similarly, as before-mentioned, current studies on cybersickness and presence also focus primarily on the influence of technical aspects of omnidirectional video streaming. In the work of Zou et al. (2018), a framework for assessing spatial presence of omnidirectional video within VR was presented [376]. The three-layer hierarchical structure was layered from the bottom up as follows: technical influencing factors layer, perception layer and spatial presence layer. Additionally, the user's perception is multi-dimensional and entails the following dimensions: visual, auditory and interactive [94, 152, 337]. Due to the significant impact of the technological parameters of a VR system on the objective level of sensory realism [297], each of the technical influencing factors (i.e., video bit-rate, resolution, FOV, audio sampling rate) was linked to one of the following dimensions in the perception layer: video quality, audio quality, visual realism, acoustic realism, proprioceptive matching and spatial presence. The findings suggest a linear relationship between traditional

video and omnidirectional video, indicating that traditional video quality metrics are sufficiently applicable to assess omnidirectional video quality, even when provided through a HMD. Their results also indicate a linear relationship between video and audio quality in terms of visual and acoustic realism. Lastly, high resolution content was found to improve the overall user experience.

A more recent study of Tran et al. (2017) evaluates the different QoE aspects relevant to 360-degree video, and takes into consideration factors of encoding parameters, viewing modes, rendering devices and content characteristics [328]. Their results were analysed utilising the previously presented Absolute Category Rating (ACR), which is considered the standard for quality assessment [54, 144]. By measuring the Mean Opinion Score (MOS), where users provided subjective measurements using Likert scales, their work presented interesting findings. Firstly, Tran et al. (2017) found that when using HMDs instead of mobile viewports for viewing 360-degree video, the perceived quality and presence scores were higher on average. Secondly, their results further indicate that the viewer's sense of presence is significantly affected by content-specific characteristics, including camera motion, which is already strongly correlated with the sense of presence and cybersickness. Videos with moderate amount of camera motion received the highest presence scores, while videos with fast amount of camera motion scored lower and increased the feeling of cybersickness. This poses a crucial problem, specifically for 360-degree videos containing high levels of camera motion. As such, it is imperative to the improvement of the viewing experience to reduce the risk of cybersickness by taking into account the temporal influences of the 360-degree video sequences [328].

Utilising techniques to reduce cybersickness is complex. Notably, the before-mentioned techniques of viewport prediction, commonly utilised to employ foveated rendering techniques, can also be used to reduce cybersickness [145, 237]. However, viewport prediction remains a challenging technicality and field of research, despite the extensive research on detecting saliency. Current methods such as motion-based saliency estimation and user behaviour modelling are promising [101]. However, it is evident there remains plenty of areas for continued development in terms of long-term viewport prediction. One of these areas is the inclusion of the user's unique visual attention by further studying gaze (i.e., through saliency mapping) in regards to visual features, content characteristics and their independent influence on viewing behaviour.

1.4 Eye-Tracking and Visual Gaze Patterns in VR

The study of eye behaviour – oculesics – poses an important area of research, as the study of eye movement holds significant implications for many domains. It allows for communication as well as interaction, functioning as an important tool for perception [5]. Oculesics behavioural studies enable higher levels of understanding on human perception, complex non-verbal communication channels and deeper levels of interaction. Existing literature focuses on a variety of related aspects such as winking, blinking and eyebrow movement. However, this section focuses primarily on the specific eye-related aspects, such as pupil dilation, pupil position, gaze direction and gaze position. A quintessential tool in oculesics research is the integration of eye-tracking technology.

1.4.1 Eye-Tracking

Eye-tracking is a method often used for the allocation of visual attention, through recording eye-motion and gaze location (eye-focus) during varying activities. With eye-tracking technology becoming more accessible, it poses an established tool in many research domains and in real-world scenario training [53, 87, 142, 143, 228, 278, 349]. The foundations of eye-tracking are based on Charles Bell's discovery, whom demonstrated the physiological connection between neurological- and cognitive processes and the movement of a person's eyes [24, 339]. While not absolute, the eyes reflect the mental processing of a person to some extent [6, 213, 243, 256, 309]. Therefore, eye-tracking is considered a significantly valuable tool that enables exploration and insight into underlying cognitive processes. Moreover, the human's inability to remember or perceive involuntary eye movement raises the importance of measuring eye movement, enabling insight in subconscious

control and involuntary behavioural responses [58, 169]. Throughout the past century, many attempts have been made to objectively measure eye movement. However, these methodologies were mostly limited by technology, financially challenging and ethically questionable [45, 77, 363]. Current technology enables more user friendly and affordable solutions. As described by Carter et al. (2020), most video-based eye-trackers enable highly accurate gaze detection by measuring the corneal reflection of an infrared light relative to the pupil [49]. The (infrared) light is projected on the eye, which produces a reflection on the cornea which can be identified by the eye-tracking software. Many eye-tracking systems require a direct perpendicular orientation between the light sensor and the user's eye, necessitating a controlled environment. Furthermore, gaze calibration is necessary for higher degrees of accuracy, prompting the user to look at a series of points on the screen to enable the software to take baseline measurements. A brief elaboration on the physiology of eye movement is provided in the following subsection.

1.4.2 Physiology of Eye Movements

Similar to a camera with an aperture, lens and photosensitive image sensor, the eye gathers and transduces light through the pupil and focuses it on the retina, utilising the cornea and the lens [245, 339]. The fovea centralis, a small area in the centre of the retina, is responsible for detailed and colour vision through its high concentration of colour sensitive photoreceptors (cones) [170]. The rest of the retina, including the parafovea and periphery, is less sensitive to detail and colour [253, 254]. Visual information is sent to the brain through the optic nerve and then processed in different areas of the cortex for interpretation and reaction. This visual information is processed, as soon as the eyes fixate on a single target. Due to the smaller fovea, the eyes move more frequently to acquire high quality information from the entire visual field, resulting in shorter fixations [253]. The movement between fixations is called a saccade, during which visual input is suppressed. This occurrence renders our vision relatively blind [44, 50, 261]. Other various involuntary movements can occur during fixations, such as tremor, drift and microsaccades [86, 173]. Furthermore, ocular motions can be made deliberately (i.e., smooth pursuit and vergence) or are reflexive (i.e., optokinetic response [80] and vestibulo-ocular [125]).

A distinction, as defined by Duchowski et al. (2017), can be made in the applicability of eye-tracking technology. Diagnostic eye-tracking studies make use of the participant's gaze to determine duration and viewing order, by recording the eye-position throughout. This method is particularly effective in studies that rely on visual stimuli as an important variable (i.e., faces, scenes, text, video and web pages) and is therefore mostly adopted [86]. Interactive eye-tracking studies are less common, however they focus primarily on the high temporal- and spatial sensitivity of eye-trackers to use gaze position as an input source to generate preprogrammed responses. An example of this is by applying display changes based on user gaze position (e.g. revealing a picture on screen after an x amount of time has passed during the fixation). Some VR applications that rely on eye-tracking make use of foveated rendering, utilising the position of the eye to ensure high quality rendering of the exact area the user is looking at, increasing performance by reducing the render quality of peripheral areas [237]. However, the majority of eye-tracking studies are performed within the field of psychology (58.13%) [49]. Regardless, eye-tracking is applicable within most research domains. Notably, eye-tracking studies are less represented in the field of mathematics and computer science with only 13.51% of eye-tracking publications pertaining to the field. In the technology field, this percentile is a mere 11.89%.

1.4.3 User Behaviour Analysis and Gaze Tracking

Common practice for system evaluation is through analysing gaze to determine the user's visual attention [219]. By measuring and aggregating gaze data from the user, it is possible to extract valuable insights into viewing behaviour. However, analysis of gaze behaviour remains a complex and challenging task. A generalisable metrics framework, that can be applied to most studies, remains limited within the current body of literature. Therefore, many studies which involve gaze behaviour and gaze patterns as a substantial factor develop specific metrics only applicable to

their specific study parameters. Common practice of analysing content characteristics to measure changes in gaze patterns is through the use of Regions / Points of Interest (ROIs / POIs). For example, Serrano et al. (2017) established an important parameter to achieve this: by measuring the degree of (mis)alignment of regions of interest (ROIs) [281]. The ROIs were defined as the areas in the 360-degree frame in which the action takes place (i.e., a character or event). Their use of calculating the degree of (mis)alignment, baseline measurements, scanpath errors and use of frame-paths enabled a very targeted approach. By identifying and highlighting ROIs as metrics, their work was able to study the influence of continuity edits on user behaviour in VR. Their findings suggest significant changes in gaze behaviour based on the positioning and (mis)alignment of ROIs. Another approach, made by Singla et al. (2017) and Bao et al. (2016), utilised the gyroscopic data from the HMD to analyse user behaviour, using the pitch-, roll- and yaw-values to calculate the absolute difference between two of these values [21, 290]. The analysis on gaze behaviour by Singla et al. (2017) indicates that video quality has no significant impact on exploratory behaviour. Moreover, due to content-specific characteristics, it was found that different types of content can evoke different pitch-values, i.e., some contents have higher pitch-values than yaw-value. Rewatchability proved an interesting aspect in their study, where participants showed different behavioural patterns when rewatching the same video. Lastly, Simone et al. (2006) used the HMD to evaluate the objective and subjective measures of HMD performance and self-reported user ratings [289]. Their study demonstrated an effective method to evaluate the relationship between objective sensor metrics and self-reported subjective post-hoc ratings, utilising sensor-based data to analyse user behaviour.

Visualisation of gaze data is done through aggregated plots, statistical graphs and heatmaps [29, 88, 195]. Another visualisation technique, contrary to the aggregated plots including temporal information, is the use of timelines and scan path visualisations [225]. While studying user behaviour without gaze tracking is achievable [15], utilising the eye-tracking data offers an extra dimension to the analysis of user behaviour when interacting with the three-dimensional scene, enabling insight into the user's perception and underlying cognitive processes. Analysis of 360-degree video is predominantly based on heatmap visualisations which further require subjective visual interpretation of the data to extract meaningful insights [16]. The heatmap, or saliency map, is layered on top of the equirectangular projection of the 360-degree video. Also known as attentional landscapes [246], these heatmap visualise the area of which viewers direct their visual attention. Early heatmaps were generated by the summation of fixation data points, which were Gaussian distributed, resulting in a landscape where peaks represented the amount of fixations. Modern heatmaps still utilise fixation data points to generate a heatmap, but are often visualised as a three-dimensional height field or superimposed as a modified layer on top of the original image. This modified layer differs in colour or transparency is used to compose a set of multiple layers of bitmaps with partially transparent top layers [141]. While there exist many variations of heatmaps imagery (e.g. deviation map, difference map and significance map), the most common heatmap is known as the attention map, which represents the distribution of viewer's attention on an image. The generation of attentional heatmaps is achieved by constructing the heatmap based on fixation points in the eye-tracking data [32, 129]. As before-mentioned, a post-test subjective visual analysis is required to better understand user's attention, rendering human interpretation of the heatmaps a less precise method for extracting the nuances in gaze behaviour. While the eye-tracking analytics serve as an effective analytical platform for 360-degree video interactions pertaining simple narrative structures, they remain limited in the ability to process and assess complex narrative structures. There has been some progress in the development for analytical platforms for 360-degree content which are better suited for complex narrative structures, e.g. IVRUX [16], however their effectivity remains irresolute.

The immersive nature of 360-degree video poses a particular challenge when it comes to visualisation techniques for gaze data; the assumption that all users observe the exact same stimulus is not applicable to omnidirectional video interactions [194]. The user's ability to control FOV enables a different viewing experience every time. This limits the established gaze visualisation techniques, which are commonly used for omnidirectional scenes in a virtual environment that require static three-dimensional stimuli [240, 251, 310, 326]. A promising development has been

made by Löwe et al. (2015), which proposes new specialised visualisations and an analytical workflow for the analysis of head movement and gaze data of immersive 360-degree video [194]. In traditional video interactions, viewing behaviour and attentional synchrony are analysed by annotating ROIs, which is incredibly time-consuming and complex in 360-degree video analyses. Mostly due to the level of distortion caused by the projection techniques which "flatten" the spherical image (i.e., equirectangular projection), the selection of ROIs is complicated: an ROI that moves around the observer requires the selected area to transfer over from the right edge border of the ERP and continue on the left edge side. Therefore, in the study by Löwe et al. (2015), they base attentional synchrony on the similarity of the individual viewing directions. Attentional synchrony refers to the moments in which the viewing direction of multiple users is drawn to specific regions within the video [194, 301]. Two main aspects of the user behaviour analysis entail the FOV joins and FOV branches, which occur when the attention of multiple observers is synchronised to a common direction or when they are diverged, respectively. The convergence and divergence of FOV is, therefore, an important tool in the analysis of user behaviour, allowing for the identification of spatiotemporal agreements between the viewing directions of users. The implications made by this study indicate that these tools can be utilised to review and enhance narrative storytelling and immersion in immersive film-making by taking into account different video genres and their influence on gaze behaviour. However, genre and visual narrative techniques were not considered in the study. This is not only a limitation in the work by Löwe et al. (2015), but is commonly overlooked in many studies within the domain. The study by Fearghail et al. (2019) introduced an interesting study on intended viewing areas (IVAs) and estimated saliency, which aims to close this gap in current research by taking into account director's intent [99]. The saliency models used in the study generated estimates and predictions on areas that ought to grasp user attention. These saliency models can often be utilised to make estimates about user behaviour. However, their findings indicate that estimations generated by the saliency models did not correspond with director's intended viewing areas, further emphasising the importance of understanding the content-specific elements and their impact on user behaviour. This has also been demonstrated by Grindinger et al. (2011) demonstrating the distinction between estimated saliency and actual gaze behaviour regardless of instructions (tasked viewing vs. free viewing), suggesting consistent mis-identification of computational saliency models between human ROIs and artificial ROIs [116].

1.5 Quantitative Imagery Analysis in Computer Vision

Visualisation of gaze data can be achieved through a multitude of techniques, during which the distribution of data points is exploited [29, 88, 195, 225]. Statistical graphs, aggregated plots, scan paths and heatmaps are common techniques to achieve this effectively. The implementation of gaze data visualisation is crucial in the analysis of gaze behaviour, yet a quantifiable comparison remains a common limitation due to human vision [222]. As an evolutionary system, human vision enables to sense and process visual stimuli. The human interpretation and processing of visual stimuli enables sophisticated analyses, i.e., the identification of differences across visual stimuli.

As such, humans are adequately capable of utilising vision to identify distinctions and analyse imagery. However, achieving a comparable objective image interpretation to the HVS requires very complex computations. The interdisciplinary field of computer vision enables quantification methodologies for image processing, i.e., techniques to enhance gaze data analysis. Derived from scientific contributions within the field of computer science and remote sensing, computer vision facilitates the interpretation of indexed multi-dimensional spatial data [37, 181, 222]. Algorithms and techniques are deployed, aimed to emulate the human visual system, allowing for computer systems to perceive, understand and extract useful information from imagery. While a perfect replication of the human visual system remains challenging, computer vision techniques can improve upon the human vision system. Advancements in computational power and machine learning techniques enable complex image processing, through machine and deep learning methodologies, such as:

- Object recognition and detection [159, 231, 312, 330]
- Optical character recognition (OCR) [211, 241]
- Feature extraction [79, 222, 348]
- Motion estimation and tracking [12, 78]
- Stereo vision and depth perception [191, 322]
- Image restoration and enhancement [331, 369]
- Facial and pattern recognition (i.e., support-vector machines and random forest) analysis [95, 149, 161, 285, 329]
- Mathematical morphology [83, 120]
- Scene reconstruction [303]
- Video analysis and understanding [11, 113, 347]
- Image segmentation, similarity and feature matching [47, 62, 96, 199, 327].

These complex image processing techniques can be implemented in a multitude of research domains, advancing the development of unique practices and applications. As such, computer vision techniques have been employed in the development of intelligent application domains, i.e., vision algorithm development for intelligent environments [209, 350], enhancing complex processes in intelligent processes and enabling activity monitoring, content generation and educational tools. Computer vision as an area of research has made significant progress within the domain of healthcare, further solidifying its complex capabilities and use [165, 201, 208, 210]. Notably, the advancements made in analysing intraoperative video, in which the development of deep neural networks has allowed for the accurate identification of surgical phases and instruments, even surpass the accuracy of some surgeons [347].

In recent years, computer vision has emerged as a significant area of research with diverse applications across various domains, such as human-computer interaction (HCI) and game and media technology (GMT). Among the numerous techniques in computer vision, image segmentation and similarity assessments have gained significant attention in these domains. Image segmentation techniques divide an image into numerous sections, each of which represents an independent area of the image. The technique can be employed to isolate an image into foreground and background components, to identify and follow objects, or to extract characteristics for further analysis [51, 377]. In the field of HCI research, image segmentation has been employed to detect and classify facial expressions to identify emotions, as well as to segment and track eye movements for eye-based human-computer interaction purposes [61, 167, 198, 311, 317]. As implied by the works of Păsărică et al. (2017) and Tesfamikael et al. (2021), image processing techniques have shown to be effective in improving the accuracy and reliability of eye-tracking systems [235, 317]. Additionally, computer vision facilitates the advancements made in tracking hand gestures and movements for touch-less interfaces [234, 364].

Image similarity computation involves assessing the degree of similarity between two or more images. This technique is used in a variety of applications, such as image retrieval, object recognition, and content-based image retrieval. The technique is particularly useful when dealing with large image datasets, as it facilitates efficient image indexing and retrieval based on similarities in visual features. Additionally, image similarity techniques can be employed in computer vision systems to support tasks such as image classification and clustering.

The uniqueness of the 360-degree video format and range of application domains (i.e., entertainment and education) necessitates a higher degree of efficient techniques for imagery analysis and processing of 360-degree video. Computer vision has shown great potential in analysing 360-degree video by enabling object recognition, tracking, and activity recognition. While a powerful tool in image and video analysis, computer vision faces significant challenges in video prediction and generation. The task of generating accurate and realistic video sequences remains elusive, despite recent advances in generative models [359, 372].

The analytical advancements made in computer vision and machine learning have led to significant progress in imagery analysis. The deep learning algorithms for image indexing and retrieval, capable of accurately identifying and quantifying complex patterns in imagery and visual data, demonstrate the capability of computer vision techniques in conducting complex analyses of visual data such as heatmaps. While humans are capable of identifying the visual differences between heatmaps, the exact differences cannot be subjectively expressed in numbers and requires complex computational techniques to quantify these image differences. Therefore, these techniques can be applied to quantify the similarity of imagery, comparable to the work by Fakheri et al. (2012), which proposes a method for improving image indexing and retrieval algorithm by utilising shape and texture properties of the image [96]. Similar methods in computer vision and machine learning allow for complex imagery quantification, implemented in a variety of applications such as physiological monitoring, medical diagnostics, and sports analytics [25, 155, 321]. As a result, research in computer vision provides techniques that can be utilised to analyse and quantify the viewing behaviour of users in the field of VR and 360-degree video interactions. In addition, imagery analysis techniques have already been used developed to compute data that can be utilised in the analysis of 360-degree scene characteristics [105, 144, 268] and expression of viewing behaviour [38, 68, 98].

1.5.1 Computation of Spatiotemporal Complexity

The quantification of the spatiotemporal complexity of a video sequence is a common technique to compute quantitative data representative of the human vision system, which is sensitive to spatial and temporal changes in a video sequence [4]. The quantification method, as presented in the framework by ITU-T Rec. P.910, computes the spatial- and temporal image complexities of a video sequence, enabling spatiotemporal evaluation in video quality assessment [144]. Cui et al. (2021) studied the effect of gaming genre on the user experience, employing an analysis based on the resting-state electroencephalogram (rs-EEG) spatial- and temporal image complexity [72]. As such, the rs-EEG micro-state and omega complexity imply significant complexity changes due to genre characteristics. The spatiotemporal complexity assessment of neuro-imaging further demonstrates the effectiveness of employing computer vision techniques in imagery analysis [72, 106, 171]. As such, omnidirectional video content also involves sustained cognitive load on various behavioural systems, indicative of the genre-specific complexity changes as demonstrated in Cui et al. (2021). Moreover, the work of Yu et al. (2018) utilises spatiotemporal analysis of 360-degree video to identify highlights (e.g. important moments) from the omnidirectional content, emphasising the suitability of such analyses in the domain of omnidirectional content [367].

The computed SI- and TI-values provide a frame (i.e., spatiotemporal matrix) for obtained relevant data analysis, as demonstrated by Konuk et al. (2013) [172]. The location of the video sequence within the matrix can be utilised to identify the video sequence according to its position on the spatial and temporal planes, as well as ensure sufficient coverage of the spatiotemporal matrix. Specifically, the work by Singla et al. (2017) employs the computation of spatiotemporal complexities to assess the selected sequence of omnidirectional content [290]. Spatial complexity of a video refers to the visual richness and intricacy of its content, particularly in terms of the number and complexity of objects, colours, textures, and patterns within each frame and across consecutive frames. It is a multifaceted concept that can be influenced by various factors, including image resolution, contrast, dynamic range, noise, scene composition, lighting conditions, visual effects, and artistic style [55, 59, 72, 134, 288, 344]. Spatial complexity can be quantified using a range of objective measures, such as entropy, fractal dimension, spatial frequency spectrum, and compression ratio, as well as subjective assessments by human observers [124, 178, 184]. The temporal complexity of a video denotes the amount of visual change occurring in a video over time, reflecting the degree to which the video varies from frame to frame [17, 69, 117]. High temporal complexities represent significant variations in the visual information over time. Similarly, low temporal complexity reflects less significant variation in the content. The temporal complexity of a video sequence can be influenced by changes in camera motion, scene changes, frame rate, video codec and compression.

The level of spatiotemporal complexity can have important implications for video processing and analysis, as well as for viewer engagement, attention, and perception [205, 230]. The spatial- and temporal image complexity of a video sequence can be measured by computing the spatial and temporal perceptual information, respectively [144]. These measures provide single-valued representations of the complexity of each frame in a video sequence. The variability of these measures over time can also be studied to better understand the spatiotemporal characteristics of a video scene or sequence.

1.5.2 Spatial Perceptual Information Measurement (SI)

The spatial complexity or spatial perceptual information measure (SI) is determined by applying a Sobel-filter to each video frame (luminance plane) at time n, Fn. The Sobel-filtered frames are then used to compute the standard deviation over the pixels in each frame. This process is repeated for all frames in the video sequence, resulting in a time series of spatial information. The SI-value is determined as the maximum value of the standard deviations of the Sobel-filtered frames at time n, represented by the equation:

$$SI = \max_{time} \{\text{std}_{\text{space}} [\text{Sobel}(F_n)]\} \tag{1}$$

The implementation of the Sobel filter involves convolving two $3x3$ kernels over the video frame and obtaining the square root of the sum of the squares of these convolution results. Let the input image pixel at the i-th row and j-th column be denoted as $x(i, j)$, and let $y = \text{Sobel}(x)$. The $Gv(i, j)$ and $Gh(i, j)$ kernels represent the results of the first and second convolutions, respectively.

$$\begin{aligned} Gv(i,j) = &-1 \times x(i-1,j-1) - 2 \times x(i-1,j) - 1 \times x(i-1,j+1) + \\ &+0 \times x(i,j-1) + 0 \times x(i,j) + 0 \times x(i,j+1) + \\ &+1 \times x(i+1,j-1) + 2 \times x(i+1,j) + 1 \times x(i+1,j+1) \end{aligned} \tag{2}$$

$$\begin{aligned} Gh(i,j) = &-1 \times x(i-1,j-1) + 0 \times x(i-1,j) + 1 \times x(i-1,j+1) + \\ &-2 \times x(i,j-1) + 0 \times x(i,j) + 2 \times x(i,j+1) + \\ &-1 \times x(i+1,j-1) + 0 \times x(i+1,j) + 1 \times x(i+1,j+1) \end{aligned} \tag{3}$$

With calculations performed for all $2 \leq i \leq N-1$ and $2 \leq j \leq M-1$, where N denotes the total number of rows and M denotes the total number of columns in the video frame, the Sobel filtered image output at the i-th row and j-th column is:

$$y(i,j) = \sqrt{[Gv(i,j)]^2 + [Gh(i,j)]^2} \tag{4}$$

1.5.3 Temporal Perceptual Information Measurement (TI)

The measurement of temporal complexity, or temporal perceptual information measure (TI), is derived from the motion difference feature, $Mn(i, j)$. This feature calculates the difference between pixel-values (from the luminance plane) at the same spatial location in consecutive frames of a video sequence. $Mn(i, j)$ is a function of time (n), and is defined by the following formula:

$$M_n(i,j) = F_n(i,j) - F_{n-1}(i,j) \tag{5}$$

$Fn(i, j)$ represents the pixel located at the i-th row and j-th column of the n-th frame in time. The temporal information (TI) metric is obtained by calculating the standard deviation over space (std_{space}) of the motion difference feature, $M_n(i, j)$, for all i and j. This computation is performed

for each frame over time, resulting in a time series of temporal information. The maximum value of this time series (max_{time}) is considered to be the TI of the video sequence. As such, a higher TI-value represents higher levels of motion in adjacent frames of the video sequence.

$$TI = \max_{time} \{\text{std}_{\text{space}} [M_n(i, j)]\} \tag{6}$$

1.5.4 Structural Similarity Index Measure

Similar to the work of Cui et al. (2021), computations in imagery analysis enable the quantification of complex dataset visualisations and graphical representations, such as rs-EEG and heatmaps [72]. Image similarity techniques based on computer vision can be employed to compare graphical representations of a dataset and determine the degree of similarity between them [18, 214, 249]. A computational model for deriving the structural similarity between two images can be acquired by utilising the proposed structural similarity index measure SSIM, as introduced by Wang et al. (2004) [221, 345].

Wang et al. (2004) proposed the use of the structural similarity index (SSIM) as a full-reference image quality assessment (FR-IQA) measure [345]. This measure is based on the idea that the human visual system (HVS) is adept at extracting structural information from visual scenes. By incorporating this characteristic as an intrinsic component of the IQA measure, the authors were able to outperform not only the measures based on mean squared error (MSE), but also the existing state-of-the-art perceptual image quality measures. Moreover, the SSIM measure a more significant correlation with subjective evaluations provided by human observers, such as the mean opinion score (MOS) [14].

The increased performance, mathematical formulation, differentiability and high degree of computational parallelisation resulted in SSIM becoming a highly adopted FR-IQA measure within the scientific community, being utilised as a proxy evaluation for human assessment in image processing and computer vision applications. The high correlation with human perception of images enables SSIM to be implemented as a method for image denoising [216, 373], dehazing [27, 283, 342], artefact-free cloud removal [324], image enhancement [374], raindrop removal [248] and medical imaging segmentation [26, 266].

The Structural Similarity Index (SSIM) is a commonly used image similarity index measure that quantifies the degree of similarity between two images by evaluating their structural information. SSIM computes the similarities between local image regions by comparing their luminance, contrast, and structure based on the correlation between pixels [345]. The luminance comparison measures the brightness similarity of the pixels, while the contrast comparison calculates the difference in pixel intensity. By modelling the image distortion as a combination of the factors of correlation loss, luminance distortion and contrast distortion, the SSIM index relies less on conventional error summation techniques and correlates more with the human visual system (HVS) [131].

SSIM computes a comparison between a reference image x and a version of the same image y based on the three components of luminance, contrast and structure, extracted at a single spatial scale (i.e., resolution) [14, 41, 345]. The luminance comparison, as a measure of luminance closeness between x and y is estimated as the mean intensity:

$$\mu_x = \frac{1}{N} \sum_{i=1}^{N} x_i \tag{7}$$

The luminance comparison function $l(x, y)$ can be expressed by the mean values μ_x and μ_y of the two images x and y. The standard deviation is utilised as an unbiased estimate of the signal contrast, measuring the closeness of the contrast, expressed as:

$$\sigma_x = \left(\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)^2 \right)^{1/2} \tag{8}$$

As such, the contrast comparison function $c(x, y)$ is the comparison between σ_x and σ_y. The structure comparison function $s(x, y)$, measuring the correlation coefficient between images x and y, is conducted on the normalised signals by dividing by σ_x and σ_y, respectively. The function $s(x, y)$ is then expressed as $(x - \mu_x)/\sigma_x$ and $(y - \mu_y)/\sigma_y$. The combination of signal luminance, contrast and structure results in the following expression for the overall similarity measure:

$$S(x, y) = f(l(x, y), c(x, y), s(x, y)) \tag{9}$$

The functions of $l(x, y), c(x, y), s(x, y)$ and $f(\cdot)$ are defined based on the conditions of symmetry $S(x, y) = S(y, x)$, boundedness $S(x, y) \leq 1$ and unique maximum $S(x, y) = 1$ if and only if $x = y$ (in $x_i = y_i$ for all $i = 1, 2, ..., N$). Subsequently, the luminance comparison is defined as

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{10}$$

where C_1 ensures numerical stability when $\mu_x^2 + \mu_y^2$ nears zero, as C_2 and C_3 do in the following equations for contrast and structure, respectively. The constants C_1, C_2 and C_3 avoid the null denominator. According to Weber's Law, the human visual system (HVS) is sensitive to relative luminance change, not absolute [298]. Therefore, the luminance signal is denoted as $\mu_y = (1+R)\mu_x$.

To quantify the luminance, contrast, and structure of images, the dynamic range of two scalar constants $K_1 \ll 1$ and $K_2 \ll 1$, as well as the of pixel-values L (which is set to 255 for 8 bits / pixel greyscale images), are used. These quantities are utilised to determine the positive constants C_1, C_2, and C_3, which are given by $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$, and $C_3 = C_2/2$. Standard deviation σ is used to represent signal contrast, defined as:

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{11}$$

The process of structure comparison requires luminance subtraction and variance normalisation. As such, the mean values of each image are subtracted from their respective pixel-values to obtain a new set of values representing the differences in luminance. These new values are then normalised by dividing them by the standard deviation of the original pixel-values. The resulting normalised values are unit vectors that lie in the hyperplane, defined by:

$$\sum_{i=1}^{N} x_i = 0 \tag{12}$$

The correlation between the unit vectors from the normalised values $(x-\mu_x)/\sigma_x$ and $(y-\mu_y)/\sigma_y$ is equivalent to the correlation coefficient between x and y, defining the structure comparison as follows:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{13}$$

where σ_{xy}, the covariance between x and y in discrete form, can be estimated as:

$$\sigma_{xy} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y). \tag{14}$$

The resulting function of SSIM(x, y) can be expressed as a combination of the luminance, contrast and structure comparison functions, defined as:

$$\mathrm{SSIM}(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma}. \tag{15}$$

Substitution of respective functions $l(x, y), c(x, y), s(x, y)$ gives:

$$= \left[\frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \right]^{\alpha} \cdot \left[\frac{2\sigma_{xy} + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \right]^{\beta} \cdot \left[\frac{2\sigma_{zy} + C_3}{\sigma_x^2 + \sigma_y^2 + C_3} \right]^{\gamma}. \tag{16}$$

The expression of the SSIM index is weighted with α, β and γ, where $\alpha > 0, \beta > 0$ and $\gamma > 0$. The parameters can be used to adjust for the relative importance of each of the components. Finally, by setting $\alpha = \beta = \gamma = 1$ and $C_3/2$, the SSIM index expression is simplified in definitive form as:

$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{17}$$

The SSIM index takes on positive values in the range [0, 1]. A value of 0 indicates no correlation between the images, while a value of 1 signifies that the two images being compared are identical ($x = y$). For a comprehensive overview of the mathematical construct of the structural similarity index measure, please refer to the paper by Wang et al. (2004) [345].

1.5.5 Advanced Structural Similarity Index Computations

Numerous adaptations of the conventional SSIM model have been proposed in the literature, to enhance the model's robustness and applicability to diverse visual applications. These adaptations are often aimed at addressing limitations associated with the single-scale nature of the SSIM model and include various multi-scale extensions, such as Multi-Scale Structural Similarity (MS-SSIM) [14, 221, 346], Multi-Component SSIM (MC-SSIM) [187], Colour-Comparison SSIM (CM-SSIM) [122], Structural Dissimilarity (DSSIM) [110, 276], Complex Wavelet SSIM (CW-SSIM) [107, 273], and continuous SSIM (cSSIM) [200]. Additionally, other variations of SSIM have been proposed to improve its performance for specific applications, demonstrating the ongoing efforts to refine the SSIM model for diverse visual applications and highlighting the importance of customised adaptations for specific domains.

1.5.6 Multi-Scale Similarity Index Measure

In the domain of eye-tracking research, colour heatmaps are commonly implemented as effective and reliable graphical representations of gaze data. The data values are encoded as colours on a two-dimensional plane, in which the colour intensity is indicative of the density of data points. As such, higher intensity colours represent areas of greater interest, i.e., fixation points. The saliency detection algorithm, used to generate saliency values to each pixel, are used to visualise a multi-scale representation of the image. Achieved through a Gaussian pyramid decomposition, the saliency detection algorithm computes saliency values at each scale of the pyramid, enabling up-sampling and aggregation of individual heatmaps [1, 180, 334].

An enhanced version of the SSIM index has been developed which operates across multiple scales of both image signals [122, 250, 265, 304, 336]. The Multi-Scale Similarity Index Measure MS-SSIM calculates the differences in contrast and structure at each scale, while the comparison of luminance is only computed at the highest scale M. The highest scale M is obtained after $M-1$ iterations, of which scale 1 denotes the index of the original signal. During the $M-1$ iterations, MS-SSIM applies a low-pass filter to the input images signals, followed by a down-sampling of the processed image by a factor of 2. This process results in a revised expression and substitution

of the comparison functions of l, c and s in (15). Contrast comparison $c_j(x, y)$ and structure comparison $s_j(x, y)$ are calculated at the j-th scale. Subsequently, luminance comparison $l_M(x, y)$ is calculated only at scale M. The MS-SSIM index is expressed as:

$$\text{MS-SSIM}(x, y) = [l_M(x, y)]^{\alpha_M} \prod_{j=1}^{M} [c_j(x, y)]^{\beta_j} [s_j(x, y)]^{\gamma_j} \tag{18}$$

where the parameters are selected such that $\alpha_M = \beta_j = \gamma_j$ for all j and $\sum_{j=1}^{M} \gamma_j = 1$. The exponents α_M, β_j, and γ_j enable the assignment of different weights to the segmented regions-of-interest within the image signals.

The different colour intensity levels in heatmap visualisations cannot be disregarded in the similarity assessment of two heatmap image signals, as the colour intensity represents the density of gaze data points. Therefore, the use of the multi-scale structure of MS-SSIM generates results with a higher degree of accuracy and reliability for image and video databases as opposed to single-scale SSIM [82, 305, 346]. MS-SSIM is capable of incorporating colour information and similarity on multiple scales, making it a more reliable and accurate computation for the nuances of colour in heatmap visualisations [336], as well as one of the most precise FR-IQA measures [14]. While both SSIM and MS-SSIM operate on greyscale image signals, colour image signals require additional image processing for compatibility with SSIM and MS-SSIM [221, 336, 346]. The conversion of colour image signals to the YCbCr colour space for MS-SSIM computation generates more precise results, as opposed to the conversion from the RGB colour space to greyscale for single-scale SSIM [221, 345].

Gaze tracking poses an important tool in the overall development of omnidirectional content analyses and development of HMDs, enabling high quality user behaviour analyses and inciting technical development in areas such as viewport prediction [359]. As established in Zink et al. (2019), the viewing experience of 360-degree video is dependent on the extent of viewing guidance. Next-generation HMDs are improving eye-tracking technology and gaze detection, resulting in more precise predictions and provision of spatial information to guide user focus and viewing behaviour [375]. However, despite the efforts of optimising oculesics, aspects from related fields such as cognitive science and film studies still impose significant influence to the overall viewing experience and guide the viewer while viewing 360-degree video content. Therefore, it is important to understand the relationship between the perception of conventional video and omnidirectional content, as well as the cognitive and cinematographic influences in guiding user attention.

1.6 Attention Guidance and User Engagement in VR

Traditional videos can be specifically edited to tailor its target audience. Through complex post-production processes, camera angles, and cinematography, the viewer's attention is guided. The traditional videos often evoke high levels of attentional synchrony, where most viewers look at the same targets at the same time [193]. During traditional video development, the director aims to guide the user's attention in a carefully designed way, either to omit specific plot points or to drive the narrative. This way, the viewer is taking on a passive role as a static element in the interaction process. However, with the introduction of VR and 360-degree video, the viewer's attention is no longer strictly guided by camera angles and framing of the scene. The unrestricted orientation allow for exploratory behaviour and freedom [262, 325]. Instead, the viewer acts more like an actor, able to adjust perspective and viewing angle as they see fit, controlling the camera and adding an extra dimension to the interaction [281]. As identified by Rothe et al. (2019), traditional attention guidance mechanisms may not be as applicable to omnidirectional video [262]. The enhanced perceptual load (i.e., increased amount of visual searches) and potentially missed content bears weight on the importance of better understanding how omnidirectional video content is perceived as opposed to traditional video content. Due to the complexity of cinematography as an art-form, this section will focus specifically on cinematography as a tool to direct and guide user attention, as well as explores the attention guidance mechanisms in 360-degree video experiences.

1.6.1 Perception of Conventional and Omnidirectional Content

Since the invention of video, which introduced the projected series of still images, the format has been revolutionary. The format remains widely adopted today, ranging from cinema to video games. Traditional cinematography utilises the video format and adapts it to drive a visual narrative forward, acting as an efficient tool to guide viewer attention. As an art-form, it uses a set of visual elements such as lighting, framing, composition, camera motion, camera angles, frames, lens choices, depth of field, zoom, focus, colour, exposure and filtration [202] to achieve this. Carefully navigating these visual elements allows the cinematographer to ensure the director's vision is translated onto the video-format. Specifically, in the interaction with 360-degree video interaction, the director's intent and resulting visual attention of the viewer do not usually align [99]. Hence, the cinematography of a video plays a pivotal role in guiding the attention of the viewer. Conventional cinematography makes use of continuity editing, a method of editing the different camera shots into a coherent sequence of events [34]. The result is a sense of situational continuity in which the flow of non-linear information, which often due to time and location dependency, is perceived as a single event [232]. Many techniques, such as the 180-degree rule and action cuts, allow for increased efficiency of spatial perception and the overall success of continuity editing [300]. While this method has become a complex and well-adapted system for video editing, describing the vast set of rules, mechanisms behind continuity editing in traditional cinematography is beyond the scope of this research. However, it is important to highlight the core components of spatial perception that allows for continuity editing to be very effective. A key component of this effect is the ability to break up a continuous flow of information into a series of meaningful events. In the field of cognition and neuroscience, this ability to segment is better known as event segmentation theory [28]: the cognitive ability to predict the immediate course of events by creating an memory-based interconnected representation based on the segmented series of meaningful events [175, 272]. Recent development in film studies suggest that this predictive process is applicable to the film industry, emphasising the importance of prediction in spatial perception [197]. Professional film editors tend to utilise editorial cuts to disrupt event continuity expectations among the viewers [34], which is a direct result of the predictive process suggested by event segmentation theory. As defined by Zacks et al. (2010), event segmentation theory is an aspect of our perceptual processing [175, 197, 257, 272, 368]. This process is responsible for segmenting the flow of information into a hierarchical set of discrete events. These segments are used to create a mental representation that is used to predict the course of events, and when new events are registered due to changes in time, space or action, this mental representation is updated with new information. This allows the viewer to conceptualise events in relation to each other and is an important aspect in understanding why continuity editing is an effective tool for video editing and achieving a guided viewing experience for conventional video-content.

However, as previously defined, virtual reality content differs from conventional video, with the most important distinction being the user's ability to interact with their point of view (POV). This takes away an important aspect of continuity editing: camera angles, thus eliciting questions on the effectivity of continuity editing for guiding viewer attention VR. The before-mentioned study by Serrano et al. (2017) suggests that various types of edits in VR are equally well understood in terms of continuity as compared to traditional video content [281]. While other studies related to editing tools, and their effect on viewing experiences, are predominantly designed for two-dimensional viewing experiences [56, 150, 355], the study by Serrano et al. (2017) provides a stark contrast by focusing predominantly on three-dimensional viewing experiences in VR. Notably, the foundation of event segmentation theory still applies to VR editing and VR content. The study used a HMD with eye-tracking to study gaze patterns and viewer behaviour, and compared the data among video edits of three different classes:

- E1: edits that are discontinuous in space or time, and discontinuous in action (action discontinuities);
- E2: edits that are discontinuous in space or time, but continuous in action (spatial / temporal discontinuities);

- E3: edits that are continuous in space, time, and action (continuity edits).

The study suggests that Regions of Interest (ROIs) attract more attention after an E2 edit than after type E1. Moreover, a peak in exploration was found at the beginning of each clip and after an edit, suggesting that users need time to adapt to new visual content. Similar to traditional cinematography, this study demonstrated that action discontinuities are the strongest predictors of event boundaries in VR and that continuity edits maintain the perceived continuity despite visual discontinuity. For instance, aligning ROIs across edits is recommended strongly for fast-paced action movies, while ROI misalignment evokes more exploratory behaviour. The study also found an exponential relation between misalignment across edit boundaries and the time it takes for viewers to fixate, with large misalignments affecting viewer behaviour, even after they fixated on the new Region of Interest (ROI). The findings of their study provide insights into the potential responses elicited from certain edit configurations, such as the importance of aligning ROIs across edits for fast-paced action movies and the effect of misalignments on eliciting exploratory behaviour. This implies that the effects of event segmentation theory in traditional video content also apply to virtual reality content.

As a major distinguishing characteristic of 360-degree video, controllable POV and perspectives are inherent to the user experience. An extra degree of freedom is offered when viewing 360-degree video content, by enabling the viewer to control POV and / or perspective. Therefore, it is important to understand the effects of a variable perspective and POV in terms of viewing experience. The extra degree of freedom elicits the problem in which users can potentially miss out on content. Specific areas of the spherical projection that are not within the FOV of the user risk not being seen, which forms a major problem when key events are missed by the user [262]. With the implications made by Serrano et al. (2011), it is assumable that, in terms of viewing experience, event segmentation theory still holds place during both virtual reality content as with traditional video content. Research by Swallow et al. (2018) studied the extent to which perspective and POV have an effect on the viewer's ability to segment the content into separate events [313]. The study focused primarily on the relationship between visual features and segmentation in first-person and third-person videos. The results show that while the visual features changed across the different perspectives, they had small and inconsistent effects on the way participants divided the activity into parts. These findings are contradictory to the visual feature dependent hypothesis, which states that if segmentation is tied to the low-level visual features of a video, such as actor posture and visual change, they should be strongly related to segmentation for both first- and third-person videos [313]. By ruling out the visual feature dependent hypothesis, these results suggest that segmentation is relatively robust to changes in visual input. Instead, these findings indicate that segmentation mechanisms appear to flexibly use visual information to identify the underlying structure of the activity in a manner that is mostly viewpoint invariant. Despite these results indicating a cause through higher-level cognitive mechanisms that require further research, the implications made by studying event segmentation theory in the context of viewing experience for virtual reality content remains extremely valuable. By identifying the similarities of video editing techniques on the viewing experience in both conventional video and omnidirectional video, the implication can be made that video-editing techniques used in conventional video content have a very similar effect on viewing experience and overall perception of 360-degree content in virtual reality as with traditional viewing experiences. The before-mentioned studies also explore, on a cognitive level, the mechanisms that enable a similar viewing experience in terms of sense of continuity in both traditional video-content and virtual reality content. The findings from these studies indicate that event segmentation theory is responsible for the viewer's sense of situational continuity and that this cognitive mechanism is as applicable to virtual reality content as it is to conventional video content. The study by Serrano et al. (2017) further explores how continuity editing in virtual reality content can influence gaze patterns and offers guidelines as to how ROIs can be utilised to provoke specific user behaviour [281]. The implication that event segmentation and viewing experience are coherent in both traditional video content and virtual reality content indicate that other effects on viewing experience, using conventional video-content, are to be expected with virtual reality content as well. The small and inconsistent effect of perspective

(first-person vs. third-person) on the participants ability to segment the content into parts made by Swallow et al. (2018) indicates that perspective does not play a significant role in the viewers sense of continuity and ability to properly perceive the presented content, which, according to Serrano et al. (2017), is applicable to virtual reality content as well.

Virtual reality content, specifically 360-degree video, is conventionally viewed in a variety of ways. One of the key challenges in 360-degree video delivery are the unpredictable delivery circumstances and usability contexts. Users have the ability to view the omnidirectional content in different ways and use other viewports that vary in their level of quality [375]. However, as distinguished by Zink et al. (2019), there exist perceptual commonalities among viewers that are coherent in spatial, temporal, locational and behavioural ways. Spatial, temporal and locational coherence all refer to the source file transportation from edge servers to the clients. Proposed solutions for these distribution challenges include: prefetching and caching of spatially adjacent tiles and the prediction of a client's viewport trajectory to reduce the motion-to-photon delay. While there is a need for more research on how these coherences can be utilised in an efficient and scalable manner, it is beyond the scope of this thesis. Continuing, behavioural coherence describes the principle that the viewport of users (e.g. HMD, mobile viewport, display) is correlated to behavioural responses, regardless of the arbitrary ways in which the user is able to view the content [23, 63, 375]. This suggests that when using cinematographic rules to guide viewer attention, it is possible to exploit this principle. As 360-degree videos often have points of interest (POIs) or areas of interest (AOIs) which tend to grasp viewer attention, it is possible make predictions on the user's gaze behaviour.

1.6.2 Attention Guidance Mechanisms in 360-Degree Video Experiences

The lack of a predefined perspective, or view, enables to viewer to experience a close-to-life environment. The before-mentioned cinematographic rules limit 360-degree video creation, constraining the director / content creator in their ability to guide viewer attention. A significant amount of research has been done in the field of attention guidance mechanisms and the effectiveness thereof [263, 284, 306, 341]. Strongly aimed at guiding viewer attention, current literature also suggests a strong influence and correlation with overall user experience [220, 262]. With foundation in film theory literature [114, 299], diegesis, as a construct, is often implemented to improve attention guidance. A diegetic mechanism implements visual artefacts of the narrative or environment to guide attention [262, 306]. Similar to cuts in traditional cinematography, the use of these internal story- / scene elements support attentional continuity by utilising character motion and non-verbal behaviour to guide the viewer, evoking a natural orientation towards the target of attention [100, 299]. However, elements that are external to the scene or story (i.e., graphical symbols and pointing arrows) are prone to be less effective in guiding attention. These non-diegetic mechanisms usually only guide the user's attention to a specific POI or AOI, rather than evoking naturally guided behaviour [262]. Current research explored the inclusion of diegetic mechanisms and compared its effect against non-diegetic mechanisms in terms of user experience, sense of presence and user preference [48, 220, 306]. It was found that the inclusion of diegetic mechanisms evoked higher levels of all three aspects. However, in terms of task performance, both mechanisms proved effective. A study by Norouzi et al. (2021) evaluated the effectiveness of using virtual animals as diegetic attention guidance mechanisms, which acknowledge the presence of the user within the 360-degree experiences, and compared it to non-diegetic mechanisms [223]. Their findings indicate that both mechanisms were effective in guiding users towards target events. However, they found that diegetic artefacts induced a higher sense of presence, and yielded better user experiences overall. The user-acknowledging behaviour of diegetic artefacts (i.e., virtual animals) and conspicuous appearance of non-diegetic artefacts both positively enhances user engagement and influences behaviour [224]. This result can be explained through the increased levels of presence, thus reducing risk of eliciting the Swayze Effect: the sensation of feeling no tangible relationship with the (virtual) environment, despite being present [338]. However, despite the promising results, their study did not take into account various content types (e.g. educational vs. entertainment), suggesting

lack of research devoted to the influence of content type on the guidance of user attention in virtual 360-degree environments.

Guiding viewer attention in 360-degree video requires less obtrusive techniques to ensure the viewers ability to control their gaze. In a study by Sheikh et al. (2016), a variety of unobtrusive techniques for directing attention in VR were evaluated [284]. Their findings suggests that integrating both audio and visual cues from a target area (POI or AOI) is the most effective technique for guiding attention in 360-degree video. Separation of audio and visual cues can lead to a decreased effectiveness, whereas integration of both leads to all participants effectively noticing the target area. Notably, out of all four techniques implemented, the most unobtrusive technique entailed a diegetic artefact: a bystander walking across the action towards the target area. In addition, their study studied the effect of distance at which action occurs on the level of immersion and viewer enjoyment. Participants avoided invasion of private space, aiming at maintaining a "safe" distance from the target area as to not be too distant or close. This observation is in line with the findings from Wilcox et al. (2006), which state that people tend to respond similarly to invasion of personal space in both virtual as well as real-life settings [284, 352]. Furthermore, this can be directly linked to Hall's model of proxemics [119] and the findings of Keskinen et al. (2019), which studied the effect of camera height on user experience [164]. They found that a close proximity and low camera placement negatively affected user experience, similar to Saarinen et al. (2017) [269]. Lastly, in the study by Skeikh et al. (2016), it was found that participants felt more immersed in the 360-degree video content when diegetic elements of the content interacted with the participant (i.e., characters making eye-contact). Although diegetic mechanisms are proven effective in subtly directing user attention, these diegetic artefacts also have a greater likelihood of being overlooked [48, 220, 262], caused by higher levels of plausibility illusion (Psi) [294].

1.6.3 Place and Plausibility Illusion

Viewing omnidirectional content through a HMD is perceived as an immersive experience. The level of immersion that the user perceives relies on a variety of parameters. As identified by Slater et al. (1997), the parameters that are strongly associated with immersion are – but not limited to – the following: graphics frame rate, extent of physiological tracking, tracking latency, image quality, field of view, render quality, dynamics, and sensory modalities [297]. The use of such immersive environments rely heavily on sensorimotor contingencies (SCs): the notion that users know how to act, in order to perceive. When interacting with a HMD, this relates to the notion that users are aware that changing head-direction or moving around changes orientation within the virtual environment [227], mimicking a physical environment. The increased levels of presence that result from SCs evoke place illusion (PI), defined as "the strong illusion of being in a place in spite of the sure knowledge that you are not there" [294]. While sharing similar qualities, PI varies from immersion and is distinguished by Slater et al. (2009) as such that immersion provides the boundaries within which PI can occur. Immersion can be seen as a property of physics, whereas PI specifically denotes the sense of "being there". The extent to which users explore the system and its physical boundaries can evoke disruptions in the occurrence of place illusion [108]. Contrary to place illusion, plausibility illusion (Psi) denotes the illusion of "perceiving that virtual events are really happening, even though the you know it is not real". The disruption of place illusion can be restored through technical adjustments or resumed activities (i.e., corrections in head-tracking or correctly rendering scenes). In particular, PI refers to how the virtual environment is perceived, whereas Psi refers to what internal representations are perceived. Psi doesn't require physical realism, as was demonstrated by Milgram's paradigm [30, 295]. Equally applicable to virtual environments, as demonstrated by Slater et al. (2006), as participants showed physiological responses when exposed to external events directed at them but not caused by them. Regardless of physical or virtual realism, events and actions directed at the user (acknowledgement) evoke physiological responses, similar to how they would respond in real life. The extent of which depends on the level of realism and virtual acknowledgement of the user. This correlation principle, between external events and a user's own interoceptive / exteroceptive sensations, is fundamental for the occurrence of Psi. Multiple studies have demonstrated this cause-effect relationship [103, 104, 294,

295, 296, 332] and occurs through the action-reaction interaction between virtual events and the perceiving user.

1.6.4 User Engagement

The increasing supply and demand of (omnidirectional) video content presses firmly against the sense of responsibility among content providers to deliver high-engagement and high-quality video content to the users. This also entails optimising content, further underlining the importance of effective attention guidance, dictated by attention economics [287]. The similarities in the cognitive perception of traditional video and omnidirectional video strongly imply external validity from studies on the impact of traditional video qualities on user engagement. Key quality metrics often relate to technical attributes, such as bit-rate, buffering ratio, join time, and rendering quality. However, the development of metrics on autonomous and independent influence of 360-degree content on the user interaction remains unexplored. Since a higher quality of experience relates to higher user satisfaction with content quality, common objective metrics for measuring engagement when viewing video is through play time. Such metrics provide the content creator with precise information on what type of content receives more overall engagement [81].

The emphasis on understanding beyond usability, within the field of human-computer interaction, is to create more engaging experiences [123, 147, 182]. As demonstrated by O'Brien et al. (2008), the theoretical foundations that underpin and expand upon the traditional attributes of engagement (user activities, user attitude, mental models, motor skills, intrinsic interest, attention and motivation) [52, 157, 270] are based on flow theory, aesthetic theory, play theory, and information interaction [205, 230]. The work of O'Brien et al. (2008) approaches engagement as a quality of user experience and as such, takes into account the influential aspects (threads) of the user engagement experience. Known as the experience threads, the sensual-, emotional- and spatiotemporal threads pertain to a variety of attributes of the user experience. The sensual thread entails the visual, auditory and interactive components. The emotional thread comprises the affective experiences related to the user interaction. Lastly, the spatiotemporal thread involves the dimensions of time and space in which the interaction takes place. These defined threads of experience are essential in understanding the attributes that initiate the point of (re)engagement, maintain engagement and lead to disengagement. Combined with the analysis on different application areas of engagement, their synthesis of the theoretical framework has resulted in a model of engagement that contains the following attributes: aesthetic and sensory appeal, attention, awareness, control, interest, novelty, challenge, feedback, positive / negative affect, motivation, usability, perceived time, interruptions, and interactivity. Moreover, the framework encompasses the behaviours, cognitions and emotions of the user in the context of design, interactive features and content application.

Many studies suggest that the engagement enhancing quality of 360-degree video increases the sense of presence, involvement, empathy and enjoyment [274, 286]. However, 360-degree video can also evoke a sense of apathy, distraction and cybersickness, inhibiting the user experience [39, 188, 242, 267, 302]. As defined by Wang et al. (2018), current challenges of user engagement in 360-degree video entail cybersickness [2, 121, 207], physical discomfort [74, 112], cognitive barriers [39, 267], attention [160, 188, 189, 238, 242], satiation [57] and visual quality [190, 236]. As previously stated, the independent role of content is often an overlooked aspect in 360-degree video research, despite its significant influence on user engagement and experience [31, 177, 188]. In some cases, these content-related components have a higher impact on the user experience than the technological factors, highlighting the significant influence thereof [19, 20, 102]. The studies by Koehler et al. (2005) and Bleumers et al. (2012) demonstrate the complex interaction between the media format, narrative, and video style, indicating a significant distinction in the suitability of various genres for 360-degree video [31, 168]. The user's ability to self-determine their focus calls for content-specific adaptations. For instance, a slower pace might be more suitable for the viewing experience of documentaries. Furthermore, the work by Wang et al. (2018) studied the influence of content genre on audience engagement [343]. Results from their study demonstrated the same complex interaction effect, previously defined by Koehler et al. (2005), indicating significant

differences in level of engagement between content genres. For example, genres such as "sports" and "science" were found to be distinctively engaging than other genres. In general, these findings suggest that each type of content has different effects on the level of engagement [7, 343]. However, by only analysing objective measures, they did not take into account the subjective measures that hold significance in better understanding the influence of 360-degree content on user behaviour and engagement. Moreover, the unique behavioural actions related to VR system interaction (i.e., head movement and gaze) were not included in the studies.

1.7 Conclusion

The literature study explored the current state of research within the field of 360-degree video user interaction. The body of literature presented in this study emphasises the complex interaction process between user and 360-degree video in VR, and further presents the underlying theories and conceptual models that entail the interaction process. The theoretical foundations presented in this literature study make use of the 360-degree video interaction model as a multi-dimensional construct to help understand the complex interplay of perceptions, cognition, usability and user behaviour during 360-degree video interaction [105, 356, 371]. By including cognitive principles, attentional guidance mechanisms, cinematographic concepts and perceptual attributes of user engagement, this research closely examines the representative 360-degree video interaction domains – cognitive perception and behavioural consequences – and explores influential components of image complexity on user behaviour across the levels of direct perception, interaction and usage situation [42].

The findings by Singla et al. (2017) demonstrated the distinctive behavioural responses while interacting with various 360-degree content using a head-mounted display, emphasising the pivotal role of 360-degree content in evoking behavioural responses [290]. Despite the significant effect of content-specific attributes on gaze behaviour [31, 177, 188], current research remains predominantly focused on the technical limitations of the 360-degree video format [109, 132, 151, 282], disregarding the use of content-aware approaches and neglecting the significant effects of content-specific characteristics on the user interaction [31, 177, 188]. Attributed by complex spatial and temporal information, the content-specific 360-degree image complexity as an essential part of the system interaction remains overlooked, similar to the exclusion of user preferences and usability context [9, 90]. The various output modalities, in which 360-degree content can be consumed and interacted with (e.g. VR, mobile viewport, seated and standing), further underpin the importance of considering the influential aspect of usage situation and usability context. Essential to the understanding of behavioural consequences of 360-degree video content is how the usability context, presence of attention guiding mechanisms and cognitive engagement interact to shape gaze behaviour [223, 293, 328, 356].

The distinct image complexity across 360-degree video comprise visual changes on both the spatial and temporal planes, leading to a considerable range of spatiotemporal complexity manifested in diverse visual and cinematographic features, unique to each 360-degree video sequence. Spatial complexity remains an excluded yet significant aspect of 360-degree video interaction research, attributed to visual richness and cinematographic characteristics associated with distinct genres [31, 55, 59, 72, 177, 188, 194, 288]. Similarly, the relevance of temporal image complexity is grounded in the both technical and behavioural implications, such as bit-rate variability and cognitive load [2, 262]. Therefore, it is important to take the nuanced temporal factors, such as camera motion, into consideration as it denotes content-specific temporal information [17, 69, 117, 328].

Furthermore, this chapter elaborated on the implications and significance of content-aware approaches, suggesting a higher impact of content-specific characteristics on the user's gaze behaviour compared to technological factors [19, 20, 102]. Moreover, understanding the relationship between spatiotemporal image complexity of 360-degree video and gaze behaviour could hold implications that extend beyond the user interaction and which could ameliorate research on technical challenges of 360-degree video such as optimising viewport prediction and enhancing foveated rendering techniques [39, 101, 145, 360]. The varying extent to which the viewer perceives the

360-degree video suggests that each type of content impacts the viewer differently, eliciting distinct behavioural responses [7, 343]. Unique to the 360-degree video format is the lack of perception of the entire 360-degree frame, as large areas of the spherical projection reside outside the user's FOV. The fear of missed content (FOMC) on specific content increases due to the amount of visual searches [223, 262]. Consequently, the interaction with 360-degree content significantly stresses on the user's cognitive and perceptual load, bearing weight on the importance of better understanding how 360-degree video content is perceived, as well as the significance role of usability context on the interaction.

Understanding these effects is imperative to the development of 360-degree video content that aims to guide user attention [99, 118]. This is further implied by the novel and experimental nature of 360-degree content development, inciting the proverbial plateau of latent potential when it comes to optimising and tailoring content to specific user behaviour. Building on this understanding, this thesis concerns the extent of which the spatiotemporal complexity of a 360-degree video sequence in VR impacts the user's gaze behaviour amidst the multifaceted interaction process in which the complex dynamics of cognition, perception and usability cannot be discarded. Consequently, the main research objective of this thesis is defined as:

> To discern the extent to which spatiotemporal image complexity of a 360-degree video sequence in VR influences gaze behaviour within the multifaceted interaction model, while factoring in the complex dynamics of cognitive perceptions and usability context.

A series of sub-questions have been devised to facilitate a more comprehensive research and to provide additional insights in addressing the primary research objective.

1. How can computer vision techniques, paired with eye-tracking data, be employed to quantify gaze patterns?
2. To what degree do cinematographic principles and attributes of cognitive perception impose a confounding effect on the user's behavioural response?
3. How is gaze behaviour affected by spatial image complexity?
4. How is gaze behaviour affected by temporal image complexity?
5. To what degree is the effect of spatial- and temporal image complexity on gaze distribution mediated by usability context?
6. How does the user's self-perception of conscious gaze behaviour compare with gaze data?

This literature study also presented viable methodologies and relevant mathematical concepts (i.e., eye-tracking, gaze analysis, M-ACR, MS-SSIM) for physiological, objective and subjective analysis of user behaviour during the 360-degree video interaction. The latter of which poses a common limitation in current research, as significant subjective measures are commonly excluded, despite the equal importance in better understanding the influence of perceptual attributes on user behaviour [31, 223, 262]. Moreover, developments in the field of computer vision enable quantification of spatiotemporal complexities and image structures, as well as computational methods to assess the unique visual characteristics of 360-degree video content, more representative of the human visual system.

In conclusion, studying user behaviour and visual gaze in 360-degree video is imperative to the optimisation of 360-degree content. The scope of this thesis and findings hold implications that are essential for enhancing the overall user experience, improving video design, deepen the understanding of cognitive processing, and evoking higher user engagement in 360-degree video interactions in VR. Therefore, this chapter emphasises the importance of studying these components to fully leverage the potential of 360-degree video.

Chapter 2

Research Methodology

The research methodology discussed in this section presents the methodologies that were applied during this study. As described in section 1.7, the aim of this thesis is to discern the extent to which spatiotemporal image complexity of a 360-degree video sequence in VR influences gaze behaviour, closing the gap in current literature by considering 360-degree video content as an autonomous and independent factor in the interaction process. This section elaborates on the research study that was conducted as a means to approach this. The structure of this chapter is as follows: firstly, an overview of the overarching aspects and general research design is presented in section 2.1. The significance of studying both physiological, objective and subjective metrics [10, 91, 289] was translated into two parts of the study. As for the physiological and objective analysis, an eye-tracking study was performed and detailed in section 2.2. The second part of the study focused on the subjective user-centric evaluation and was performed through a post-test user evaluation, presented in section 2.3. The subsections thereafter encompass overarching elements of the study, such as population and procedure, and are discussed in sections 2.5 and 2.6, respectively. Lastly, section 2.7 provides insight into the data analysis methods, respective variables and data preparation, concluding this research methodology chapter.

2.1 Research Design Overview

The main research objective of this thesis is to study how user behaviour is impacted across varying 360-degree video image complexities amidst the multifaceted interaction process. As evident from the implications from current literature, the complex nature of 360-degree video interaction calls for an analytical approach which takes into account the multi-dimensional nature of the interaction process between user and 360-degree video. The assessment of multi-modality media, as defined by Wu et al. (2009), can be achieved by assessing the representative dimensions of cognitive perceptions and the subsequent behavioural consequences [356]. As discussed in the related works, many studies on 360-degree video user experience focus solely on technical parameters such as video or audio quality [260, 354]. However, research of the user interaction with 360-degree video require a complex and multifaceted approach, utilising both technical aspects and content characteristics to achieve a comprehensive analysis that encompasses both cognitive perceptions and behavioural consequences [9, 91, 289]. The significant influence of user preferences and external conditions heavily influence the perception of 360-degree video, thus rendering a purely objective approach ineffective and necessitating the inclusion of a subjective evaluation. User behaviour can be assessed through objective and subjective metrics, and are application specific [356]. In this thesis, and as defined by the sub-questions from § 1.7, the approach of this thesis entails physiological, objective and subjective analyses, as sensor-based physiological data was shown to be correlated to self-reported subjective measures. By utilising physiological data, the gap between objective and subjective measures is crossed, adding an extra dimension to the analyses [10, 91, 271].

The above-mentioned approaches have been translated into two parts that entail the entirety of the research design. The first part is a physiological and objective approach, employing an eye-tracking study (see § 2.2) to obtain measurements on the user's gaze based on fixation data. Participants were presented a series of 360-degree video content [194, 328], varying in respective spatiotemporal image complexities, and were prompted to freely interact with the system during which the eye-tracking software measured gaze data. The selection of content was based on respective spatial- and temporal image complexity of the 360-degree video sequences, selected to ensure sufficient coverage across the spatiotemporal matrix as detailed in § 2.2.3 [172, 290]. The eye-tracking study was followed by a user-centric evaluation study (see § 2.3), which enabled subjective analysis of the user's interaction and perception. As such, the user's gaze when viewing varying 360-degree video content, and how it varies depending on respective spatiotemporal complexities, was observed. This was followed by a subjective analysis of users' perception to gain further insight into the relationship between seating, gaze patterns and perceptual attributes such as engagement, attention, and spatial awareness. Moreover, as denoted by Ebrahimi et al. (2009), the significant influence of usage situation and usability context (i.e., environment) remains of-

ten disregarded in comparable literature [90, 362]. The unique challenge posed by the immersive nature of 360-degree videos is that the assumption of all viewers observing the exact same stimulus does not hold true for omnidirectional video interaction. This is further emphasised by the work of Brunnström et al. (2013), which defines the distinct levels of interaction applicable to omnidirectional video content in VR [42]. The freedom to choose where to look within the video results in variations in the visual experience and can potentially lead to different interpretations of the content [194]. Therefore, the implemented study design employed the inclusion of seating type as a third variable, as it can potentially influence the viewer's visual experience and subsequent interpretation of the content. In particular, the use of a rotating chair omits any limitations in the participant's movement, isolating the behavioural effect of variation in spatiotemporal image complexity. By including a limiting contextual factor, the use of a fixed-position chair, the effect size of spatiotemporal complexity on gaze behaviour can be compared across different seating types, adding an extra dimension to the findings. As such, the proposed research methodology enables user behaviour analysis across the three distinct domains of direct perception, interaction and usage situation [42].

This approach aims to reconcile the dichotomy between objective and subjective measurements by integrating both quantitative and qualitative methodologies, thereby providing a more comprehensive and nuanced understanding of the user interaction. To accommodate the inclusion of these variables, a mixed-methods design was implemented, enabling an extensive analysis of both within-subjects and between-subjects effects. This methodology allows for the manipulation of the independent variable (i.e., spatial complexity and temporal complexity) within-subjects, while the other independent variable (seating type) is manipulated between-subjects. The implementation of a mixed-method design over a 2×2 factorial design was decided based on considerations in generalisability and statistical power.

Due to the time-consuming nature of the eye-tracking study, it was decided to adopt a within-subjects study design for the sequenced viewing of 360-degree content, in which the participants viewed an entire sequence of subsequent 360-degree videos in VR. In order to minimise the learning-/ order effect, prominent in a within-subjects study design due to the transfer of knowledge, the order of videos in which they were presented was randomised using a Latin square design. Moreover, by implementing a within-subjects study design, the amount of random noise and occurrence of confounding variables in the data set was minimised. The participants were split in two groups, one which viewed the 360-degree video content on a rotating chair while the other group was seated on a fixed chair, by implementing a between-subjects design method. The variable chair type was not explicitly disclosed to the participants to avoid any potential bias in excessive rotating caused by their awareness of it. The specific group division and sequencing method, as well as content selection process are described in § 2.2.5 and § 2.2.3, respectively.

During the execution of the eye-tracking study, participants were required to wear an head-mounted display in order to view the 360-degree video sequences. As such, the participants were exposed to the risk of cybersickness or any other sense of dizziness or nausea caused by the discrepancy between their own movements and the motion of the virtual scene [40, 126, 328]. Furthermore, participants might experience a sense of fatigue, physical strain or discomfort from looking around the virtual scene extensively during the viewing sessions. To minimise the risk and impact of these effects, the following precautions were put in place.

During the recruitment process, participants were asked about risk inducing parameters, such as sensitivity to motion sickness or similar sense of dizziness, recent medical injuries or surgeries, and back, neck or other physical conditions that might be of risk to these effects. Participants with any of these conditions or sensitivities were excluded from participation, see § 2.5 for further elaboration on the population. The recruited participants were extensively informed on the above-mentioned physical risks and the potential of them occurring regardless. To further minimise the risk of these effect occurring during the eye-tracking study, a pre-test parameter study was performed. The main objective of the pre-test parameter study was to establish the parameters of the main eye-tracking study and minimise a sense of physical strain, discomfort or cybersickness. As a result, only short VOD content [81] was selected for the eye-tracking experiment. Furthermore, the eye-tracking experiment was conducted while being seated on an ergonomic chair to avoid

unnecessary physical exertion caused by standing up. The seated position on a rotating chair enabled full 360-degree rotations, while also reducing physical strain on the user's neck and spine. The material and apparatus utilised during the study is presented in § 2.4. An ergonomically adjustable HMD was used throughout the viewing sessions, and users were allowed to take breaks whenever necessary. Moreover, a short intermission was implemented between the eye-tracking study and the user evaluation to reduce over-stimulation and provide a moment of relaxation before conducting the user-centric evaluation. Lastly, only methodologies that were proven to minimise risk of cybersickness, such as the M-ACR method (see § 2.2.1), were utilised throughout this study.

2.1.1 Pre-Test Parameter Study

As mentioned, prior to conducting the main eye-tracking study, a pre-test parameter study was conducted. The pre-test parameter study was considered a pilot study, prior to the main eye-tracking study, and was primarily designed to define the parameters of the study in regards to the duration of events. The findings from relevant literature [126, 137, 158] indicate that the use of a HMD can result in cybersickness, fatigue or any other form of physical strain or discomfort, prompting the following question:

> Is there a realistic risk of a user experiencing physical or psychological harm or discomfort during the proposed experiment design of this research?

As such, it was decided to run the pre-test parameter study to minimise risk of these effects occurring as much as possible by using a small sample size to determine the optimal parameters. The pre-test parameter study was conducted as follows.

The pre-test parameter study incorporated the same set-up and procedure from the eye-tracking study and user evaluation. The complete eye-tracking study design is elaborated in § 2.2, as well as the user evaluation in § 2.3. There was no distinct difference with the main experimental process in terms of study design, procedure, and execution of the pre-test parameter study. Figure 1 visualises the additional SSQ assessment task as part of the pre-test parameter study, as well as a compressed version of the entire study procedure (see § 2.6). Firstly, during the recruitment process, users were asked about risk inducing parameters (e.g. motion sickness sensitivity or other physical conditions that form a risk). Furthermore, users were informed extensively and explicitly on the exploratory nature of the pre-test parameter study. They were informed about the exploratory nature of the study in which it was designed to establish the risk-reducing parameters of the main eye-tracking study, and that, while cautiously designed, participation in the pre-test parameter study could still result in the before-mentioned effects. They were also ensured that throughout the duration of the study, they would be closely monitored and that terminating the viewing session was possible at all times. Furthermore, the users were informed that participation was completely voluntary and at own risk. Moreover, due to the learning effect, users that participated in the pre-test parameter study were excluded from participation in the main study. The recruitment correspondence and information sheet of the pre-test parameter study can be found in appendices D13 and D14, respectively. Lastly, they were asked to provide consent, similar to the main eye-tracking study. This information was also provided during the introduction and start of the pre-test parameter study. The provided digital consent form and information sheet for the pre-test parameter study can be found in Appendix E17 and E15, respectively.

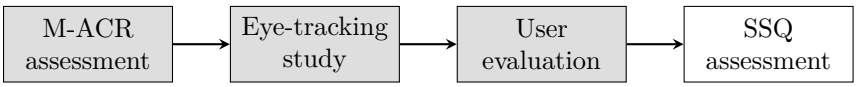


Figure 1: Pre-test parameter study procedure.

Following the introduction, users were instructed to perform the main eye-tracking as described in § 1.4. Throughout the experiment, the total duration of the experiment was monitored. Due

to the emphasis on free viewing and exploration, the users were free to watch as much of the 360-degree video as they pleased, resulting in distinct total duration times for each user. After watching the entire sequence of 360-degree video content during the eye-tracking study, and after performing the user evaluation (see § 2.3), they were instructed to fill out the additional simulator sickness questionnaire (see Appendix F19) [163]. The simulator sickness questionnaire, from the work of Singla et al. (2017) [290], was used to assess how much the users were prone to nausea or similar forms of discomfort when participating in the proposed experiment design. While filling out the SSQ, users were asked to provide the severity of a list of symptoms they might have experienced, such as fatigue, headache, eye strain, blurred vision, or increased salivation. Lastly, users were asked to rate the total duration of the experiment on the 5-point Likert scale presented in Table 1.

1	2	3	4	5
Too short	Short	Neutral	Long	Too long

Table 1: Duration rating scale.

The findings from the pre-test parameter study were used to establish the duration of the experiment and takes into account the results from the SSQ to establish parameters for main study design to minimise risk of cybersickness, physical strain or other physical discomfort. The resulting adjustments made as part of the eye-tracking study design can be found in § 2.2. Additionally, the pre-test parameter also monitored the duration, relevance and efficiency of the metrics used as part of the user evaluation (see § 2.3). The evaluation metrics, consisting of questions and statements, were revised due to observed considerations of redundancy, ambiguity or deviancy. Any reported unclear or digressive metrics were revised or omitted from the user evaluation questionnaire (UEQ) and semi-structured interview (SSI). The resulting accommodations as part of the pre-test parameter study are discussed in § 2.3.

Results The pre-test parameter study was conducted by 9 participants, comprising of 5 male and 4 female users. Their respective age ranged from 19 to 26 ($\mu = 22.11$ years, $\sigma = 2.09$). The majority of users had some slight experience with VR, with a mean score of $\mu = 2.11$ and $\sigma = 0.81$ on a 5-point Likert scale ranging from never (1) to once a week or more (5). All users had normal or corrected-to-normal vision, and no users reported visual impairments or had any experience with epilepsy/motion sickness or any physical condition that may be aggravated by using a VR headset. One user stated previous experience in similar research in the past. All users provided their informed consent.

The users could indicate the severity of the symptoms that could occur during VR usage on a 4-point Likert scale as presented in Appendix F19. Overall, the users reported experiencing minimal discomfort during the pre-test parameter study, as evidenced by the mean score for general discomfort ($\mu = 1.67, \sigma = 0.47$). Fatigue ($\mu = 2.00, \sigma = 0.94$) and eye strain ($\mu = 1.67, \sigma = 0.67$) were the subsequent most commonly reported symptoms, followed by headache ($\mu = 1.67, \sigma = 0.67$), difficulty focusing ($\mu = 1.67, \sigma = 0.82$) and blurred vision ($\mu = 1.44, \sigma = 0.68$). The remaining symptoms were reported less frequently with a relative lower severity. Furthermore, burping was least frequently reported across all symptoms.

While showing slight discomfort due to the duration of the experiment, the pre-test parameter study proved useful in identifying the risk-inducing parameters (i.e., fatigue and eye strain) that occurred while conducting the study. The results from the pre-test parameter study were taken into account and the eye-tracking part and user evaluation part were accommodated accordingly, as detailed in § 2.2 and § 2.3, respectively.

2.2 Eye-Tracking Study

The approach employed to acquire physiological and objective measures on the user's gaze behaviour was done utilising an eye-tracking study. As demonstrated by Arndt et al. (2014),

eye-tracking poses a valuable tool to assess QoE through the use of physiological sensors [10, 91]. Aside from allowing exploration and insight into cognitive perceptions [6, 213, 243, 256, 309], eye-tracking also enables the measurement of eye movement outside of conscious control [58, 169]. As before-mentioned, the focus of the eye-tracking study was to analyse the users' gaze behaviour when presented with a variety of 360-degree video content. The eye-tracking study was designed utilising the framework established by Carter et al. (2020), which takes into account the reliability, validity and technical challenges [49]. Due to the reliability on visual stimuli as an important variable, it was decided to apply a diagnostic eye-tracking approach to the study design [86].

The eye-tracking study aimed to analyse gaze behaviour as instinctive and naturally occurring as possible. The implications a comprehensive understanding of the behavioural impact of 360-degree video has on the future development of omnidirectional content are significant, similar to the significant influential factor of context in which they content is viewed [90]. Therefore, it was decided to subsequently study the difference in effect size across different usability contexts, utilising different seating types. Since the effect of different viewing context is beyond the scope of this thesis, it was decided to measure the most natural occurring viewing behaviour while maintaining a contextual continuity, enabling a higher external validity of the thesis. This also resulted in the eye-tracking study being conducted in a neutral setting, without any changes made to the location or viewing context. Due to the significant distinction between tasked viewing and free viewing [116], and to furthermore evoke the most natural occurring viewing behaviour from the user, it was decided to provide users with as much control over their own actions as possible. This was further realised by omitting video-specific tasks from the study, meaning that users were not instructed to complete a specific task (e.g. search-related tasks) other than viewing the video. As identified by Said et al. (2004), the user's behavioural actions are manifestations of engagement. Therefore, the unrestricted exploration of the virtual environment, combined with the user's ability view as much of the video as they pleased, addresses the spatiotemporal thread of engagement in order to provoke gaze behaviour [230, 270]. The other threads of engagement and cognitive influences are discussed in § 2.3. Lastly, the users were able to stop the viewing session if they feel disengaged with the video. It was not required to watch the entire video. However, to enable consistency in the interaction across all users, it was only allowed to terminate the viewing session. Other controls, i.e., play, pause or playback, were not allowed. The entire procedure, including the eye-tracking study specifics, is described in § 2.6.

Pre-Test Parameter Accommodations During the eye-tracking part of the pre-test parameter study, users were presented content with a duration of 2 to 3 minutes. This led to an increased duration of the experimental run, increasing fatigue ($\mu = 2.00, \sigma = 0.94$) among users and increased risk of physical discomfort ($\mu = 1.67, \sigma = 0.47$). Moreover, some users ($n = 5$) specifically stated not being interested enough in watching the entirety of each video. The majority of users found the total duration of the experimental run too long ($n = 6$). To accommodate these findings, reducing fatigue, physical risks and maintaining engagement, it was decided to limit each 360-degree video to a duration of 60 seconds.

2.2.1 M-ACR

The user's perception of the video poses an important factor in the eliciting of viewing behaviour. As established by Singla et al. (2017), exploratory behaviour is not only impacted by different levels of video resolution quality. As such, video quality of experience remains an important aspect of the user's QoE assessment. Video quality, as a term, commonly encompasses attributes such as video size and resolution. However, during this QoE assessment, this terminology will be used to define the quality of experience evoked by a 360-degree video. Regardless of the negligible influence video resolution has on viewing behaviour, the difference in perceived QoE holds significant impact on levels of immersion and exploratory behaviour (i.e., high quality leads to more exploratory behaviour) [376]. Furthermore, the subjective assessment of QoE entails sensor-based, perceptual and usability aspects [91, 105, 271, 277, 292, 371]. Therefore, this study aims to take into account the difference of perceived QoE of each user across all videos, enabling

analysis on the behavioural effects caused by perceived QoE. Moreover, increasing the internal validity of the study, it was essential to measure the QoE of each video across all users. Measuring the QoE of each video per user enabled the examination of individual variability in the subjective experience of each user. This approach is particularly useful in distinguishing how behavioural consequences are influenced by QoE values at the user-level. By analysing the individual QoE scores, variations can be identified in how the users respond to different videos, emphasising the unique distinctions in perception, interpretation and preference of each user.

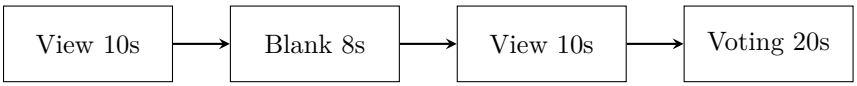


Figure 2: M-ACR presentation sequence of one fragment stimulus.

The study by Yaqoob et al. (2020) identified various subjective approaches for 360-degree video with respect to perceptual, cybersickness and sensor-based QoE aspects [362]. An extremely effective approach with regards to the before-mentioned QoE aspects and the inclusion of content characteristics, is the utilisation of the Absolute Category Rating (ACR) method, as recommended by Tran et al. (2017), Yaqoob et al. (2020), and related works [105, 137, 239, 328, 362, 371]. However, ACR is not designed to take into account to the omnidirectional nature of 360-degree video and longer-length viewing sessions on an HMD. Therefore, it was decided to adopt the M-ACR scoring method [233, 290], before starting the eye-tracking viewing session, to this study, as presented in the work by Singla et al. (2017). Users were asked to subjectively assess the videos they were about to watch.

Traditionally, ACR presents the sequence one at a time, commonly with a duration of approximately 10 seconds, after which the user is asked to evaluate its quality so that the sequences are rated independently of each other. However, the M-ACR adaptation of the traditional ACR scoring method presents each fragment twice, enriched with a short blank sequence in between and followed by the voting time afterwards. After the double presentation of each video fragment, users were asked "How is your assessment about the perceptual quality of the video on the scale from 1 to 5?". Users then said aloud their rating, which was noted by the experimenter, allowing the user to continuously wear the HMD throughout the M-ACR assessment, similar to the study by Singla et al. (2017) [290]. Users were given a maximum of 20 seconds to vote. Voting scores were acquired through the Mean Opinion Score (MOS), enabling users to give subjective measurements on a five-level Likert scale, see Table 2 [144]. The M-ACR procedure is visualised in Figure 2. This repetition of sequences is due to the priming effect, as most people are not used to watching 360-degree videos regularly. Conducting the M-ACR assessment prior to the eye-tracking study enabled users to get acclimated to the VR environment, mitigating learning bias. Since most viewers are not accustomed to watching 360-degree videos frequently, the initial sequence serves as a primer to familiarise them with omnidirectional video content. This approach enhances the validity of the ratings obtained when the sequence is shown for the second time. The M-ACR assessment prior to the longer-length viewing session during the eye-tracking study also counteracts any unpredictable or curious viewing behaviour evoked from the relative novel experience of watching a 360-degree video with VR-equipment. The inclusion of M-ACR session and the entire procedure can be found in the § 2.6. The Modified Absolute Category Rating was designed and conducted following the guidelines and framework as established by ITU-T Recommendation P.910 for subjective video quality assessment for multimedia applications [144, 233].

Other approaches were considered during the selection of the subjective assessment method, namely DCR and DSIS. Degradation Category Rating (DCR) is usually focused on the objective quality of a system, making it a more suitable method for testing fidelity or transparency, rather than subjective perception [144]. Alternatively, an approach more applicable to omnidirectional video is the Double-Stimulus Impairment Scale (DSIS), which is considered less reliable than M-ACR and increases the risk of cybersickness occurring, which this study aims to minimise [144, 293]. Due to these reservations, DCR and DSIS were not implemented during this study.

1	2	3	4	5
Poor	Bad	Fair	Good	Excellent

Table 2: ACR subjective quality scale.

2.2.2 Database

The 360-degree content that was presented during the eye-tracking study was acquired from the YouTubeVR dataset, and was selected using the database categorisation by Afzal et al. (2017), in which a substantial portion of the YouTube 360-degree video dataset was categorised and catalogued based on genre [2]. The content characterisation by Afzal et al. (2017) distinguishes a total of 14 genres across the video dataset. The dataset consists of $n = 2285$ videos, of which the genres of roller coaster ($n = 325$), scenery ($n = 315$), animals ($n = 216$), cartoon ($n = 197$) and video game ($n = 197$) account for the top 5 genres in terms of quantity. However, due to the commercialisation and standardisation of 360-degree content, the categorisation of Afzal et al. (2017) does not accommodate for the exponential growth of the YouTube video database nor includes novel genres [8, 54, 319, 320, 375].

EAC to ERP Conversion The 360-degree video content, acquired from the YouTube database, was by default in the equiangular cubemap (EAC) format. In order to implement it in the iMotions eye-tracking software and use it throughout this study, the EAC format had to be converted to equirectangular (ERP) format. To ensure a minimal overhead conversion and elicit high-quality results, the OpenCV library was implemented, which is a highly optimised computer vision and image processing library. The conversion of a 360-degree video from the EAC format to ERP included the implementation of a custom Python script, which utilises the OpenCV library. The custom Python script for EAC to ERP conversion is provided in Appendix B4.

The script reads the EAC format input video, extracts the cubemap representation from the top half of the video frames by processing the frames and combining the top and bottom halves, and writes the resulting equirectangular frames to a new output video file. This is achieved by resizing the video frames using OpenCV's `resize` function and bicubic interpolation. By using OpenCV's `VideoCapture` and `VideoWriter` classes, the input video file was read and output video file is written. Moreover, the script checks for each 360-degree video if it adheres to the EAC format.

Initially, attempts were made to utilise FFmpeg, a widely-used multimedia framework, to perform the conversion. As supported by FFmpeg, the v360 filter is capable of converting formats, filters and codecs across the 360-degree video formats. To achieve this, the following command syntax was used: `ffmpeg -i EAC.mp4 -vf "v360=eac:equirectangular" ERP.mp4`. However, due to compatibility issues with the required v360 filter, the FFmpeg approach could not be successfully executed. Consequently, the custom Python script using OpenCV was implemented. To ensure compatibility with the iMotions eye-tracking software, the `.mp4` files were converted to `.wmv` files.

2.2.3 Set of 360º Content and Systematic Selection

The eye-tracking study was designed on the premise of users viewing a selection 360-degree video content, which represented significant changes in spatial- and temporal complexity. However, to acquire suitable 360-degree content, adhering to this requirement, a systematic selection was conducted. This subsection elaborates on the 360-degree video content selection process, in which a systematic approach based on primary filtering and visual criteria was taken to select six 360-degree videos, that would isolate and identify videos of varying levels of spatial- and temporal complexity from the database. The selected 360-degree video sequences utilised during the eye-tracking study are discussed in this subsection and visualised in Figure 3, containing equirectangular projected frames from each video. Table 3 presents the identification of each selected 360-degree video.

(a) A1 (b) A2 (c) B1

(d) B2 (e) C1 (f) C2

Figure 3: Frames of the selected 360-degree content in ERP.

Each video was cut into shorter clips of 60 seconds to adhere to the total experiment duration parameters of the study (see § 2.2). The visual elements and contents of each of the 360-degree videos are detailed in Chapter 4. A brief description of the 360-degree video sequences is presented below:

- A1: A seated perspective while a lion approaches, in an African landscape.
- A2: Fast-paced skiing down a mountain slope covered in snow.
- B1: First-person perspective of the stationary blocks in a game of Tetris.
- B2: Running down the subway tracks in a high-speed video game.
- C1: Slow moving roller coaster-type ride through a canyon landscape.
- C2: High-speed roller coaster during a Californian sunset.

Filter The selection of content based on their respective spatial- and temporal image complexity was challenging, as computing the spatiotemporal complexity of the entire 360-degree video database proved demanding. As a means to optimise the selection process, the database was filtered on visual attributes that are indicative of various levels across the spatial and temporal planes.

As found by Cui et al. (2021), significant changes to the spatiotemporal complexity of a video sequence occur on the spatial and temporal planes, indicated by changes in genre-specific characteristics [72]. Video genre-specific characteristics display distinct visual styles and structures that influence the spatial image complexity of a video sequence, closely relating genre and spatial complexity. Therefore, video genre was used as the primary, initial indicator of significant variation in spatial complexity. The selection based on distinct genres ensured coverage across the various levels of the spatial plane. Due to the large number of experimental runs required to extensively assess each existing genre and respective spatial complexity, it was decided to use a subset of genres. Consequently, enabling the comparison of the effect of spatial complexity on viewing behaviour, three indicative genres were used as a primary filter to select videos with enough variation in spatial complexity. By first filtering the database on genre, the relative position of each video along the spatial plane could be estimated, as the specific spatial complexity was later computed to define the specific values across the spatiotemporal matrix (see Figure 4).

Furthermore, enabling comparison of the effect of temporal image complexity on gaze behaviour, selection was also based on factors that contribute to the 360-degree video's temporal complexity. Movement of the camera, object displacement, altered scenery and new perspectives

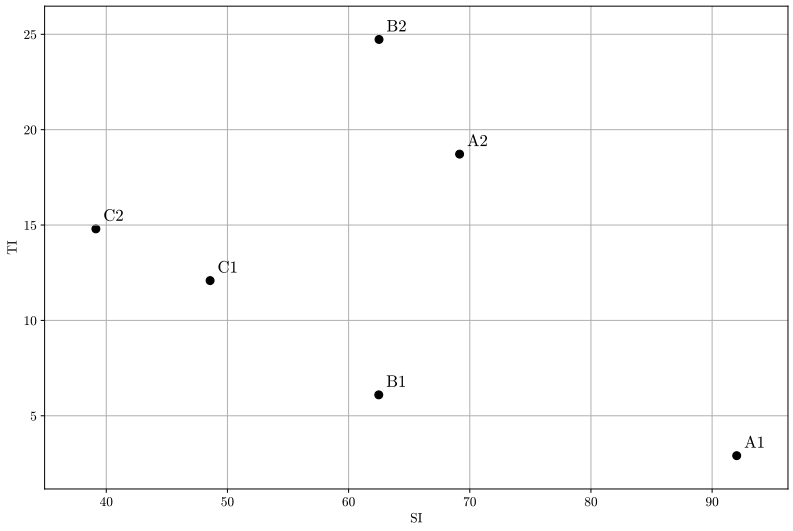


Figure 4: Spatiotemporal Matrix

contribute to the number of visual changes occurring over time in consecutive frames. As such, these visual changes over time determine the degree of temporal complexity of a video sequence. However, as described in "Visual Criteria", only type E3 continuity edits were selected, highlighting movement of the camera as a primary temporal factor. As such, camera movement was used as the primary, initial indicator of significant variation in temporal image complexity. Camera motion in this thesis refers to the amount of – relative – motion in which the observer moves throughout the 360-degree video sequence. The filtering based on camera motion enabled representation of significant changes on the temporal plane across the 360-degree video selection. Similar to selecting video sequences of varying spatial complexity, by using distinct levels of camera motion, it was possible to estimate its position on the temporal plane as the precise temporal complexity was computed later on. As such, the varying degree of spatiotemporal complexity across the selected videos was estimated by using distinct genres and amount of camera motion to ensure sufficient coverage of both spatial and temporal planes within the spatiotemporal matrix. Detailed in § 2.2.4, the selection process utilised the computation of SI- and TI-values to ensure the selected 360-degree video sequences provided sufficient coverage across the spatial and temporal planes (see Figure 4).

The six selected 360-degree videos, represent sufficient change in spatiotemporal complexity, were of genres: scenery (A), video game (B) and roller coaster (C), all containing both a relative low (1) and relative high (2) amount of camera motion. To ensure this representation and a greater degree of validity and reliability, the following considerations were made in determining the distinctive coverage of the spatiotemporal planes due to genre variation. For instance, genres scenery and animals share visual similarities, and were therefore not both included in the final selection. Scenery provides incredibly rich and detailed image structures, containing high levels of visual richness (i.e., amount of textures, objects, edges and features). Video game content contains less detailed textures and objects and varies significantly in the level of camera motion, thus being a valuable inclusion in the selection due to the variability of the genre and lower estimates of spatial complexity. Lastly, to include a high-motion type genre, roller coaster videos were selected, enabling observation of effects with higher relative camera motion as well as high levels of visual richness.

Similarly, to study the effect size of temporal complexity on viewing behaviour independently

of any effect offset by change in spatial complexity, two videos were selected from each of the three genres. Since a relative low camera motion within the roller coaster genre could be considered a relative high amount of camera motion in other genres such as scenery, the level of camera motion was considered relative throughout this thesis. The two videos from each genre contained one video with relative static / low camera motion, and another with relative dynamic / high camera motion. This approach enabled the cross-examination of the effect of temporal complexity within each genre-specific video, while controlling for the effect of spatial complexity. The cross-combination of 360-degree video with varying visual richness and changes over time provides sufficient representation of variations in both spatiotemporal complexity planes. The position of each selected video within the spatiotemporal matrix is presented in Figure 4. In total, a set of six videos were selected for this research, which is in accordance with the ITU-T Rec. P.910 framework [144]. A specification of the six 360-degree videos is presented in Table 3.

ID	Title	Genre	Camera Motion	SI-value	TI-value
A1	Lion	Scenery	Low	92.029	2.908
A2	Ski	Scenery	High	69.156	18.718
B1	Tetris	Video game	Low	62.489	6.097
B2	Subway	Video game	High	62.510	24.729
C1	Canyon	Roller coaster	Low	48.569	12.087
C2	California	Roller coaster	High	39.143	14.799

Table 3: Selected 360-degree content and their feature specifications.

Visual Criteria In § 1.6, a theoretical foundation on the perception of omnidirectional content, attentional guidance mechanisms, and user engagement in relation to their significant influence on viewing behaviour is presented. The resulting implications from relevant literature and comparable studies were crucial in the selection process, as visual factors (e.g. objects in the scene) significantly influence viewing behaviour. This subsection discusses the considerations made during the selection process to control for confounding effects of visual artefacts.

Firstly, Serrano et al. (2017) found that the misalignment of ROIs after a cut or edit evoke higher levels of exploratory behaviour [281]. Therefore, to minimise the effect of cinematographic principles on viewing behaviour, it was decided to select single-cut videos of a type E3 continuity edit in which space, time and action are continuous, omitting any potential ROI misalignment. Importantly, the majority of video content is edited using action discontinuity (type E1). As such, a significant part of the total database was disregarded.

The study by Tran et al. (2017) demonstrated the effect of camera motion on the user's self-reported sense of presence [328]. The linear relationship between presence and user engagement [156, 192, 230, 376] further necessitated the inclusion of distinct types of camera motion in the subset of videos. Since a moderate level of camera motion pertains to higher levels of presence, it was decided to include both static and dynamic video types, specified as fixed camera position and relative moderate camera motion, respectively.

The work by Zou et al. (2018) furthermore emphasises the positive effect of high resolution content on the overall user experience and on exploratory behaviour [376]. However, the analysis of gaze behaviour by Singla et al. (2017) indicates that video quality exhibits no significant impact on exploratory behaviour [290]. To account for the discrepancies in these publications, and to induce consistency across all stimuli, all selected videos were FHD (1920x1080p) resolution and MPEG-4 AAC codecs. Moreover, due to the temporal impact of using varying frame rates, only 360-degree video content with a frame rate of 30 fps was selected.

Lastly, to allow for control over consistency in camera motion and perspective, only egocentric videos were selected. Egocentric videos, as proposed by Xu et al. (2018), positions the observer as an action doer [359]. In this first-person perspective, the digital camera is mounted on the observer's head or body. The observer poses as an active element, rather than a passive element,

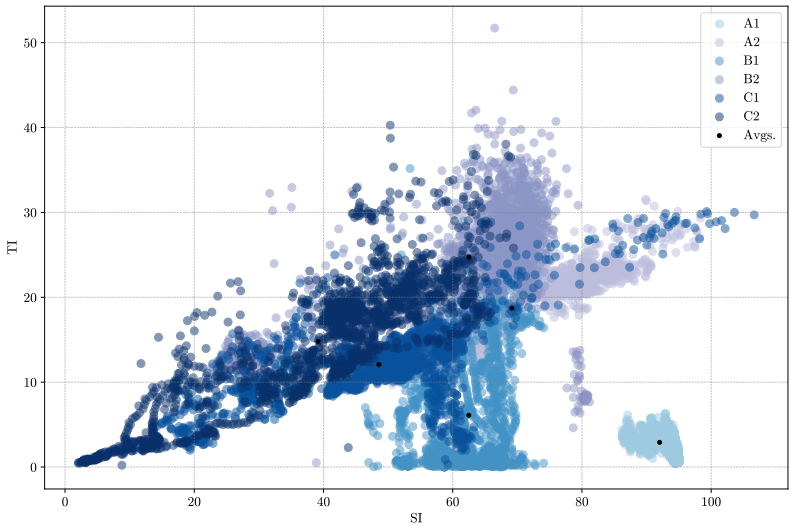


Figure 5: Spatiotemporal matrix (per-frame distribution).

and either navigates through the environment or observes events in their vicinity [186, 361]. Furthermore, the first-person perspective enables accurate relative visual changes based on the head and body movements of the observer. The resulting higher levels of presence and immersion elicit a more natural and intuitive user interaction and further attributes the observed behavioural effect to the changes in genre and camera motion. Consequently, the control over differences in perspective and camera motion consequently increases the internal validity of the study.

The main aim of the selection process was to isolate as much influential content characteristics to increase validity of the results and ensure the behavioural effect is induced by variations in spatiotemporal complexity. While the before-mentioned visual criteria were applied to carefully select the set of 360-degree video sequences, the selected videos still naturally contain unique content-specific visual artefacts of influential significance to user behaviour which can not be neglected. The reliance of user gaze on the amount of guided attention, achieved through content-specific visual artefacts and diegesis [223, 262, 306], necessitates consideration of present attention guidance mechanisms in each selected video. In addition, to elicit natural and instinctive viewing behaviour, it was decided to exclude content pertaining non-diegetic artefacts (i.e., arrows and text). Furthermore, it was decided to only select content containing minimal diegetic artefacts, as excluding diegesis in its entirely significantly limited the remainder of available content. Due to this inextricable nature of diegesis and the influence thereof on presence and user engagement, it was decided to explicitly highlight and take diegetic artefacts and visual attention guidance mechanisms into consideration in the analysis of gaze behaviour. Therefore, a separate, extensive diegetic assessment was performed. The diegetic assessment holistically approaches, codes and analyses the visual artefacts present in each selected 360-degree video and assesses the potential influence on gaze behaviour based on attention guiding mechanisms and influential diegetic attributes. The conducted diegetic assessment is presented in Chapter 4.

2.2.4 Spatiotemporal Complexity Specification

Selection of 360-degree video content was done in consideration with respective spatiotemporal complexity of each video sequence, in accordance with the defined framework in ITU-T Rec. P.910 [144]. The quantification framework, as presented by ITU-T Rec. P.910, was employed

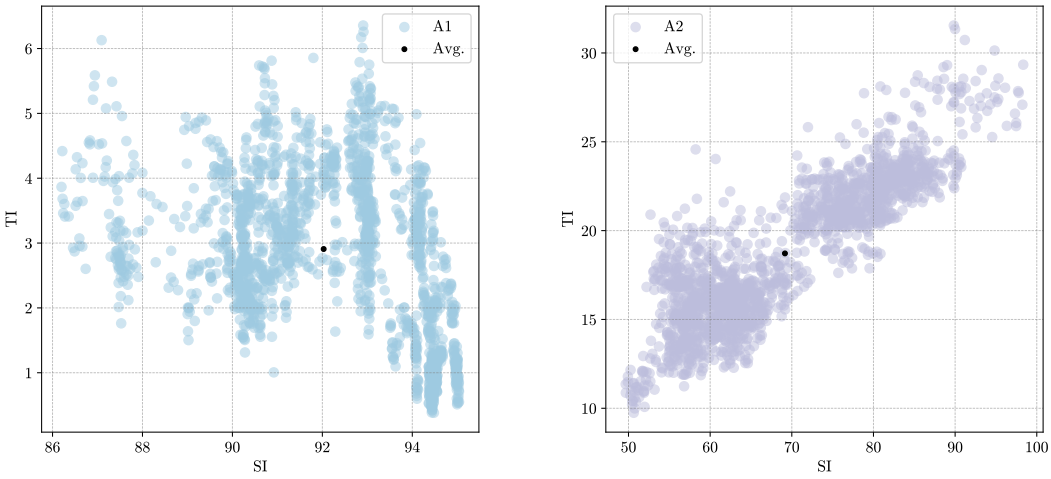
to compute the spatial- and temporal image complexities of each of the six selected 360-degree video sequences and was used to ensure comprehensive representation across the spatiotemporal planes [144]. The selected 360-degree videos represent various configurations of both low and high SI- and TI-values. The computation and mathematical construct for both spatial- and temporal image complexities of a video sequence is detailed in § 1.5.1. As such, 360-degree content was selected which sufficiently contained varying complexities across the spatial and temporal planes, visualised by the spatiotemporal matrix in Figure 4. The location of each selected 360-degree video sequence in the spatiotemporal matrix represents the degree of image complexity, and is a quantitative representation of the 360-degree video in terms of space and time.

The spatial image complexity of each 360-degree video sequence was computed using the Spatial Perceptual Information measure (SI). This approach entails applying a Sobel filter to each frame of the video at a given time, denoted as Fn, to generate a time series of spatial information. The Sobel filter implementation involves convolving two $3x3$ kernels, $Gv(i,j)$ and $Gh(i,j)$, over the video frame, where the input image pixel is represented as $x(i,j)$, and the output is $y = \text{Sobel}(x)$. After calculating the standard deviation of the pixels in the Sobel-filtered frames, the SI-value was determined as the maximum value of these standard deviations (1). By considering the variations in pixel intensities, which are indicative of the presence of edges and texture in the video frames, Consequently, the spatial complexity of a video sequence was computed by taking the variation in pixel intensities, indicative of the presence of edges and texture in the video frames, into consideration.

Similarly, the temporal complexity of each 360-degree video sequence was computed by leveraging the Temporal Perceptual Information measure (TI). This method leverages the motion difference feature, $M_n(i,j)$, which calculates the difference in pixel values (from the luminance plane) at the same spatial location in consecutive frames of the video. The feature $M_n(i,j)$, as a function of time (n), identified variations in pixel intensities across adjacent frames, capturing the motion present in the video sequence. The computation of the standard deviation over space (std_{space}) of the motion difference feature, $M_n(i,j)$, for all i and j, generated a time series of values. The maximum value of this time series (max_{time}) represents the TI of the video sequence (6). Therefore, a higher TI-value signifies greater levels of motion between adjacent frames, effectively quantifying the temporal complexity of the video content.

In total, the SI- and TI-values of 13 360-degree video sequences were computed, of which only the final selection of six videos ensured sufficient coverage, as other videos did not offer sufficient variance in spatial- and temporal complexity. A custom Python script was developed to compute the spatial- and temporal complexities of each 360-degree video sequence. The script utilises `OpenCV` (Open Source Computer Vision Library) to perform the video processing tasks. OpenCV is an open-source computer vision and machine learning software library. OpenCV was used to read each video frame, apply the required Sobel filters and convert each frame to greyscale images. Furthermore, the script uses `NumPy` to perform the required array-based operations and calculations. Lastly, the `Matplotlib` library was used to plot the spatial- and temporal complexities in a scatter plot. For each of the six selected 360-degree videos, a minimum of $n = 1800$ frames have been processed. The cumulative per-frame plot is presented in Figure 5. The Python script for calculating the spatial- and temporal complexity of each video sequence is presented in Appendix B5. The exact value range and spatial- and temporal complexities of the 360-degree videos is detailed as follows.

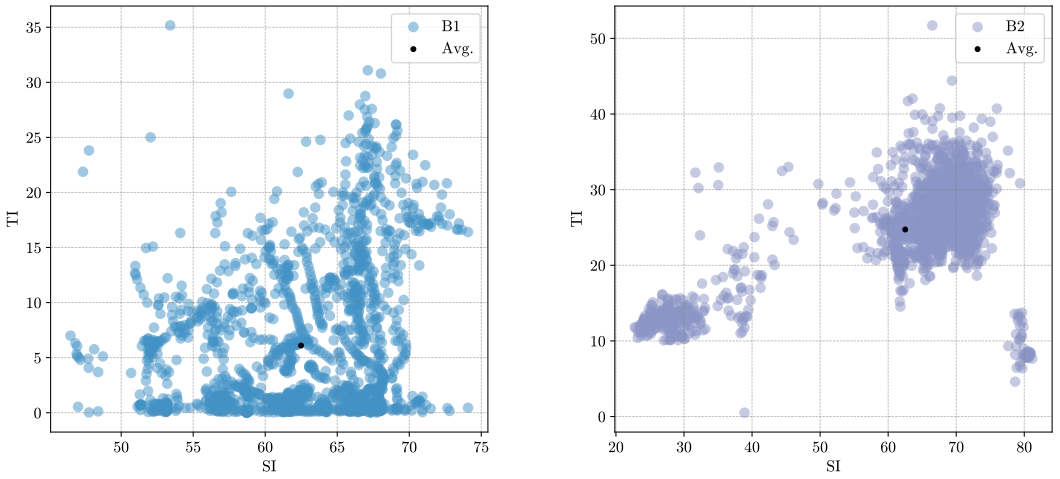
A1 The spatiotemporal complexity of video A1 was computed over a total of 1812 individual 360-degree video frames in equirectangular projection. The $SI_{min,max}$ range $[86.199, 95.039]$, with an average SI-value = 92.029, denotes the degree of image complexity on the spatial plane. The $TI_{min,max}$ range $[0.383, 6.355]$, with an average TI-value = 2.908, denotes the degree of image complexity on the temporal plane. Figure 6a presents the per-frame distribution of SI- and TI-values.



(a) A1 (b) A2

Figure 6: Spatiotemporal matrices of A1 and A2 (per-frame distribution).

A2 The spatiotemporal complexity of video A2 was computed over a total of 1885 individual 360-degree video frames in equirectangular projection. The $SI_{min,max}$ range $[49.658, 98.341]$, with an average SI-value = 69.156, denotes the degree of image complexity on the spatial plane. The $TI_{min,max}$ range $[9.747, 31.546]$, with an average TI-value = 18.718, denotes the degree of image complexity on the temporal plane. Figure 6b presents the per-frame distribution of SI- and TI-values.

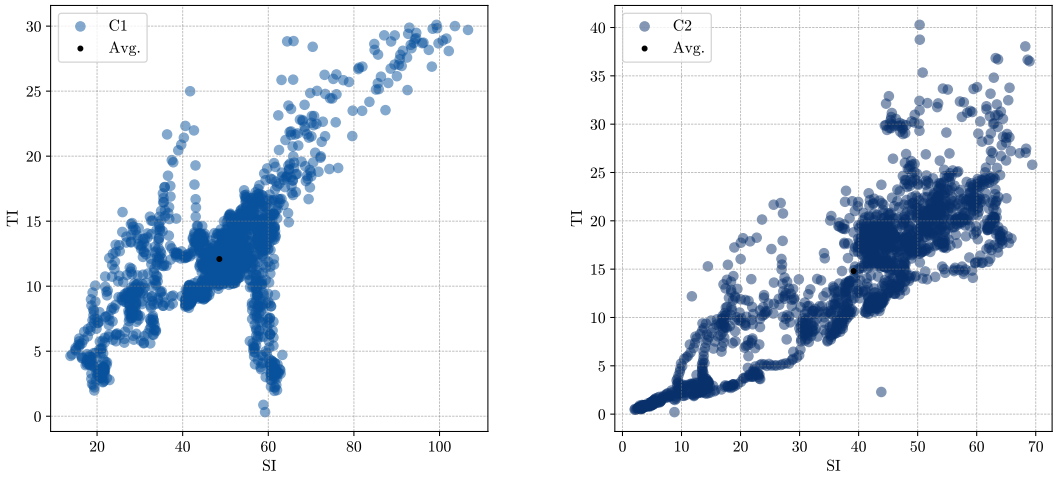


(a) B1 (b) B2

Figure 7: Spatiotemporal matrices of B1 and B2 (per-frame distribution).

B1 The spatiotemporal complexity of video B1 was computed over a total of 1948 individual 360-degree video frames in equirectangular projection. The $SI_{min,max}$ range $[46.481, 74.082]$, with an average SI-value = 62.489, denotes the degree of image complexity on the spatial plane. The $TI_{min,max}$ range $[0, 35.167]$, with an average TI-value = 6.097, denotes the degree of image complexity on the temporal plane. Figure 7a presents the per-frame distribution of SI- and TI-values.

B2 The spatiotemporal complexity of video B2 was computed over a total of 1850 individual 360-degree video frames in equirectangular projection. The $SI_{min,max}$ range $[22.757, 81.169]$, with an average SI-value = 62.510, denotes the degree of image complexity on the spatial plane. The $TI_{min,max}$ range $[0.509, 51.715]$, with an average TI-value = 24.729, denotes the degree of image complexity on the temporal plane. Figure 7b presents the per-frame distribution of SI- and TI-values.



(a) C1

(b) C2

Figure 8: Spatiotemporal matrices of C1 and C2 (per-frame distribution).

C1 The spatiotemporal complexity of video C1 was computed over a total of 1931 individual 360-degree video frames in equirectangular projection. The $SI_{min,max}$ range $[13.794, 106.655]$, with an average SI-value = 48.569, denotes the degree of image complexity on the spatial plane. The $TI_{min,max}$ range $[0.316, 30.099]$, with an average TI-value = 12.087, denotes the degree of image complexity on the temporal plane. Figure 8a presents the per-frame distribution of SI- and TI-values.

C2 The spatiotemporal complexity of video C2 was computed over a total of 1812 individual 360-degree video frames in equirectangular projection. The $SI_{min,max}$ range $[2.071, 69.406]$, with an average SI-value = 39.143, denotes the degree of image complexity on the spatial plane. The $TI_{min,max}$ range $[0.209, 40.278]$, with an average TI-value = 14.799, denotes the degree of image complexity on the temporal plane. Figure 8b presents the per-frame distribution of SI- and TI-values.

2.2.5 Sequencing Method

Due to the within-subjects design of the study, all selected videos were presented in sequence to each user. Regardless of the priming effect of the M-ACR method, see § 2.2.1, the amount of

exploratory behaviour is influenced due to the order sequence. This is further emphasised by Serrano et al. (2017), which found peak levels of exploration at the beginning of watching VR content [281]. To accommodate for the learning and order effect, it was decided to counterbalance the order of videos across users. During the viewing session, each user was assigned a specific – randomised – order in which the selected 360-degree video sequences was presented. This subsection discusses the specific sequencing method that was applied to optimise the order randomisation, and discusses the group specification for the comparison of gaze dynamics across different usability contexts.

The order of videos was determined using Latin Square Design (LSD), enabling higher levels of control over the effects of the extraneous variables (i.e., learning effect) [162, 258]. By implementing a Latin square to the experiment design, the main effects of both spatial- and temporal complexity were significantly isolated, while also enabling high levels of control for the effect of any confounding variables that might affect the validity of the results. The order-effect is minimised as each 360-degree video appears the same amount of times as the first, second, third, fourth, fifth and sixth in the sequence. The Latin square design resulted in a matrix containing the ordering sequence, in which each stimulus occurs only once in each row and once in each column. To minimise the carry-over effect often produced by the Latin square design, the matrix was balanced using the methodology presented by Bradley et al.(1958), generating a balanced Latin square design matrix, Ls, in which a stimulus precedes another exactly once [35]. The Latin square matrix Ls was used to determine the sequencing order as follows:

$$Ls_R = \begin{bmatrix} A1 & A2 & B1 & B2 & C1 & C2 \\ A2 & B1 & A1 & C2 & B2 & C1 \\ B1 & C2 & A2 & C1 & A1 & B2 \\ C2 & C1 & B1 & B2 & A2 & A1 \\ C1 & B2 & C2 & A1 & B1 & A2 \\ B2 & A1 & C1 & A2 & C2 & B1 \end{bmatrix}$$

The Latin square matrix randomised the order in which users viewed the videos, ensuring that each video appeared an equal number of times in each position in the sequence. The resulting Latin square array is denoted as:

$$\begin{aligned} A_{Ls_R} = \{&(A1, A2, B1, B2, C1, C2), (A2, B1, A1, C2, B2, C1), (B1, C2, A2, C1, A1, B2), \\ &(C2, C1, B1, B2, A2, A1), (C1, B2, C2, A1, B1, A2), (B2, A1, C1, A2, C2, B1)\} \end{aligned} \tag{19}$$

The array was used to establish a balanced design with six videos and two factors (video type and sequence position), ensuring that each video appeared exactly once in each position within a block of six videos. The array only contains the first six sequences, the array was repeated across all users, such that the seventh user was presented the first sequence, the eight user the second sequence, and so forth. Due to the exclusion of results acquired during the pre-test parameter study, the Latin square was not used to generate a randomised sequence of 360-degree video during the pre-test parameter study.

For the sequencing of the 360-degree video content, other approaches were considered equally. A prominent alternative method was to implement an orthogonal array [33, 179, 183] to determine the sequencing order. This approach was disregarded based on the following considerations. Firstly, an orthogonal array requires a large sample size and amount of experimental runs to achieve an equivalent level of precision as the balanced Latin square. Secondly, the Latin square design allows for a design that controls for the effects of the known nuisance factors, whereas orthogonal arrays account for a multitude of factors as $n > 2$, where n is the number of nuisance effects. Lastly, orthogonal arrays can result in differences in variance between the different combinations of factor

levels. To ensure homoscedasticity, i.e., homogeneity of variance, it was decided to adopt the Latin square design.

2.2.6 Group Specification

The moderating effect of usability context was studied using a between-subjects approach, utilising two different seating types. This was achieved by splitting the sample size into two distinct groups R and F, each utilising either a rotating chair or a fixed-position chair, respectively. The identification and distinction of acquired data between both groups was done by using the distinct group ID R and F, as described in § 2.7.2. Users were split equally amongst the two groups, in order R, F, R, F. As such, the same Latin square matrix was adopted across both groups, of which the sequencing order for group R is presented above, where $Ls_R = Ls_F$ and $A_{Ls_R} = A_{Ls_F}$. The resulting Latin square matrix, applicable to group F, is denoted as:

$$Ls_F = \begin{bmatrix} A1 & A2 & B1 & B2 & C1 & C2 \\ A2 & B1 & A1 & C2 & B2 & C1 \\ B1 & C2 & A2 & C1 & A1 & B2 \\ C2 & C1 & B1 & B2 & A2 & A1 \\ C1 & B2 & C2 & A1 & B1 & A2 \\ B2 & A1 & C1 & A2 & C2 & B1 \end{bmatrix}$$

The resulting Latin square array for group F is denoted as:

$$\begin{aligned} A_{Ls_F} = \{&(A1, A2, B1, B2, C1, C2), (A2, B1, A1, C2, B2, C1), (B1, C2, A2, C1, A1, B2), \\ &(C2, C1, B1, B2, A2, A1), (C1, B2, C2, A1, B1, A2), (B2, A1, C1, A2, C2, B1)\} \end{aligned} \tag{20}$$

2.3 User Evaluation

Highlighted by the implications from existing literature, and reiterated in § 1.7, the 360-degree video interaction encompasses a complex and nuanced interplay of cognition, perception, usability and user behaviour. As such, the multifaceted interaction process necessitates both objective and subjective methodologies to adequately assess the complex dynamics thereof.

The objective approach, realised through the eye-tracking study, was fundamental in the generation of gaze data. However, as emphasised by Wu et al. (2009), the complex assessment of cognitive perceptions and subsequent behavioural consequences necessitates the inclusion of a subjective evaluation, as objective approaches remain limited in the ability to encompass a comprehensive analysis that takes into account the significant influence of user preference and perception [9, 356]. As further defined by Wu et al. (2009), the multi-dimensional construct of QoE [76, 356] represents the cognitive and behavioural responses while watching 360-degree video, which also require the assessment of context influential factors, such as user expectations and perception [90]. Therefore, the significant influence of these factors in impacting cognitive perceptions, eliciting behavioural responses, cannot be disregarded in the behavioural assessment.

The inclusion of a subjective evaluation is further emphasised by the work of Holmqvist et al. (2011), which highlights that generated eye-tracking data (i.e., gaze coordinates and heatmaps) solely provide information on fixations, lacking the necessary insights into the underlying cognitive perceptions that influence the user's gaze [130]. Similar to the work by Egan et al. (2016), which presents a correlation analysis between objective gaze data and subjective self-reported measures [91], the subjective data from this study was acquired through the inclusion of a post-test user-centric evaluation.

The user evaluation protocol consisted of two parts: the user evaluation questionnaire (UEQ) and a semi-structured interview (SSI). The questions used were designed to address specific areas of research foci and sub-foci, as outlined in this subsection. The UEQ and SSI protocol can be found in appendices F20 and F21, respectively. The entire evaluation protocol is detailed as follows:

- Introduction: introducing the main objective of the evaluation.
- Part I: the user's overall experience and well-being after conducting the experiment is evaluated.
- Part II: the user's experience is evaluated through a set of questions and statements related to attention, engagement, spatial awareness and usability context (UEQ).
- Part III: the user's self-perception on viewing behaviour is assessed through a set of pre-defined questions (SSI).
- Part IV: based on the acquired answers, follow-up questions are provided which delve deeper into the given answers or provide clarity.
- Conclusion: thanking the user for their time and provide time for additional questions or suggestions.

The placement of the evaluation within the entire experiment process was taken into consideration during the design process. Specifically, there were several reservations that were taken into account during the decision-making of when to evaluate users on the presented 360-degree content. This resulted in the decision to run the evaluation across the entire sequence after the entire presentation of content, rather than after each individual 360-degree video sequence. Firstly, running the evaluation after each presented video required the disassembly of the HMD setup, which increases risk of physical discomfort. Secondly, the constant change in virtual orientation and real-life spatial orientation further increased risk of motion sickness due to the discrepancy in spatial orientation. Lastly, running the subjective evaluation after the presentation of all content variations enabled subjective analysis on the correlation and differences across content types, as users were able to compare across the content sequence. Coupled with the within-subjects design, this approach enabled the evaluation of various stimuli under the same cognitive perception levels per user. However, this methodology does depends on stimulated recall across the users, which is based on a selected interpretation of the content experience, rather than an exact replica of the experience [205]. Therefore, the subjective measurements rely on the user's ability to recall. This limitation, as identified by McCarthy et al. (2004), is further discussed in Chapter 6. However, during events of intense cognitive activity, such as the proposed eye-tracking experiment in this study, self-reported measures that relate to highly memorable experiences are significantly more reliable as compared to alternative methodologies such as think-aloud protocols [93]. While the evaluation data relies on the user's recollection and interpretation, the reflective data on their experience provides valuable insights into the memorable aspects of their interaction. It was therefore decided, in the context of this research, to run the evaluation after completion of the entire eye-tracking study.

The first part of the user evaluation was conducted using a user evaluation questionnaire (UEQ). In accordance with the framework presented by Wu et al. (2009), the subjective evaluation of behavioural consequences necessitates the use of metrics tailored to the specific application domain. As such, a set of targeted metrics were employed, focusing on the subdomain of exploratory behaviours [356]. The tailored metrics comprise subjective evaluation on the level of interaction, level of direct perception and level of usage situation, as detailed in the work of Brunnström et al. (2013) [42, 308, 323]. The assessment of the level of interaction was based on attributes such as engagement and attention [230]. Furthermore, the level of direct perception was evaluated based on the attributes of spatial awareness, i.e., the perceptual information created during content consumption. Lastly, subjective evaluation of the usage situation was conducted with regards to the usability context (i.e., seating) [90].

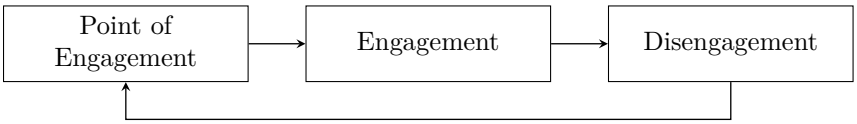


Figure 9: Model of Engagement

The time-consuming nature of the eye-tracking experiment necessitated a time-efficient evaluation process, as it was important that users maintained engagement during the experiment without the risk of boredom or fatigue. Therefore, it was decided to implement a set of statements throughout the questionnaire to acquire subjective measurements, which helped to efficiently gather insight on the user's perception and experience in a relative short amount of time, similar to Norouzi et al. (2021) [223]. The users were able to rate their level of agreement on each of the statements utilising a 7-point Likert scale, similar to efficient subjective assessment by Wu et al. (2009) [226, 356]. The UEQ includes a set of questions and statements regarding the perception and interpretation of the viewed content during the eye-tracking session, and determines – on the level of perception – how attributes of engagement, attention, spatial awareness, fear of missed content and usability context contribute to the behavioural impact of spatiotemporal complexity and seating type.

Pre-Test Parameter Accommodations During the semi-structured interview, users of the pre-test parameter study were evaluated on the presented questions, statements and duration of participating in the evaluation. This led to the exclusion and modifications of some questions and statements to accommodate a more optimal experiment duration, and to avoid unnecessary ambiguous questioning. As a part of the pre-test parameter study, the UEQ questions pertaining engagement, attention and spatial awareness were measured separately for each 360-degree video sequence. However, it proved difficult to obtain the nuanced differences of these attributes between the videos, as users showed difficulty recalling those measures for each video specifically. Furthermore, repetitive questions about each of the 6 presented videos significantly increased the duration of the evaluation. Consequently, the UEQ was translated to a more general approach, and the video-specific evaluation was only done during the semi-structured interview.

2.3.1 Engagement and Attention

As defined by O'Brien et al. (2008), engagement is a part of the user experience and thus poses an important considerable aspect of the interaction [230]. Viewing behaviour (i.e., gaze distribution) is the result of the process of engagement, where high levels of user engagement incite more exploratory behaviour. Hence, when analysing the user's perception of the 360-degree video content, it is essential to consider the engagement process [205, 230], as illustrated in Figure 9. The point of engagement is initiated by user interest, accompanied by attributes of motivation, novelty and aesthetic appeal of the system. Given that this study pertains users evaluating distinct 360-degree videos, attributes of aesthetics (i.e., QoE) and motivation (i.e., experiential goals) were not considered in the subjective analysis. Furthermore, since relative experience with VR or 360-degree video was a part of the inclusion criteria for participation, novelty of the system was also excluded from the analysis. As a result, initiation of engagement in this experiment design was sustained by user interest, sensory appeal and aesthetics, until disengagement occurs by interruption. With the latter two being evaluated through QoE assessment, the user's level of engagement was predominantly based on the attributes of interest.

Due to the variety in presented 360-degree video content, the level of user interest in the specific video was considered a significant influential factor on the level of engagement the users could experience during that specific video. As implied by the same process of engagement, content that fails to interest a user is a considerable barrier to engagement [148, 230]. This is further emphasised by the work of Dobrian et al. (2011) and Davis et al. (1989), demonstrating

the correlation between level of interest, tolerance and the impact of user preference on user assessment [76, 81]. While some users may have favoured one genre over the other, others might have been indifferent to the content. As a result, users were instructed to evaluate how interesting they found the entire selection of 360-degree videos during the post-test evaluation. Determining to what extent the presented content accomplishes user interest, to establish engagement, was done by asking the users to rate each video on a 7-point Likert scale. The users were asked to rate their level of interest in each of the videos to determine the relationship between gaze behaviour and how interested they were in the content.

Moreover, as presented in McCarthy et al. (2004), the experiential threads of engagement denotes various attributes of engagement and encompasses the sensual, emotional and spatiotemporal threads of experience [205, 230]. As such, user interest is considered one of many attributes of engagement. It defines the emotional thread of experience, accompanied by the level of enjoyment. The sensual thread of experience is measured through perceived richness of the graphics, on which the spatiotemporal complexity of the 360-degree video sequence significantly relies. Lastly, the spatiotemporal thread of experience denotes the amount of feeling situated in the story, fast perception of flow of time, sense of being in control and lack of being aware of other people. As such, to determine the overall level of user engagement, the attributes from each of the experiential threads were taken into account. A set of questions was devised to measure each of the attributes from the three before-mentioned threads of experience. The overall level of engagement was acquired by averaging the 7-point Likert scores from each of the associated questions, resulting in a singular score. Similarly, acquiring a subjective insight into the user's level of attentional focus during the viewing session was done using 7-point Likert scale questions and statements. In particular, the questions related to attentional focus were aimed at the user's sense of feeling drawn to specific AOIs and maintaining attentional focus.

2.3.2 Spatial Awareness and Usability Context

Another measure, to understand the relationship between gaze behaviour and cognitive processing thereof, was the amount of spatial awareness each user pertains. The study of spatial awareness and usability context is critical to gaining a deeper understanding of the relationship between gaze behaviour and cognitive processing while watching 360-degree video. Spatial awareness, defined as the ability to perceive and understand the spatial dimensions of one's environment, plays a crucial role in navigation and orientation within the virtual environment. Furthermore, the user's ability to identify spatial dimensions is a significant influential factor in generating a sense of presence and immersion, as implied by Zou et al. (2018) [376]. Spatial awareness was assessed using the user's ability to locate ROIs and understanding of the virtual layout. Similar to the level of engagement, the scores were averaged across the 7-point Likert scales to generate the user's overall level of spatial awareness of the virtual environment. The study of spatial awareness enabled a comparative analysis of the amount of visual attention required for user's to perceive and comprehend the spatial layout of the virtual environment. Additionally, as suggested by Rothe et al. (2019), the areas of the spherical projection present outside the user's FOV results in content being left out of viewer perception, leading to a sense of fear of missed content [262]. The user's awareness of this phenomenon strongly stresses both perceptual and cognitive load. Consequently, the users were asked about the experienced sense of FOMC (fear of missed content) [223].

Furthermore, the influence of usability context – seating type – denotes an important level of interaction, as identified in the work of Brunnström et al. (2013) [42, 90]. Therefore, to gather additional insights on the confounding influence of various usability factors on gaze behaviour, the users were asked about the level of comfort, overall usage enjoyment and limiting influence of the utilised chair type. In addition, the difference in seating between groups R and F, as well as the unrestricted exploration of the virtual environment, could result in different interpretations of the presented content as implied by Löwe et al. (2015) [194]. Therefore, the users were asked to provide a brief description of each of the videos, and were encouraged to recall distinct elements from each of the 360-degree videos.

2.3.3 Perception of Viewing Behaviour

The second part of the evaluation entailed the semi-structured interview (SSI), designed to delve deeper into thoughts and active gaze behaviour of the user, similar to O'Brien et al. (2008) [230]. The semi-structured interview was used to identify trends in the user's perception of their own gaze behaviour, adding an extra dimension to the subjective analysis. As such, this second part of the evaluation contained a series of open-ended questions aimed at understanding the relationship between the user's self-reported perception of conscious gaze behaviour and acquired gaze data. While the UEQ focused on the general viewing experience, the semi-structured interview contained video-specific questions about the user's perception of gaze behaviour during each 360-degree video. Furthermore, during the semi-structured interview, the users were asked about how they perceived any behavioural changes across the various presented content. In particular, users were asked to describe and reflect on their gaze behaviour related to changes in camera motion and different video genre content, as well as motivate their answers. Moreover, the enhanced perceptual load, as occurs while interacting with 360-degree content, could result in unconscious behavioural responses and as such, could go unnoticed by the user [196, 223, 262]. The implementation of a semi-structured interview enabled insight into key trends of how the user perceived their own viewing behaviour and examine a potential dichotomy between objective and subjective gaze behaviour. The questions were oriented at differences in genre and camera motion as representative variations in spatial- and temporal complexities, ensuring comprehension among users. Depending on the given answers, follow-up questions were provided to elicit further insights.

The user evaluation utilises the discussed literature and theoretical foundation in Chapter 1 to expand upon the complex variety of theoretical concepts. As such, the user evaluation comprises various concepts and theoretical frameworks, found within interrelated research domains. Table 4 presents an overview of the related concepts and theoretical associated with each question and statement provided during the user evaluation. The SSI protocol is presented in Appendix F21.

2.4 Material and Apparatus

The apparatus utilised during the study and the experiment setting is described in the following section. The split-part experiment design required different material and apparatus for each of the two parts. Firstly, the eye-tracking study was conducted using specific hardware and software to accommodate the main objective of the research, which is described as follows.

The VR environment was developed in HTC Viveport and displayed through the HTC Vive Pro Eye HMD. The HTC Vive Pro Eye HMD features two 3.5-inch dual-AMOLED displays, producing a cumulative resolution of 2880x1600 pixels, equating to 1440x1600 pixels per eye. The HMD utilises a PenTile Diamond subpixel layout, implementing two subpixels per pixel, which produces a sharp image due to the pixel density of 615PPI. Each display operates at a sample rate of 90HZ. The advanced tracking capabilities of the HTC Vive Pro Eye supports 6 DOF marker-based tracking, facilitating accurate motion tracking in the virtual environment. Moreover, the HMD utilises Tobii eye-tracking technology, which outputs gaze data at 120Hz with an accuracy and precision between 0.5º - 1.1º. Using a 5-point calibration system, both eyes are tracked. The HMD renders a 107.06º horizontal FOV, a 107.71º vertical FOV and a 110.48º diagonal FOV. The visible FOV accumulates to 98º for both horizontal and vertical dimensions. The device features ergonomically adjustable elements, including adjustments for eye-lens distance, interpupillary distance (IPD) headphones and a comfortable strap. Lastly, various sensors such as SteamVR tracking, G-sensor, gyroscope, proximity and and IPD sensor were built-in.

Secondly, the user evaluation required different material to acquire the subjective evaluation data from each user. After conducting the eye-tracking process, the user evaluation protocol was followed. This consisted of the UEQ (see Appendix F20) to provide subjective measurements on the perceived content. Subsequently, a semi-structured interview was conducted between the researcher and user, based on the questions presented in Appendix F21. The UEQ measurements and semi-structured interview data were digitally acquired by the researcher using a laptop, and were developed using Qualtrics software.

Research Focus	Sub-Focus	Questions, Statements
Engagement	Enjoyment experience	How enjoyable did you find the overall 360-degree video experience?
	Graphic quality	How would you rate the richness and quality of the graphics in the 360-degree videos?
	Passage of time	How quickly did time seem to pass while watching the 360-degree videos?
	Sense of control	How much control did you feel you had over your viewing experience while watching the 360-degree videos?
	User Interest	How interesting did you find the 360-degree videos?
Plausibility Illusion	Narrative immersion	To what extent did you feel situated in the story being depicted in the 360-degree videos?
Place Illusion	Awareness of others	To what extent were you unaware of the presence of others while watching the 360-degree videos?
Attention	Attentional focus	To what extent were you able to maintain your attention on the 360-degree video throughout the entire viewing experience?
	Influence of temporal change on attention	"I found myself getting distracted by the background elements when watching the videos with a static camera." "I found myself more focused on the details of the scene when the camera was moving slowly."
Spatial awareness	AOI identification	How well were you able to locate and identify important objects or landmarks within the virtual environment?
	Understanding virtual environment layout	How well were you able to understand the layout of the virtual environment?
	Virtual navigation	How well were you able to navigate through the virtual environment?
	Influence of camera motion on spatial awareness	"I had a better understanding of the layout of the environment when watching the 360-degree content with a static camera." "I found it difficult to orient myself and understand the layout of the environment when watching the 360-degree video with a moving camera."
Usability context	User experience of seating type	How comfortable was the use of the [chair type] during the viewing session? To what extent did the [chair type] affect your overall enjoyment of the 360-degree video?
	Usability of fixed chair	"I felt limited in the amount of exploring I could do due to the fixed chair." "I found it harder to keep track of the camera movements because of the fixed chair."
	Usability of rotating chair	"I felt more encouraged to look around because of the rotating chair." "The rotating chair made it easier for me to follow the camera movements."
Perception of content	Interpretation	Can you provide a brief description of each of the videos you watched in this study? Please provide elements or details you found interesting or remember vividly from each of the videos.
	Fear of Missed Content	To what extent did you experience the sense of FOMC, due to loss of information or missed out content? If so, can you describe when and why?
Perception of gaze behaviour	Awareness of temporal influence	Some of the videos were more dynamic, with the camera moving relatively fast or having more movement. Other videos were more static, with the camera moving relatively slow or remaining stationary. Can you describe the effect this had on how you viewed the content?
	Awareness of spatial influence	How would you describe your viewing behaviour between the different genres of videos (scenery, roller coaster, video game)? How was it different and why do you think that was the case?

Table 4: User evaluation research focus and sub-focus areas.

A specific computer setup was used to accommodate the HMD hardware and required software. The computer used met the minimum CPU and GPU requirements, with an Intel Core i5-4590 or AMD FX 8350 processor, and an NVIDIA GeForce GTX 970 or AMD Radeon R9 290 graphics card or better. The HMD setup consisted of the before-mentioned HTC Vive Pro Eye Headset, two VIVE base stations, two VIVE controllers, as well as the required cables and power adapters. To record and analyse eye-tracking data, iMotions eye-tracking software [140], including the VR eye-tracking module, was implemented.

Setting The study was conducted in the eye-tracking laboratory of the Utrecht University, the Netherlands. In order to minimise potential sources of interference and distractions, the study was performed in the dedicated laboratory area with only the user and researcher present. Users were seated behind the computer, and were placed on an ergonomic chair, depending on the assigned group. The chair was also adjustable in height and amount of recline, to accommodate the best and most comfortable seating position for each user. As described in § 2.2, and with regards to the significant influence of context influence on viewing behaviour [90], it was decided to maintain contextual continuity across all users to ensure measurements of most natural occurring viewing behaviour. Furthermore, the researcher was seated close to the computer in order to manually control the input as well as closely monitor the experiment. There was sufficient space between the researcher and the user to allow full 360-degree rotations on the dedicated chair type. Other users, in case present, were seated on the other side of the monitor to ensure they didn't have access to the content prior to their experiment session.

Compensation To minimise the risk of users providing false information about their physical well-being in return for any monetary compensation, it was decided to rely on intrinsic motivation of the user and not provide compensation of any monetary value. Refreshments were offered throughout the experiment process, and additional parking expenses were compensated for. Mutual participation in another research, as a quid pro quo for their participation, was offered.

2.4.1 Implementation

Prior to the data collection, the required software components were set up. Implementation required the installation of the following software packages: `Steam VR 1.15.10` [65] and `Vive_SRanipal SDK 1.3.6.8` [135] in conjunction with `iMotions 8.2.2` [140] and the LTS version of `Unity 2020.3` [316]. To enable the communication between the eye-tracking drivers of the HTC Vive Pro Eye and iMotions software, `VIVE eye-tracking SDK` and `SR Runtime` were employed, as well as the VIVE and SteamVR software. The required VIVE base stations, which use infrared signalling to track the HTC Vive Pro Eye HMD and controllers, were set up diagonally across at a distance of 4 meters. Installed at an height of 2 meters and at an angle of 35 degrees, the base stations individually provide an FOV of 120 degrees. Subsequently, the IP address and TCP ports from `SRanipal SR Runtime` were entered in the host address field of the iMotions sensor and API settings, enabling the connection between iMotions and the eye-tracking sensors. The capture rate was increased to 30 fps to achieve higher temporal resolution for screen recording, as well as match the frame rate of the content. The 360-degree video content was converted and to an equirectangular monoscopic 2D format and uploaded, as supported by the iMotions software. It is important to note that in iMotions, only certain image formats are supported, including Mono, Top/Bottom, Side by Side, and Cube EAC. Loading the reference image as a Mono format was recommended to ensure optimal gaze mapping accuracy. Consequently, and as before-mentioned, the 360-degree video content in EAC format obtained from the YouTube database was initially converted into the ERP format. This conversion process is described in § 2.2.2. The reference image, required for data visualisation, was created from the original monoscopic format 360 stim image instead of FRAME. The gaze mapping procedure mapped all of the gaze data into this new reference image, enabling the creation of a single heatmap from the 360 stim image. Since heatmaps were generated at the aggregate level within the iMotions data visualisation software, individual heatmap visualisation required separate gaze data tracking for each of the 360-degree videos across all users. Therefore, a total of 6 eye-tracking runs were performed per user. Lastly, the system was calibrated at the start of each experimental run to increase reliability and precision in the gaze tracking data.

2.5 Population

The following section provides an overview of the population involved in this study. This section includes a detailed description of the criteria for eligibility, the descriptive statistics of the demographic information, the sample size, the sampling and recruitment methods used, as well as the adhered to information provision and consent procedures.

2.5.1 Criteria

The following participation criteria was used to recruit users and provide precise selection:

- Be over 18 years old
- Have normal or corrected-to-normal vision (incl. colour-blindness)
- Have no history of motion sickness or epilepsy
- Have no physical conditions that may limit or be aggravated by using a VR headset
- Haven't participated in similar research in the past.

Since the eye-tracking study aims to elicit as much natural viewing behaviour as possible, it was decided to recruit users that had no previous experience with this type of research. This reduced risk of users behaving differently because they are aware of what is expected of them. This is further emphasised by the notion that the users would know they are being observed and as such, a small chance exists that they would behave differently because of that. The inclusion of these criteria aims to reduce this Hawthorne effect as much as possible [279]. Furthermore, it was important that users were somewhat experienced with VR, to reduce the novelty effect on gaze behaviour [281]. The results from Serrano et al. (2017) further underpins this criterion, as occasional or rare use of VR is unlikely to impact results. People who had undergone eye surgery or had eye diseases, wear heavy makeup, or had high myopia were excluded from participating in the study due to the potential effects on eye-tracking performance.

2.5.2 Participants

The participants in this research study comprised a diverse demographic set, as the sample was primarily characterised by their experience with VR environment and usage, and physical ability to participate in the experiment. The selection was done during the recruitment using the established criteria, see Appendix D14. As such, the sample is representative of a subset of a large population of users whom are interested in using VR technology and not physically impaired to use the technology.

A total of $n = 52$ users conducted the proposed experiment, of which $n = 33$ males and $n = 19$ females. While an older age was not necessarily considered an exclusion criterion, it was considered heavily correlated with lower levels of experience with VR usage. Consequently, all users were of age between 18 to 29 years ($\mu = 22.5, \sigma = 2.57$).

The selective procedure recruited users whom all had prior experience with VR. As such, the VR experience frequency was distributed as follows: $n = 9$ users uses VR once a year or less, $n = 21$ a few times a year, $n = 17$ once a month or more and $n = 5$ once a week or more. All users were extensively informed about the research and provided written consent, see Appendix E16 and E17, adhering to the ethical obligations of research participation.

Notably, the sample size ($n = 52$) adheres to the minimal sample size for studies containing video quality assessment, suggested by Konuk et al. (2013) [172]. It is important to highlight that throughout this work, the terms users and participants are used interchangeably, as the term 'users' refer to the users of the experimental setup of this research.

2.5.3 Sample Size

The sample was further divided into two equal groups R ($n = 26$) and F ($n = 26$), enabling the between-subjects design for studying usability context. Group R contained $n = 19$ males and

$n = 7$ females. The age ranged from 18 to 28 ($\mu = 22.9, \sigma = 2.68$). Group F included $n = 14$ males and $n = 12$ females, with their respective ages varying from 18 to 29 years ($\mu = 22.1, \sigma = 2.43$).

2.5.4 Sampling and Recruitment

For this research, a purposive sampling method was used to recruit users. The purposive sampling method was chosen to ensure that users who met the eligibility criteria and were available and willing to participate were selected for the study. Recruitment was predominantly done among students and other colleagues of the Utrecht University, as well as interested individuals not affiliated with the Utrecht University. The users were approached via digital platforms, such as WhatsApp, iMessage, email and social media, to provide them with sufficient time to consider their participation. The recruitment communication included a brief description of the study and its purpose, along with an invitation to participate in the research. The complete invitation, as part of the the sampling correspondence, can be found in Appendix D.

The link embedded in the recruitment text lead an information web-page, containing all inclusion criteria and necessary information regarding the research. Interested individuals could voluntarily provide their email address, enabling correspondence on further details of the study, such as scheduling of the experimental sessions. The web-page was developed using Qualtrics software. Consideration time was included to ensure that potential users had enough time to consider their participation and make an informed decision. The selected users were then contacted to confirm their participation in the study.

Initial recruitment was done 4 weeks prior to the scheduled experimental runs. The study was conducted over a total time period of 27 days with continuous recruitment during this time period, to ensure a sufficiently large sample size was obtained to accommodate for the between-groups design.

2.5.5 Information and Consent

To ensure transparency and ethical conduct of the research, informing participants and obtaining their informed consent was essential. Participants were introduced to the study during recruitment and were provided extensive elaboration on the research at the start of the study. To avoid any confusion or misunderstandings, participants were informed extensively on the implications of their participation in the research. It was also emphasised that the study did not involve any concealment or deliberate misleading of users in any way. While users were informed on the fact that the experiment required watching a variety of 360-degree video content, the specific variation of which (i.e., spatiotemporal complexities) was not stated beforehand, reducing risk of the Hawthorne effect [279]. Further provision of detailed information about the nature of the study, the specific procedures involved and the potential risks of participation were provided using a digital information sheet, which can be found in Appendix E16.

Explicit consent was obtained from all users before participation in the research. The consent form, as can be found in Appendix E17, presents information about the nature of the study, how their data will be collected, stored, and analysed, including sensor recordings such as eye-tracking data. Sufficient time was provided for any questions or concerns before providing consent.

A collection of all information sheets, the consent form used to inform users of the research objective, participation guidelines, requirements, risks, and terms and conditions can be found in Appendix E. This includes a modified information sheet used specifically for the pre-test parameter study (see Appendix E15). The information sheets and consent form were developed and distributed using Qualtrics software.

The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences was conducted to assess potential ethical and privacy concerns related to this research, as presented in Appendix C. All provided information and ethical considerations were processed accordingly. Whilst the Quick Scan identified issues, this project was allowed to proceed after additional human assessment (see approval email in Appendix C12)

2.6 Procedure

The experiment design, as before-mentioned in § 2.1, encompasses multiple approaches and methodologies. The inclusion of the M-ACR quality assessment method, eye-tracking study and user-centric evaluation resulted in an extensive and complex experiment process, which is visualised in Figure 10. The flowchart utilises geometry to represent processes, input, output and points-of-decision. The rounded rectangular nodes represent the experiment's start- and end-points. The transparent parallelogram nodes define input points in the process, while the greyscaled parallelogram nodes indicate output points. The involved processes are visualised as rectangular nodes. Furthermore, decisions are represented as a diamond node, containing a prerequisite. Depending on whether the prerequisite is met during the process, the flow will continue in a predefined direction. Lastly, the arrows indicate the direction and flow of the processes. This subsection further elaborates on the experiment process as a whole and the flow of various sub-processes within, while further elaborating the different stages of the entire procedure. The entire experiment had an approximate duration of 30 to 40 minutes per user.

2.6.1 Preliminaries

Prior to conducting the experiment, the recruitment process facilitated the sampling of the participants. The recruitment of participants and sampling method are described in "Sampling and Recruitment" (see § 2.5). The experiment could be conducted one participant at a time, and therefore necessitated efficient scheduling. As such, during recruitment, participants interested in conducting the experiment were asked to provide a time-slot preference. On the scheduled experiment time-slot, participants were welcomed to the University and the experiment would start. The set-up and preparation of the required hard- and software was done beforehand. This includes the HTC Vive Pro Eye and iMotions eye-tracking software, as well as other apparatus described in § 2.4.

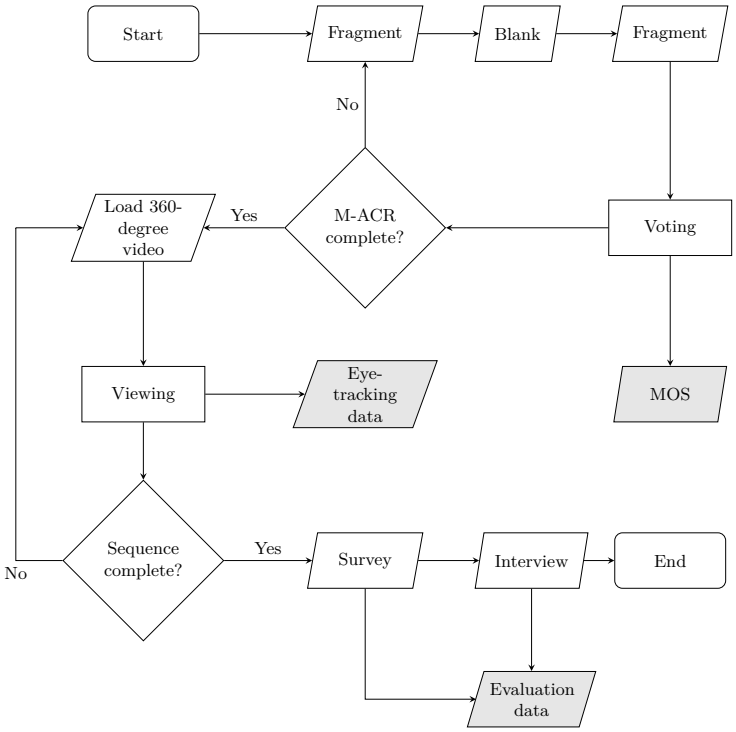


Figure 10: Flowchart of the experiment process.

Start The start of the experiment contains the introduction of the experiment, gathering of demographic information, asking of consent and provision of instructions. During the introduction, the main aim of the study was further elaborated, and the head researcher was introduced. After the introduction, participants were informed on the terms and conditions of their participation, as described in "Information and Consent" (see § 2.5). The introduction of the study and all essential information was presented using a digital information sheet, provided to the participant. The information sheet can be found in Appendix E16. Furthermore, participants were asked to provide consent using a digital consent form prior to the experiment. The digital consent form can be found in Appendix E17. Enough time was provided for each participant to read and understand the scope of the research, related procedures and potential risks. Lastly, measurements, consent and demographic information were acquired through a short questionnaire, which can be found in Appendix F18. The start and introduction of the study took approximately 5 minutes.

Eye-Tracking After the introduction of the experiment, participants proceeded with the eye-tracking section of the experiment. The researcher carefully placed the HMD on the participants head, which was ergonomically adjusted including focus lens adjustment. As proposed in the framework by Carter et al. [49], calibration was performed once at the start of each eye-tracking run to maintain consistent quality of the fixation data. The first part of the eye-tracking process included the subjective quality assessment of all stimuli, through the M-ACR methodology. The participant was presented a 10 second fragment of each 360-degree video that they would be watching in its entirety later in the experiment. After the 10 second fragment, a short blank intermission of 8 seconds was presented. After this brief intermission, the participant watched the same fragment as before again, after which the voting round started. The participant kept the HMD on throughout, and the researcher asked the participant to give the presented fragment a rating as described in § 2.2.1. The output data from the voting process generated the MOS of each participant for each of the 360-degree videos. The M-ACR assessment process was repeated until all six 360-degree videos were rated by the participant.

After completing the M-ACR assessment process, the participant began watching the entire sequence of 360-degree video content. The videos had the same maximum duration and the sequence order was randomised across all participants, as described in § 2.2.5. The researcher prepared the software for data generation and carefully instructed the participant to use the HMD. After loading the predetermined video content, the participant watched the presented content. This process generated eye-tracking data, predominantly gaze origin, gaze direction and timestamps. The participant continued watching the video until requested to stop watching the content, or when the video had ended. The subsequent video, as predetermined in the sequencing for each participant (see § 2.2.5), was then loaded as long as the sequence had not yet been completed. The software was continuously monitored by the researcher to ensure all generated data was safely stored accordingly. When all videos from the sequence had been seen by the participant, the eye-tracking section of the experiment was concluded and all installed hardware (i.e., the HMD) was removed from the participant. Before proceeding to the user evaluation, a short intermission was provided to allow participants to adjust from the exposure of the 360-degree video and the virtual reality environment. The eye-tracking study took approximately 15 minutes, including the brief intermission.

User Evaluation Once the brief intermission was over, and the participant was ready to proceed, the user evaluation was performed utilising the user evaluation questionnaire and semi-structured interview. Firstly, the participant was asked to fill out the digital user evaluation questionnaire as presented in Appendix F20. The UEQ contained several statements and questions relating to the perception of presented content, the details of which are presented in § 2.3. Secondly, the semi-structured interview was conducted with the participant, enabling further insight into the participant's perception and behaviour. The answers provided by the participant were digitally written down by the researcher. The SSI protocol can be found Appendix F21. The output data generated from the user evaluation questionnaire and semi-structured interview

was registered as evaluation data (i.e., self-reported measures). The post-test evaluation took approximately 15 minutes.

End As part of the conclusion of the study, a short debriefing period was implemented. During the debriefing, participants were able to ask additional questions regarding the study, their use of data or any other related matter. Although the experiment was carefully designed to minimise risk of cybersickness, physical strain or any form of discomfort, the experiment was still not entirely risk-free. As such, if necessary, participants were also provided additional information regarding post-experiment care. Subsequently, participants were thanked for their participation and were compensated through refreshments provided throughout the experiment and in some cases, mutual participation in another research as a quid pro quo. Furthermore, contact information of the researcher was provided to the participant in case any additional questions or concerns would arise in the future. Lastly, the acquired data was safely stored, transferred and processed in preparation of the data analysis, finalising the experiment. The conclusion of the experiment took approximately 5 minutes.

2.7 Data Analysis

The quantitative and qualitative data generated from the study was analysed using a systematic approach. In assistance of addressing the main research objective and answering the related sub-questions, as introduced in § 1.7, tailored approaches were implemented to analyse the subsequent data and derive significant results. This subsection details the identification of present variables, both independent and dependent, in § 2.7.1. The data analysis approaches required pre-processing of the acquired data, which is discussed in § 2.7.2. Furthermore, the analytical framework, statistical approaches and implemented Python libraries are highlighted in § 2.7.3 and § 2.7.4, respectively. Lastly, the various data visualisation techniques are detailed in § 2.7.4.

2.7.1 Variables

The eye-tracking study and user evaluation were designed to generate physiological, objective and subjective data to approach the main research objective of this thesis. As such, the various variables were used to analyse relationships between these variables to determine how varying spatiotemporal image complexity in 360-degree video influences gaze behaviour in VR, while factoring in the dynamics of cognition, perception and usability context. This subsection identifies the included variables and discusses the respective roles within the thesis. A total of four independent variables were used in the research study, which were utilised to measure their effect on the dependent variables. A total of 15 dependent variables were measured and identified. An overview of the variables used in this research is presented in Table 5. Aside from the statement ratings, the Likert scale data was treated as continuous despite the inherent ordinal nature, enabling the employment of parametric tests. Due to the increased sensitivity, the analyses provide a more nuanced interpretation of results. The treatment of Likert scale data as continuous is further supported by the Central Limit Theorem, which states that the sampling distribution of the mean reaches normality in sufficiently large sample sizes [127, 176]. The demographic information acquired as part of the demographic questionnaire consists of user-related demographic information. Each of these measures were considered complementary measures in the analysis of this thesis. The data of perceptual attributes, usability factors and statement ratings was acquired using the Qualtrics software.

Spatiotemporal Complexity The spatial- and temporal complexities from each 360-degree video sequence were the primary independent variables. The respective SI- and TI-values varied, identified by the spatiotemporal matrix in Figure 4. As such, the SI- and TI-values were used to assess how variations in these values influence the amount of gaze distribution expressed by each user. The computation of SI- and TI-values is detailed in § 2.2.4.

	Variable	Type
Independent	SI	Continuous
	TI	Continuous
	SeatingType	Categorical
	δ	Continuous
Dependent	N_ψ	Continuous
	Perceptual Attributes	Continuous
	Usability Factors	Continuous
	Statement Rating	Ordinal

Table 5: Overview of the independent and dependent variables.

Seating Type The sample size was split into two groups, which was done to determine the influence of usability context on the amount of gaze behaviour. Each group of user conducted the experiment on an assigned chair type. As such, the independent categorical variable SeatingType contained two levels: fixed-position chair and rotating chair.

Diegetic Artefact Score δ Each 360-degree video contains visual artefacts that can be of influence to the amount of exploration done while interacting with the video. Therefore, visual elements, such as diegetic artefacts and attentional guiding mechanisms, pose influential in the guidance of attention and eliciting of exploratory behaviour. To control for any confounding cinematographic influence, a coding scheme was devised to determine the cinematographic influence present within each of the utilised 360-degree videos. This is further detailed in Chapter 4, which elaborates on the use of coding schemes to determine the presence of diegetic artefacts that guide user attention across the 360-degree videos. Ultimately, points were assigned on the size and duration of such artefacts, resulting in a Diegetic Artefact Score δ. The δ-value denotes the degree to which attention guiding visual artefacts are present in the 360-degree video and as such, pose as a confounding factor in eliciting exploratory behaviour. This preliminary diegetic assessment, presented in Chapter 4, also examines the complex relationship between the degree of gaze exploration and total δ point-values by utilising logarithmic and polynomial regression models as well as reciprocate ratio-values. The independent continuous variable of δ was included as a confounding variable in the statistical analyses, due to the complex association of diegetic artefacts and degree of gaze exploration.

Quadrifactorial Exploration Index N_ψ The normalised quadrifactorial exploration index N_ψ was the primary dependent variable employed in the data analysis, and was specifically constructed and formulated for the scope of this thesis. The index expresses the measure of exploratory gaze as a continuous numerical value in range [0, 1]. By utilising gaze distribution heatmap signals, the exploration index combines area coverage ratio, average pixel intensity, structural dissimilarity and entropy to quantify gaze distribution. The degree of gaze distribution N_ψ was computed for each individual heatmap across all users. For each user, a total of six heatmaps were extracted from the advanced eye-tracking software, corresponding to the six presented 360-degree videos during the study. The resulting dataset containing all computed N_ψ-values is presented in Table A.1, found in Appendix A1. The mathematical construct, formulation and computation of the quadrifactorial exploration index is detailed in Chapter 3.

Perceptual Attributes The perceptual attributes pertain influential aspects of the user's interaction and perception that were established in the literature study. The variables denote attributes and aspects of the user's perception of 360-degree video interaction. As such, the perceptual attributes to models of cognitive processing that elicit behavioural responses not accounted for by the objective measurements of the eye-tracking study, which did not contain parameters directly related to the temporal effects of human perception and judgement [9, 305, 356]. As such, the

subjective evaluation assessed aspects of cognitive perceptions and subsequent behavioural consequences, resulting in the set of perceptual attributes. The set of perceptual attributes include the levels of QoE, engagement, attentional focus, spatial awareness and fear of missed content (see § 2.3). Each of the perceptual attributes is a continuous dependent variable, as measured on a 7-point Likert scale. Some perceptual attributes, such as engagement and attentional focus, entail multiple aspects and were composed of a multitude of measures. For these attributes, a single representative score was derived based on the average Likert-scores for each of their associated questions from the UEQ and SSI, as detailed in 2.7.2.

Usability Factors The assessment of usability factors was done by measuring the level of comfort and impact on personal enjoyment of each user regarding the usage of a rotating or fixed-position chair. These quantitative metrics were used to provide additional insights on the level comfort and enjoyment of the seating type. These continuous dependent variables were measured on a 7-point Likert scale.

Statement Rating Lastly, as included in the UEQ and SSI, the users were presented a variety of statements regarding their viewing behaviour. The level of agreement on each statement was measured on a 7-point Likert scale across all users. These quantitative metrics were used to acquire additional insights, used to enhance the qualitative findings. These continuous dependent variables comprised a set of 7 measures, containing the degree of agreement for each statement.

2.7.2 Data Pre-Processing

A series of pre-processing steps were taken to prepare the raw data for further processing and analysis. The data and sensor recordings from this research study were acquired utilising the iMotions software export module, which included general information, metadata, gaze data and timestamps. The gaze distribution heatmaps were extracted using software analysis module, utilising the sensor recordings. The generated gaze distribution heatmap image signals were used in the computation of the quadrifactorial exploration index N_ψ. All analyses on image signals were performed utilising equirectangular projection [16, 54]. A total set of 330 heatmap image signals were generated. This includes each individual heatmap image for each video across all users ($n = 6 \times 52 = 312$), a subset of heatmaps using aggregate gaze data from both groups ($n = 12$), and all users combined ($n = 6$).

The pre-processing steps taken in the computation of N_ψ are detailed in § 3.1, however, briefly introduced in this subsection. For the coverage area ratio and normalised average pixel intensities, the gaze distribution heatmap image signal was superimposed on the white image frame to ensure consistent computations of N_ψ, after which the heatmap image signals were converted to greyscale. Subsequently, a binary mask was created to identify pixels with varying intensity, denoting the presence of the gaze distribution heatmap. This same approach of superimposing was applied to the computation of structural dissimilarity and entropy. Additionally, the probabilities for each intensity level were computed through the created histogram of each greyscale heatmap image signal. The white image frame, on which each gaze distribution heatmap image signal was superimposed, was processed as the reference image for structural dissimilarity. This process was employed on a total of $n = 330$ (312 unique and 18 aggregate) gaze distribution heatmap image signals. The process of computing the degree of gaze exploration N_ψ for all $n = 52$ users across the 360-degree video set is detailed in Chapter 3.

The long-format dataset was structured using identifiers for each of the users. The identifier consisted of the groupID (i.e., relating to seating type) and subsequent participant number, in order. As such, user 1 of group R was identified as R1. Similarly, user 15 of group F was identified as F15. As part of the data pre-processing, the resulting N_ψ-values were included in order $A1, A2, B1, B2, C1, C2$ in the resulting dataset (see Appendix A1) to avoid sequencing errors caused by the Latin square design. This enabled a consistent overview of the data in each column. All data was pseudonymised to ensure privacy and data security. The selection process ensured

that each user adhered to the inclusion and exclusion criteria as detailed in § 2.5. Furthermore, the dataset was checked for missed data, which was not found. Consequently, no data entries were excluded from the dataset.

Categorical Variable Encoding The categorical variables used in this research study were pre-processed using encoding strategies. As identified in § 2.7.1, SeatingType as a categorical variable was included. One-hot encoding was used to assign a new binary variable to each category of the SeatingType. Consequently, the newly created binary value denoted the specific group in which the user was assigned. Therefore, SeatingType of value 1 denoted the rotating chair, and 0 denoted the fixed chair. Consequently, the encoding strategy was implemented to prepare the dataset for the subsequent statistical analyses.

As before-mentioned, the perceptual attributes of engagement and spatial awareness were measured on a multitude of 7-point Likert scales during the UEQ. To derive a singular score for each of these attributes, the scores were averaged across the associated questions. The level of engagement was a composite score, measured over seven different questions. Similarly, spatial awareness was measured over three. Due to the 7-point Likert scales used to assess the perceptual attributes, a normalisation was performed to ensure the MOS scores derived from the QoE 5-point Likert scales could be compared on the same scale as the rest of the perceptual attributes. Using Min-Max normalisation, the range of Mean Opinion Scores (MOS) was standardised to a 7-point scale, aligning the QoE-value range with the other perceptual attributes.

$$X_{\text{new}} = \frac{(X_{\text{old}} - X_{\text{min\_old}})}{(X_{\text{max\_old}} - X_{\text{min\_old}})} \times (X_{\text{max\_new}} - X_{\text{min\_new}}) + X_{\text{min\_new}} \tag{21}$$

where $X = MOS$, X_{old} is the QoE value on scale 1-5, and X_{new} is the QoE value on scale 1-7. As such, X_{min} and X_{max} denote the minimum and maximum values on the old and new scales. Consequently, a one-unit change in score is consistent across the perceptual attributes.

Lastly, as detailed in Chapter 4, pre-processing for δ was done utilising the coding scheme of the diegetic assessment.

2.7.3 Analytical Framework

The research study design comprised two parts, as detailed in § 2.1. The eye-tracking study and user evaluation generated both quantitative and qualitative data, aimed to elucidate the main research objective and answer the related sub-questions introduced in § 1.7. As such, a series of analytical approaches were implemented to help understand the relationships and effect of each of the influential factors on the user's gaze behaviour during 360-degree videos in VR. The subsequent subsections discuss the employed statistical and analytical approaches for both quantitative and qualitative analyses. Table summarises the research objectives and related analytical approaches. The results from the quantitative and qualitative analyses are presented in Chapter 5.

Quantitative Approach

The eye-tracking study and user evaluation resulted in a set of variables, as identified in § 2.7.1. Specific statistical models and tests were employed to help answer the corresponding sub-questions, devised in § 1.7. In total, a set of 4 statistical objectives were devised to analyse the quantitative data:

- I: Relationship between the perceptual attributes and gaze exploration;
- II: Determining the influence of spatiotemporal 360-degree video complexity on gaze exploration;
- III: Assessing the interaction effect of usability context;

- IV: Defining the general consensus on the experiential statements.

The spatial- and temporal complexities of 360-degree videos were established as the primary independent variables of this thesis. The SI- and TI- values were used in the primary research objective, studying how variations in these values influence the amount of gaze distribution (N_ψ) expressed by users.

However, prior to analysing the influence of spatial- and temporal complexity on gaze distribution, the mediating effect of usability context, and the general consensus, a preliminary analysis was conducted on each of the perceptual attributes, enabling the identification of significant predictors of N_ψ. The inclusion thereof in statistical analyses further isolates the impact of spatial- and temporal image complexity and optimises the statistical models. Similar to the presence of diegesis (δ), as detailed in Chapter 4, as well as usability context (seating type), the attributes of cognitive perception were included as confounding variables in the statistical models, optimising the models in their assessment of spatiotemporal complexity on gaze dynamics.

I: Perceptual Attributes on Gaze Exploration

The multi-dimensional interaction process of 360-degree video interactions underpins the eliciting nature of cognitive and perceptual factors on behavioural responses [9, 76, 356]. The user evaluation encompassed a subjective user-evaluation, taking into account the influence of cognitive perception among users. Perceptual attributes demonstrating a significant association with N_ψ introduce risk of skewed estimations of the relationship between spatiotemporal complexities and N_ψ.

Therefore, to validate its inclusion as confounding variables in the statistical model and analysis of spatiotemporal complexity on N_ψ, the influence of each perceptual attribute on gaze exploration was analysed. The defined perceptual attributes of a 360-degree video interaction included QoE, engagement, attentional focus, spatial awareness and fear of missed content. The continuous, normalised and averaged, values were used as a measure against the quadrifactorial exploration index N_ψ. A linear mixed-effects analysis was conducted to identify significant predictors among the set of perceptual attributes. While most of the perceptual attributes were all measured once per user, both QoE and N_ψ were measured multiple times per user. A multiple linear regression model would not suffice due to this mixed structure in the dataset, as repeated measures from the same user are more prone to correlation than the other perceptual attributes. As such, violating the independence assumption of linear regression. The linear mixed-effects model is more robust against the within-subject correlation in the repeated measures, which utilises both fixed and random effects. The significant predictors among the set of perceptual attributes were used in the subsequent statistical models as confounding variables.

The following hypotheses were formulated for the analysis of the association between each of the perceptual attributes and gaze distribution:

- H_0: There is no significant relationship between [perceptual attribute] and degree of gaze distribution (N_ψ).
- H_1: There is a significant relationship between [perceptual attribute] and degree of gaze distribution (N_ψ).

By conducting the preliminary analysis, the significance of the relationship between each of the perceptual attributes on the amount of gaze distribution was assessed. The linear mixed-effects model validated the inclusion thereof in the primary spatiotemporal complexity analysis of gaze distribution, reducing unnecessary model complexities. The results of the linear-mixed effects analysis are presented in § 5.2.1.

II: Spatiotemporal 360-Degree Video Complexity on Gaze Exploration

The main research objective addresses the significance of spatial- and temporal image complexity on gaze exploration and distribution in 360-degree video user interaction. As such, the primary

focus of this statistical objective was to analyse the influence of spatiotemporal complexity on the degree of gaze exploration, imperative to the understanding of the complex dynamics between 360-degree video complexity and user behaviour in VR environments. The computation of spatial- and temporal complexities of the selected 360-degree videos was detailed in § 2.2.4.

To increase reliability of the study, it was important to ensure the effect size on degree of exploration was isolated to only variations in spatial- and temporal complexities. Therefore, prior to the analysis, a series of confounding variables were identified and highlighted. The preliminary diegetic assessment and linear mixed-effects model identified and validated the presence of diegetic artefacts and significant attributes of perception as confounding variables. The proposed study design, and implemented techniques to reduce random effects such as the Latin square design to minimise order-bias, further increased reliability of the assessment.

A mixed-effects multiple regression model was employed to accommodate the hierarchical structure of the data and repeated measures, taking into account both fixed and random effects. The hierarchical data structure was due to the repeated measures design in which each user was exposed to all six 360-degree video, resulting in multiple data recordings per user. The independent variables employed in the statistical model were the spatial complexity (SI) and temporal complexity (TI). The dependent variable implemented in the model was the quadrifactorial exploration index N_ψ. The mixed-effects multiple regression employed the significant perceptual attributes, as determined by analytical objective I, the presence of diegetic artefacts δ, and seating type as control variables.

The following hypotheses were formulated for the analysis of the association between spatial- and temporal complexities and gaze distribution:

- H_0: There is no significant relationship between spatial- and temporal image complexities and degree of gaze distribution (N_ψ).
- H_1: There is a significant relationship between spatial- and temporal image complexities and degree of gaze distribution (N_ψ).

The direction and magnitude of the relationships were explained through the interpretation of the coefficients. The R^2-value was utilised to assess how respective spatial- and temporal complexities explain the variance in N_ψ, taking into account the control variables and random effects. This analytical approach defines the understanding and facilitated a nuanced understanding of how spatiotemporal image complexity in 360-degree video impacts gaze behaviour, addressing the main research objective from § 1.7. The results of the mixed-effects multiple regression analysis are presented in § 5.2.2.

III: Interaction Effect of Usability Context

The between-subjects design enabled the acquisition of gaze data between two groups, varying in respective usability context. Motivated by the multi-modal output of 360-degree video, which highlights the notion that not every 360-degree video will be seen in a similar way, two different usage situations (i.e., seating types) were compared. Users in group R were seated on a rotating chair, whereas users of group F were seated on a fixed-position chair. The rotating chair enabled full control and provided easier usability for viewing the entire 360-degree image. In contrast, the fixed-position chair limited the user's movement. Notably, when comparing the resulting aggregate gaze distribution heatmaps between both groups, a significant difference can be observed in gaze patterns. This observation further indicates a notable influence due to usability context (i.e., seating type). The aggregate heatmaps of group R and F are presented in Figures A.3 and A.4, respectively (see Appendix A3).

A combination of both a subgroup and interaction analyses were conducted to assess the different influences of spatial- and temporal complexities on gaze distribution across the two seating types. The spatiotemporal complexities (SI- and TI-values) were employed as independent variables, and N_ψ as the dependent variable.

The subgroup analyses were conducted separately for each group, as the sample was split into two groups (R and F). The subgroup analysis examined whether the relationship between

spatiotemporal image complexity and gaze distribution was different across the two groups. For the subgroup analyses, the following hypotheses were formulated:

- H_0: There is no significant difference in the relationship between spatial- and temporal image complexity and degree of gaze distribution (N_ψ) across different seating types.
- H_1: There is a significant difference in the relationship between spatial- and temporal image complexity and degree of gaze distribution (N_ψ) across different seating types.

Additionally, the interaction analysis specifically assesses how this effect of SI- and TI-values on N_ψ was moderated by seating type. Seating type was a controlled condition in the study, and depending on the level, could potentially change the nature or strength of the relationship between image complexity and gaze. This interaction suggests a moderation, resulting in the proposed moderation analysis rather than a mediation analysis. For the interaction analysis, the following hypotheses were formulated:

- H_0: There is no interaction effect between seating type (R/F) and spatial- and temporal image complexity on gaze distribution (N_ψ).
- H_1: There is an interaction effect between seating type (R/F) and spatial- and temporal image complexity on gaze distribution (N_ψ).

The examination of the difference in effect size across groups, as well as the interaction terms, assesses the underlying mechanisms of the effect of spatial- and temporal image complexity on gaze distribution and the dependency on seating type. As such, this analytical approach enabled a better understanding of how the context of use, specifically seating type, interacts with the spatiotemporal image complexities to affect the user's gaze distribution. The results of the usability group moderation analysis are presented in § 5.2.3.

IV: General Consensus on Experiential Statements

The fourth analytical objective was to examine the general consensus across the users on each of the user evaluation statements. As part of the user evaluation, statements were provided relating to the user interaction and self-perception of gaze behaviour. In total, each user was given six statements: two of which were unique to the group they were in. The other four statements were rated by all $n = 52$ users. The provided statements are presented in Table 4 (see § 2.3). Users utilised a 7-point Likert scale to denote their respective level of agreement on each statement. The general consensuses across users enabled complementary insights in the self-perception of gaze behaviour during the 360-degree video interaction.

The distributions of the ratings were examined using the descriptive statistics of the statements, enabling a preliminary understanding of the general consensus. Furthermore, to identify if there were significant differences in the ratings between group R and F, a series of non-parametric tests were employed.

The Mann-Whitney U test was utilised to assess the differences between the groups for each of the four common statements. The choice of non-parametric tests was motivated by the ordinal nature of the ratings. For each of the four common statements, rated by all users, a test statistic and p-value was acquired. For each test, the following hypotheses were formulated:

- H_0: There is no significant difference in the median rating for a given statement x between the two seating types.
- H_1: There is a significant difference in the median rating for a given statement x between the two seating types.

The examination of the central tendencies and distribution of the ratings for each of the experiential statements enabled a more nuanced understanding of the users' self-perception of

gaze behaviour during the 360-degree video interaction. The employment of the non-parametric statistical test enabled further insight in the general consensuses across the independent groups. The descriptive statistics, as well as the results of the Mann-Whitney U tests, are presented in § 5.2.4.

Qualitative Approach

The user evaluation encompassed the use of a semi-structured interview to acquire qualitative data on the user's self-perception of expressed gaze behaviour. This part of the evaluation was predominantly focused on acquiring subjective information on how the user's would describe the behavioural impact of various 360-degree videos. As such, a grounded theory analysis was performed utilising the qualitative data acquired during the semi-structured interview, allowing for effective categorisation of the transcripts. Consequently, the following analytical objective was devised:

- V: Defining the key trends in the users' self-perception of gaze behaviour.

Variations in usability context and the unrestricted orientation of the virtual environment, as implied by Löwe et al. (2015), could lead to varying interpretations of the content [194]. As such, during the semi-structured interviews, users were asked to provide a brief description of what they had seen. This approach enabled the filtering of qualitative data acquired from users who had a distinctly different interpretations of the content compared to the other users. All $n = 52$ users had similar interpretations of the content and therefore no qualitative data was excluded from the analysis.

V: Key Trends in Self-Perception of Gaze Behaviour

Analysing the acquired qualitative data on the user's self-behaviour, and to identify the trends and patterns in the conscious and active gaze behaviour of the users. The qualitative analysis was performed using emergent coding in alignment with Straussian Grounded Theory [318, 340]. During the grounded theory analysis, the data was systematically coded, categorised and labelled. It was decided to employ the Straussian Grounded Theory methodology rather than the Glaserian Grounded Theory, due to the implementation of more explicit coding procedures [333].

The emergent coding was performed utilising three coding procedures: open, axial and selective. The open coding procedure was utilised to assign codes to different strings of text from the transcripts. This approach enabled the data to be dissected into discrete parts, which could then be compared for similarities and differences. Each data chunk was labelled subsequently. During the axial coding procedure, the data chunks were systematically categorised using a combination of inductive and deductive thinking. This process involved relating the different codes, which encompass the identified categories and concepts, to each other based on causal relationships, context, consequences and conditions. The labelled data chunks of the same code were combined into clusters based on their respective commonalities. Lastly, as part of the selective coding procedure, core categories were identified and linked to each other, serving as the central theoretical concepts around which other categories are related. As such, the relationships between data chunks, codes and categories could be closely examined. The culmination of the inter-connected concepts and relationships, resulted in a theoretical framework of key trends and patterns in the conscious gaze behaviour of the users.

The results of the grounded theory analysis are presented in § 5.3. The theoretical framework of key trends and patterns in the conscious gaze behaviour of the users are presented in § 5.3.2.

2.7.4 Python Libraries and Visualisation Techniques

The mathematical computation of spatiotemporal image complexity, formulation of the quadrifactorial exploration index, diegetic assessment and the statistical analyses were conducted using

Analysis	Sub-Question
I	To what degree do attributes of cognitive perception impose a confounding effect on the user's behavioural response?
II	How is gaze behaviour affected by spatial image complexity? How is gaze behaviour affected by temporal image complexity?
III	To what degree is the effect of spatial- and temporal image complexity on gaze distribution mediated by usability context?
IV	How does the user's self-perception of conscious gaze behaviour compare with gaze data?
V	How does the user's self-perception of conscious gaze behaviour compare with gaze data?

Table 6: Analytical Framework

Python. This subsection describes the employed statistical and mathematical Python libraries and packages.

Data manipulation and analysis was performed using `Pandas`. The `NumPy`, `SciPy` and OpenCV's `cv2` libraries were employed for various numerical calculations, array manipulations, interpolation and additional statistical functions. Additionally, the `StatsModels` library was used for the estimation and inference of the presented statistical models. Various modules, e.g. `StatsModels.stats`, were employed to facilitate additional model diagnostics and inferential statistics. Model selection and validation was done using `scikit-learn`.

For data visualisation, model checking and assessment of underlying assumptions, a variety of visualisations were constructed. Primarily, `Matplotlib` was employed for static and interactive visualisations and basic plot structures. The `Seaborn` was used for more complex statistical graphics. Additional modules, such as `Plotly` and `Axes3D` were used for the advanced three-dimensional visualisations. Each script included specific Python libraries, packages and modules, tailored to the specific objective of the script or analysis.

Chapter 3

Quadrifactorial Exploration Index N_ψ

Research on the behavioural influence of spatial- and temporal complexity on gaze dynamics in 360-degree video interactions required a comprehensive method to quantify the amount of exploratory gaze behaviour expressed by users during the eye-tracking study. While traditional gaze metrics provide information about location and duration of fixations, they fall short in expressing the complexity of gaze distribution patterns during 360-degree video interactions. Fixation-based gaze interpretation remain limited in its representation of exploratory gaze behaviour, providing either coverage area or intensity values. Moreover, due to the susceptibility of fixation points to saliency bias, these representations are not truly representative of the user's attention or exploration. The quantification of complex gaze patterns was further motivated by the ambiguity of subjective human interpretation of fixation heatmaps. While the human visual system is capable of distinguishing visual differences across various heatmaps, complex statistical analyses on the influence of spatiotemporal image complexity required a more reliable and quantifiable metric that encompasses the exact level of gaze distribution rather than categorical (i.e., low vs. high exploration).

As such, the one-dimensional approach of using traditional gaze data points to interpret gaze distribution, such as fixations, do not sufficiently capture the underlying structure and distribution of gaze patterns, nor do they support the variability of gaze behaviour across the different types of content. Quantifying the complexity of gaze distribution requires a multi-faceted approach that accounts for the spatial extent and concentration of gaze patterns, as well as for structural differences and randomness of intensity distribution. As such, this chapter aims to answer the devised sub-question from section 1.7:

> How can computer vision techniques, paired with eye-tracking data, be employed to quantify gaze patterns?

To achieve this, a structural approach has been employed to develop a novel metric, the Quadrifactorial Exploration Index N_ψ. The metric utilises the gaze distribution heatmaps, extracted from the eye-tracking software used during this study, to compute an exploration index. This index value represents the degree of gaze exploration and distribution expressed by the user while watching the 360-degree video. As such, the index is representative of the amount of exploratory gaze behaviour, measuring the degree of expressed gaze distribution.

The quadrifactorial exploration index N_ψ utilises the generated heatmaps from the iMotions VR eye-tracking software to quantify the level of gaze exploration based on a set of four factors: coverage area ratio, average intensity, structural dissimilarity, and entropy. As such, the quadrifactorial exploration index not only captures the variability in gaze behaviour across different types of content, but also enables complex analyses on the relationship between spatial- and temporal complexity and gaze distribution. This chapter details the construct and formulation of the quadrifactorial exploration index N_ψ, which quantifies the degree of exploration, enabling complex computations and an advanced analysis of viewing behaviour.

3.1 Area Coverage Ratio and Average Intensity

The iMotions eye-tracking software was used to generate a gaze distribution heatmap based on fixation points and duration, which visualises the extent and intensity of viewing exploration and map the user's attentional landscape. The quadrifactorial exploration index N_ψ uses the net heatmap coverage area and heatmap intensity as the basis indicators of exploration, as area and intensity pose important factors in the quantification process of exploration behaviour. The term area denotes the spatial extent of the gaze distribution heatmap, interpreted as the areas of the 360-degree video frame that users have seen. A large coverage area suggest that users have explored more of the 360-degree scene. Similarly, a smaller coverage area in the gaze distribution heatmap indicates a more concentrated level of exploration.

The average intensity represents the degree of focus in the user's gaze behaviour. Users who have looked extensively at specific areas of the 360-degree scene resulted in higher intensity values, while lower intensity values indicated a more evenly distributed exploration pattern. The heatmap

coverage area is representative of the distribution of gaze data, with varying levels of intensity denoting the degree of gaze concentration.

Pre-Processing To ensure uniform and reliable computations, based on gaze distribution heatmap image signals, pre-processing of the imagery was required. Firstly, the iMotions eye-tracking software was used to generate individual gaze distribution heatmaps, based on the user's fixation points (x and y coordinates) and duration (timestamps). The heatmaps were generated separately for each video across all users. Gaze distribution heatmaps were superimposed on a uniform white background (i.e., white image frame). As such, the resulting heatmaps were of the same dimensions across the entire set of heatmaps. The superimposing of the gaze distribution heatmap on a white image frame enables quantification of the coverage area by implementing binary masks, as discussed in subsequent sections. Secondly, the heatmaps were converted to greyscale images to quantify the intensity value of each heatmap pixel. Lastly, as part of the pre-processing of heatmaps, the intensity values were normalised to scale 0 - 255 to accommodate valid comparisons.

The area coverage ratio A measures the spatial extent of the user's gaze exploration. To achieve this, a binary mask was created to identify the non-white heatmap pixels n that are part of the gaze distribution heatmap and the white pixels as part of the white image frame layer. The binary mask is an array with the same dimensions as the input heatmap, and sets each pixel value to `true` or `false` based on a pixel intensity threshold value. The threshold was set at 254. Consequently, any pixel with an intensity value less than the threshold (i.e., any non-white pixel) was considered part of the heatmap. A is calculated by dividing the number of non-white pixels n by the total number of pixels N in the heatmap image signals. N is uniform across all heatmaps, due to the identical dimensions and resolution. To enable normalisation and interpretation of the final quadrifactorial exploration index N_ψ, the heatmap coverage area is expressed as a ratio, generating a value in [0, 1] and enabling comparison across all heatmap image signals. While not applicable to the heatmaps used in this study, the use of a ratio expression also accounts for variations in image size. The resulting area coverage ratio A is expressed as:

$$A = \frac{n}{N} \tag{22}$$

The area coverage ratio calculates the ratio of which pixels are part of the heatmap. As such, a value of 1 indicates that all the pixels in the image are affected by the heatmap (i.e., all areas of the 360-degree video scene have been explored). Similarly, and $A = 0$ indicates that no pixels are affected by the heatmap (i.e., no exploration has occurred).

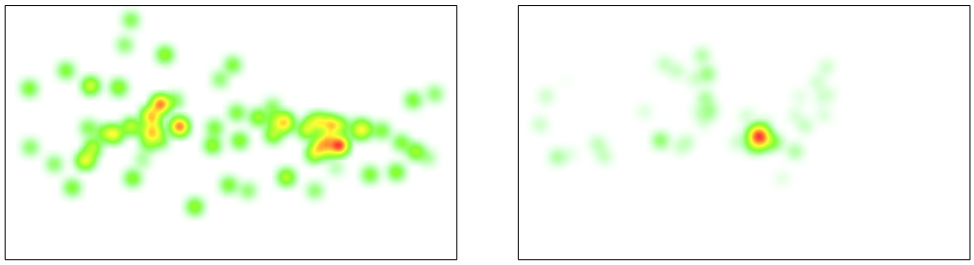
Secondly, the intensity values of the heatmap are indicative of gaze concentration, with higher intensities indicating greater levels of gaze focus. As such, the heatmap coverage area alone does not suffice in accurately representing the level of exploration. By calculating the average intensity I of the non-white pixels n from the coverage area, the overall gaze intensity within the heatmap area can be expressed. To assess the average level of gaze focus, while accounting for the number of non-white heatmap pixels n and intensity values, the average intensity I can be expressed as the sum of all intensity values of n, divided by n:

$$I = \frac{\sum_{i=1}^{n} I_x}{n} \tag{23}$$

where I_x is the intensity value of the x-th non-white pixel.

The average intensity was normalised, using the maximum intensity value of 255, to ensure the index is expressed in the same range [0, 1] as the area coverage ratio. Consequently, this was done for each of the factors to ensure the final expression of the quadrifactorial exploration index N_ψ falls within range [0, 1]. The normalised intensity (I_{norm}) is expressed as:

$$I_{norm} = \frac{I}{255} \tag{24}$$



(a) $E_1 = .271$ (b) $E_1 = .128$

Figure 11: E_1 values of heatmaps with relative high (a) and low levels (b) of exploration.

Lastly, the area coverage ratio A and normalised average intensity I_{norm} were combined to obtain a single metric E_1 that is, to some extent, representative of exploratory behaviour. Using a multiplicative approach ensured that both area coverage ratio and normalised average intensity contribute to the final index without neither component dominating the other. As such, E_1 can be expressed as:

$$E_1 = A \times I_{norm} \tag{25}$$

The expression of E_1 provides an initial measure of gaze distribution by considering both area coverage and intensity values. The E_1 initial exploration index was computed using a custom Python script, which is presented in Appendix B6. As indicated in Figure 11, the amount of gaze distribution can be expressed as a multiplicative function of A and I_{norm}. In this figure, the gaze distribution heatmap signals as superimposed on the white image frame are presented.

3.2 Structural Dissimilarity and Entropy

The combination of heatmap area and intensity denote to the spatial extent of individual gaze patterns. As such, it can be seen as an initial indicator for gaze distribution. However, the area coverage ratio does not account for the distribution or variability in gaze patterns. Specifically, a large area could have a low exploration density and a high exploration density could be identified in a relative small coverage area. The area coverage ratio provides a general sense of gaze distribution, but remains limited in capturing the complexity or diversity of gaze distribution. For example, the metric does not differentiate between a gaze distribution which is spread out and a gaze distribution that frequently moves between a few specific areas.

Furthermore, while intensity values indicate the duration of gaze points in specific areas, the intensity values do not account for the complexity and spatial distribution of such gaze patterns across the heatmap image signal. It provides a general indication of the overall level of gaze concentration, but it does not distinguish between a gaze distribution which is focused on a small area and a gaze distribution that is evenly distributed with a lower intensity at each pixel. Specifically, a user that has explored more of the virtual 360-degree environment has perceived less detail throughout the scene as compared to a user with a very concentrated gaze. Similarly, areas with higher intensity values do not always indicate extensive exploration if the gaze was only concentrated on a few confined areas. Therefore, it is important to account for the nuances gaze pattern interpretation. To overcome these limitations, image processing techniques form the domain of computer vision were be implemented.

Firstly, as detailed in § 1.5, advancements made in image segmentation and similarity techniques have led to significant progress in complex imagery analyses. As before-mentioned, the HVS is capable of identifying visual differences between various heatmap signals. However, it requires complex computational techniques to exactly quantify such differences. Inspired by the work of Cui et al. (2021), the quantification of visualisations and graphical representations of data

can be achieved through complex computations of image similarity techniques, which determine the degree of similarity between two image signals [18, 72, 214, 249].

Emphasised by Hore et al. (2010), the objective methods to evaluate image signals are commonly based on comparisons utilising explicit numerical criteria [46, 131, 218]. As such, assessing the exact structural differences between the generated heatmap signal and the white image signal requires computational methods on the per-pixel-level. Consequently, a computational model for deriving a similarity index between two images was implemented. As introduced by Wang et al. (2004), the structural similarity index SSIM, as a measure to express similarity between two images, has been proven effective in various areas of research. SSIM outperforms common metrics and state-of-the-art perceptual image quality measures, mostly due to its strong correlation with the human visual system [14, 131, 221, 345]. In situations where traditional error summation methods are not sufficient, the SSIM index has demonstrated its efficiency. Moreover, its high correlation with HVS denotes SSIM as an effective tool to analyse the heatmap image signals, making it a reliable technique to compare complex heatmap image signals. The mathematical construct and inner workings of the SSIM model is detailed in § 1.5.4.

However, implementing SSIM as a measure to express the amount of gaze distribution based on a gaze distribution heatmap signal, required a modified approach. The heatmap signals make use of colour intensity to indicate the density of data points. As such, the coverage area and colour intensity variation are essential components of interpreting the gaze visualisation in a gaze distribution heatmap. Due to the single-scale structure of SSIM, the implementation thereof to compare heatmap signals which contain multiple scales and colours does not suffice. Specifically, similarity assessments using heatmap image signals, rely strongly on the different intensity levels of the colours. To account for the the multi-scale representation of the heatmap image, as occurs through the colour encoding on a two-dimensional plane during heatmap generation, an enhanced version of SSIM that operates across the multiple scales of coloured heatmaps was implemented [122, 250, 265, 304, 336]. The Multi-Scale Similarity Index Measure MS-SSIM calculates the comparison of contrast and structure at each scale, while luminance is only computed at the highest scale. Particularly, MS-SSIM is more applicable for heatmap analyses than SSIM, as it incorporates colour information, evaluates similarity at multiple scales and is more sensitive to changes in the image structure. In the context of gaze distribution heatmap analysis, which contain multiple levels of details, this approach proves significantly more effective. Another approach, Complex Wavelet-SSIM was considered, but was disregarded due to its focus on image scaling, translation and rotation [107, 273]. The multi-scale structure of MS-SSIM provides more accurate and reliable results in the context of assessing the colour nuances across gaze distribution heatmap signals [82, 305, 346]. Mathematically, MS-SSIM is expressed as:

$$\text{MS-SSIM}(x,y) = [l_M(x,y)]^{\alpha_M} \prod_{j=1}^{M} [c_j(x,y)]^{\beta_j} [s_j(x,y)]^{\gamma_j}$$

where the parameters are selected such that $\alpha_M = \beta_j = \gamma_j$ for all j and $\sum_{j=1}^{M} \gamma_j = 1$. The exponents α_M, β_j, and γ_j enable the assignment of different weights to the segmented ROIs in the image signals. The entire mathematical construct and underlying formulation of the MS-SSIM model is detailed in § 1.5.6. A custom Python script has been developed to compute the MS-SSIM value between the two image signals, as presented in Appendix B7.

Utilising MS-SSIM to assess the degree of similarity between the heatmap image signal and the white image frame (on which the gaze distribution heatmap signal is superimposed) makes use of the same greyscaled heatmap image signals as previously generated for E_1. Due to the overlay of the heatmap on a white image frame, the change in pixel values is exclusively a result of the colour changes from the gaze distribution heatmap. While otherwise recommended for heatmaps superimposed on a video frame [221, 345], conversion to the YCbCr was not necessary in this context as the gaze distribution heatmap is responsible for all texture- and edge- changes present in the signal. Furthermore, the white image signal on which the gaze distribution heatmap is superimposed represents a completely uniform (i.e., no gaze pattern) distribution, enabling a

comparison which denotes the degree of similarity between the two image signals and enhancing the overall level of reliability of the metric. Another approach would involve comparing the individual heatmap signal to an aggregate heatmap signal or a heatmap with a predefined pattern. However, the MS-SSIM computation would be more focused on the deviation from a predefined or average gaze pattern, rather than capturing gaze distribution of the user.

Furthermore, MS-SSIM expresses an index that denotes the degree of similarity between the heatmap image signal and the white image frame. As such, a value of 1 indicates identical images and 0 indicates no similarity. However, this value does not represent the structural difference in the image signal as produced by heatmap on the white frame, but rather indicates the white pixels that are not affected by the heatmap projection. By reversing the scale, the resulting value represents the amount of dissimilarity and differences between the heatmap images signal and the white image frame. Calculating the dissimilarity results in a measure that represents the structural difference due to the presence of the heatmap image signal. The dissimilarity index d is expressed as:

$$d = 1 - \text{MS-SSIM} \tag{26}$$

where MS-SSIM = MS-SSIM(x, y), such that:

$$d = 1 - ([l_M(x,y)]^{\alpha_M} \prod_{j=1}^{M} [c_j(x,y)]^{\beta_j} [s_j(x,y)]^{\gamma_j})$$

In this context, d is employed to measure the structural dissimilarity between the white image frame and the superimposed heatmap signal. In this new interpretation, a value of 0 indicates no exploration as the heatmap image signal is identical to the white image frame, and a value of 1 indicates maximum exploration as the heatmap is completely different from the white image frame.

The motivation for implementing dissimilarity as a metric lies in the sensitivity to spatial distribution of MS-SSIM. The structural dissimilarity evaluates the degree to which the heatmap image structure deviates from the uniform distribution (no gaze pattern) of the white image frame. Moreover, the use of MS-SSIM renders the exploration index robust to variations in intensity, as changes in brightness and contrast are additionally accounted for. While not applicable to this study, as all gaze distribution heatmap images are of the same dimensions and generated the same way, this robustness enables stability even when different visualisation techniques are combined. Lastly, since MS-SSIM is sensitive to structural information, the comparison with the white image frame will highlight areas where the heatmap contains more detailed and localised patterns, enabling a comprehensive understanding of how users explore the 360-degree video scene at different spatial scales.

However, in constructing an expression that denotes the amount of exploratory behaviour based on the gaze distribution heatmap, MS-SSIM alone may not be sufficient for quantifying gaze distribution due to its design to assess perceptual similarity between two images rather than the extent of exploration. That is, d provides an understanding of the overall difference in gaze patterns, however it does not capture the distribution or variability of the gaze points within the heatmap itself. To provide a more comprehensive representation of gaze distribution, it was decided to incorporate entropy as a factor in the expression.

Entropy offers a measure of the unpredictability and randomness of the gaze distribution patterns and associated heatmap intensities. Diverse and less predictable exploration patterns, suggestive of a more broadly distributed gaze pattern, are indicated by high values of entropy. Similarly, uniform and predictable exploration patterns are indicated by lower entropy values, suggesting a higher levels of focus on specific regions or AOIs. As such, the entropy value entails valuable information on the diversity of exploration patterns, scene complexity and gaze distribution variability, enhancing the quadrifactorial exploration index.

As a measure of complexity and randomness of the heatmap signal, entropy is expressed as:

$$H(x) = -\sum_{I=1}^{L} P(I) \log_2 P(I) \tag{27}$$

where L is the number of intensity levels in the heatmap, and $P(I)$ is the probability distribution of intensity level I. As the total number of pixels present in each heatmap image signal is expressed as N, it was decided to express the total number of possible outcomes in this formula as L to denote the possible levels of intensity. The probability distribution $P(I)$ of all intensity levels was computed by normalising the histogram of all intensity levels I:

$$P(I) = \frac{n_I}{N}$$

where n_I is the number of pixels with intensity level I and N is the total amount of pixels in the heatmap image signal.

The greyscale heatmap image signal ensures that the entropy computation is not influenced by any additional colour information in the heatmap, as the distribution of gaze points is primarily represented by the intensity values [0, 255]. By using the before-mentioned greyscaled heatmap signal for the calculation of the entropy value, the summation is computed over all possible intensity levels I [0, 255], thereby:

$$H(x) = -\sum_{I=0}^{255} P(I) \log_2 P(I) \tag{28}$$

Lastly, the entropy value was normalised to the range of [0, 1] on the maximum possible entropy for a greyscale image with 256 intensity levels to achieve compatibility with d and the other index measures:

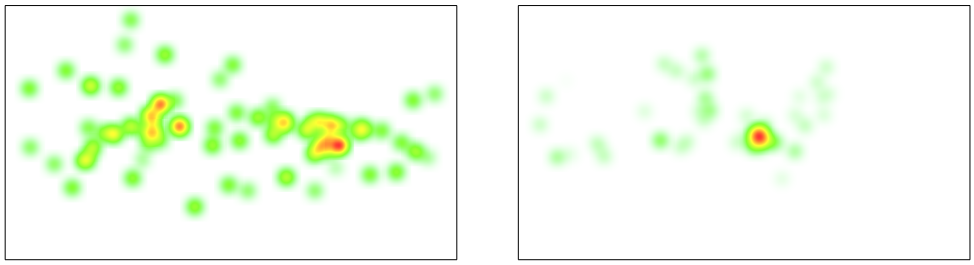
$$H(x)_{norm} = \frac{H(X)}{\log_2(256)} \tag{29}$$

Similar to the formulation of the initial exploration index E_1, the structural dissimilarity d and normalised entropy $H(x)_{norm}$ of the heatmap image signals were combined to form a complementary exploration index E_2. As such, the complementary index E_2 is expressed as:

$$E_2 = w_1 \cdot d + w_2 \cdot H(x)_{norm} \tag{30}$$

where $d = 1 - \text{MS-SSIM}$ and weights w_1 and w_2 were assigned to each metric based on factor importance, such that $w_1 + w_2 = 1$ to ensure E_2 is within range [0, 1]. Non-equal weights were assigned to d and $H(x)_{norm}$ due to the uncertain extent to which each factor contributes to the value of E_2. As structural dissimilarity was considered a direct indicator of gaze distribution based on the gaze distribution heatmap signal, the optimal value of $w_1 = .581$ was derived from the employed Principal Component Analysis on the dataset containing all values of structural dissimilarity d and normalised entropy $H(x)_{norm}$. This process is detailed in § 3.4. Consequently, to adhere to $w_1 + w_2 = 1$, w_2 was set to .419.

The combination of structural dissimilarity based on the Multi-Scale Structural Similarity Index Measure and entropy provides a complementary metric for the degree of gaze exploration expressed by the user, focused on the structural dissimilarity, complexity, randomness and uncertainty of gaze distribution in the heatmap image signals. Thereby, a maximum dissimilarity index indicates a completely difference heatmap image signal (as compared to the white image frame). Similarly, the maximum normalised entropy values indicates that the greyscale heatmap image signal has the highest possible complexity, variability, uncertainty or randomness in its pixel intensity values, indicating a gaze spread across a large area of the image with no single dominant region. Consequently, a maximum value of $E_2 = 1$ suggest extensive gaze distribution and exploration throughout the heatmap image signal.



(a) $E_2 = .276$ (b) $E_2 = .104$

Figure 12: E_2 values of heatmaps with relative high (a) and low levels (b) of exploration.

The E_2 complementary measure of gaze distribution was computed using a custom Python script, which combines the resulting MS-SSIM value from the MS-SSIM script (see Appendix B7) and the computation of Shannon entropy, as presented in Appendix B8 [357]. The script utilises the OpenCV library for image processing, the numpy library for numerical operations, and the skimage library for computing the Multi-Scale Structural Similarity Index Measure and Shannon entropy. Contrary to the computation of E_1, a binary mask was not required to compute MS-SSIM and Shannon entropy, as they rely on the continuous nature of the pixel intensities.

As a complementary measure to express gaze distribution, based prominently on the structural dissimilarity and entropy values of the heatmap signals, E_2 generated similar results as E_1. Figure 12 presents the computed E_2 values for the same set of heatmap image signals used in Figure 11, indicating a similar compatibility as a suitable metric for expressing the amount of exploration based on gaze distribution heatmap images.

3.3 Formulation of the Quadrifactorial Exploration Index N_ψ

As both suitable metrics to determine the amount of gaze distribution expressed by the user, E_1 and E_2 utilise the generated gaze distribution heatmaps to make computations based on pixel intensity, spatial distribution, randomness, variation, complexity and structural image similarities. By implementing techniques derived from imagery analysis applications in the field of computer vision, both the initial exploration index E_1 and complementary exploration index E_2 use distinct approaches to index the level of gaze exploration. Demonstrated by Figures 11 and 12, both approaches produce similar results, indicating the validity of both metrics as measures of gaze distribution. This indication is further supported by the resulting correlation matrix (see Table 7) of the two metrics, as detailed in § 3.4. However, in case of larger gaze distribution heatmaps, E_1 produces higher index values. Consequently, larger gaze distributions produce a higher degree of deviation between the two exploration metrics. This deviation can be explained by the use of the average pixel intensity in the computation of E_1, as E_2 takes into account all pixel intensities and variations.

To overcome these limitations, a combination of both metrics was proposed. By combining the four factors from each metric, area coverage ratio, average intensity, structural dissimilarity and entropy, the discrepancies produced from E_1 and E_2 were minimised, while enhancing the overall reliability of the metric. As such, by combining these four factors, the Quadrifactorial Exploration Index provides a comprehensive measure of exploration, not only capturing the spatial extent and concentration of gaze patterns, but also the structural differences and intensity distribution randomness and variation. The quadrifactorial exploration index encompasses the extent and depth of exploration as well as complexity and variability of the gaze distribution patterns, denoted as:

$$\psi$$

The name reflects the multi-factorial nature of the metric, which captures various aspects

of exploratory behaviour in 360-degree video. Using a weighted approach, the quadrifactorial exploration index is initially expressed as the weighted sum $\beta_1 \cdot E_1 + \beta_2 \cdot E_2$, thereby:

$$\psi = \beta_1 \cdot (A \times I_{norm}) + \beta_2 \cdot (w_1 \cdot d + w_2 \cdot H(x)_{norm}) \tag{31}$$

where the weighted sum of structural dissimilarity index d (26) and normalised entropy value $H(x)_{norm}$ (29) are added to the multiplication of area coverage ratio A (22) and normalised average pixel intensity I_{norm} (24).

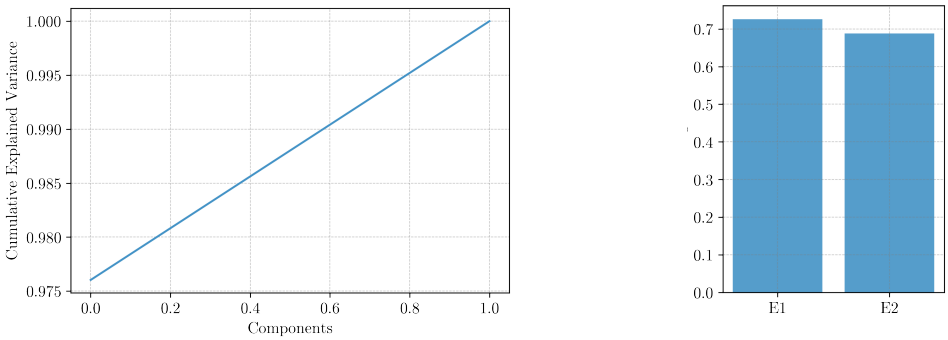
The additive model of $\psi = E_1 + E_2$ was motivated by the independent contributions of each variable. As such, the additive model ensures interpretative ways to combine a multitude of factors. Furthermore, the model offers computational simplicity and ease of implementation, specifically due to the use of a single heatmap image signal. The additive model can be adjusted and extended by adding or removing factors to the expression, enabling flexibility and adaptability for future research. In contrast, formulating a multiplicative or nonlinear model assumes specific interaction and relationships, of which the existence has not been established in this thesis, between the factors. Consequently, any added complexity and potential or overfitting might produce redundant results. Weights β_1 and $beta_2$ were implemented in the formulation of ψ to emphasise the relative importance of each factor-combination in the exploration index. Thereby, the influence of E_1 and E_2 and its relative contribution is controlled, allowing for specific tailoring to this thesis.

3.4 Principal Component Analysis

To obtain an optimal mathematical formulation of the quadrifactorial exploration index N_ψ, a Principal Component Analysis (PCA) was employed, utilised to derive the optimal values for each of the weights in the expression of ψ. By applying PCA on the dataset containing values of all four factors, the relative contribution of each variable to the variance in the dataset was assessed. As a statistical technique, PCA reduces the dimensionality of an interrelated dataset by using orthogonal transformation to convert the data into a new set of variables, i.e., principal components. The data-driven approach of employing principal components ensures that the new set of linearly uncorrelated variables retain as much of the variance present in the original dataset. In the applicability of this thesis, PCA was employed to obtain the weights β_1 and β_2, determining the relative importance of each factor in the exploration index formula. An identical process was used to derive the optimal weights w_1 and w_2 as part of the expression of E_2. The area coverage ratio, average pixel intensity, structural dissimilarity and normalised entropy values were computed for each of the six 360-degree videos across all $n = 52$ users, resulting a dataset with $n = 312$ values for each weighted factor of (31). A custom Python script was developed to perform PCA on this dataset and calculate the values of ψ for each user and each video, based on the heatmap image signals. This Python script, uses `sklearn` library's PCA function and can be found in Appendix B9. Given the multi-factorial nature of dataset, PCA was employed to generate a single principal component, representing the maximum variance of the original data.

In the PCA process, each principal component has an associated eigenvalue, representing the variance explained by that component. The eigenvalues provide the significance of each component in explaining the variance in the dataset. By computing the eigenvalues and eigenvectors of the covariance matrix, which assesses the joint variability of each pair of factors to determine the direction of maximum variance in the dataset, the principal components that explain the most variance in the dataset were selected. The eigenvalues can be computed by multiplication of the explained variance ratios of each of the principal components with the total number of variables.

The explained variance ratio of the employed PCA was [.976, .024], indicating that the first principal component in the analysis explains 97.6% of the total variance in the dataset. Subsequently, the second component only accounts for 2.4% of the variation. Figure 13a presents the scree plot, in which this interpretation is supported. Adhering to the Kaiser criterion [36, 60], only factors with an eigenvalue > 1 were retained. Since the number of variables used in this PCA equals 2 (E_1 and E_2), the resulting eigenvalues were [1.9529542, 0.0479458], meaning that only



(a) Scree Plot

(b) Loadings Plot

Figure 13: Scree (a) and loadings (b) plot of the first principal component.

the first principal component (eigenvalue = .1952) adheres to the Kaiser criterion. Consequently, only the first principal component was retained, justified by the explained variance ratio values, scree plot (see Figure 13a) and Kaiser criterion.

The loadings (i.e., weights) for the expression of ψ were extracted using the first principal component. These loadings can be interpreted as the relative importance of each variable in explaining the variance in the dataset and were used to assign values to weights β_1 and β_2. The PCA loadings resulted in weights $\beta_1 = .72568339$ for the interaction of area coverage ratio and average intensity values, and $\beta_2 = .68802879$ for the weighted sum of structural dissimilarity and normalised entropy values. As such, E_1 and E_2 contribute almost equally to the value of ψ, as was previously stated in the comparison of the index values of Figure 11 and 12. A high correlation of .952 was found between E_1 and E_2, as presented in the correlation matrix in Table 7. As such, both the initial exploration index and the complementary exploration index provide similar information about the degree of gaze distribution as expressed by the users.

	E_1	E_2
E_1	1.000000	0.951928
E_2	0.951928	1.000000

Table 7: PCA correlation matrix of E_1 and E_2

Weights β_1 and β_2 are indicative of how much E_1 and E_2 contribute to the first principle component, acting as the coefficients in the expression of the quadrifactorial exploration index ψ. A similar approach was applied for the computation of optimal weights $w_1 = .581$ and $w_2 = .419$ as part of the expression of E_2 (30). The final 3 decimal weights were calculated using the 8 decimal values of each weight resulting from the PCA. Substituting the computed weight values of β_1, β_2, w_1 and w_2 into the formula of ψ (31) as:

$$\psi = 0.726 \cdot (A \times I_{norm}) + 0.688 \cdot (0.581 \cdot d + 0.419 \cdot H(x)_{norm}) \tag{32}$$

Simplification of the expression gives:

$$\psi = 0.726 \cdot (A \times I_{norm}) + 0.399 \cdot d + 0.289 \cdot H(x)_{norm} \tag{33}$$

where the area coverage ratio A, normalised average intensity I_{norm}, structural dissimilarity d and normalised entropy $H(x)_{norm}$ values of the heatmap image signals are proportionally contributing to the index ψ. The proportions, as determined by their respective weights β_1, β_2, w_1 and

w_2, were assigned by using the data-driven approach of the principal component analysis. As such, the factors maximise the variance explained in the dataset. The dataset was not standardised, as all included factors were already in range [0, 1], ensuring that all factors contribute equally to the PCA while not influenced by differences in scales.

Due to the implementation of optimal weights, and $\beta_1 + \beta_2 \neq 1$, the resulting value of ψ contains a theoretical maximum value of 1.414. The theoretical maximum value was derived using the sum of both weights β_1 and β_2. To ensure ψ remains consistent with all four factors A, I_{norm}, d and $H(x)_{norm}$ and is expressed within [0, 1], all values of ψ were normalised to range [0, 1]:

$$N_\psi = \frac{\psi}{1.414} \tag{34}$$

where N_ψ denotes the final value of the quadrifactorial exploration.

The main objective of performing PCA was to derive the optimal weights in the expression of the quadrifactorial exploration index such that the weighted sum of E_1 and E_2 explains the maximum possible variance in the data. Consequently, by using the loadings as weights, the exploration index ψ poses a comprehensive and reliable measure of gaze distribution as it mitigates the risk of over-emphasising one factor over the other three based on arbitrary or subjective decisions. The data-driven approach of PCA ensures an accurate reflection of the inherent structure and correlations within the dataset of A, I_{norm}, d and $H(x)_{norm}$. It is important to adhere to the normalised value of ψ (N_ψ), due to the novel nature of the metric, as well adjust accordingly for future iterations of the quadrifactorial exploration index. The optimal weights were derived based on the dataset containing all values for A, I_{norm}, d and $H(x)_{norm}$, and as such, are the specific optimal values of β_1 and β_2 for this study. Larger datasets or different research applications may result in a slight alterations of the optimal weights. The normalisation process ensures that despite alternate weight-values, the index still produces a value in range [0, 1].

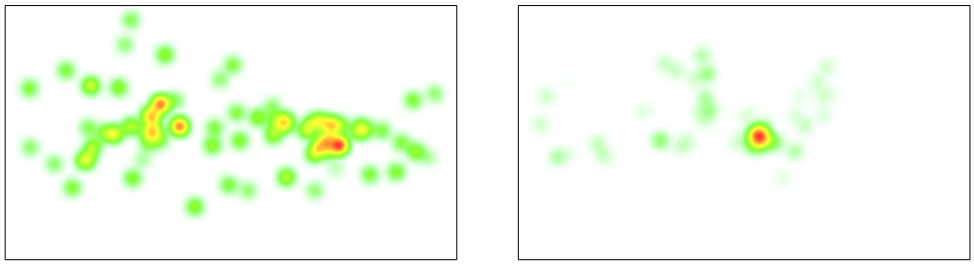
	E_1	E_2	N_ψ
Low	.128	.104	.117
High	.271	.276	.272

Table 8: Comparative matrix of E_1, E_2 and N_ψ values.

The resulting value of N_ψ, in range [0, 1] indicates the degree of gaze distribution performed by a user during a particular 360-degree video interaction in VR. A higher value of N_ψ suggests that the user's gaze was more widely and uniformly distributed across the scene, indicating a higher level of exploration. Conversely, a lower value of N_ψ indicates that the user's gaze was more focused on specific regions, suggesting less exploration and more concentrated attention. As such, the difference in level of exploration can be derived from the N_ψ-values, as presented in Figure 14, which uses the same gaze distribution heatmap signals as Figures 11 and 12. A comparative overview of the three metrics for low and high levels of gaze exploration is presented in Table 8. The required computations made to derive N_ψ have been combined in a final Python script, presented in Appendix B10. The script can be employed to compute the N_ψ-value of any heatmap image signal superimposed on a white image frame.

Reliability Cronbach's alpha was used to determine internal consistency and assess the reliability of the novel metric, imperative to the validation of the quadrifactorial exploration index N_ψ as a reliable metric of gaze distribution [67, 314]. Components E_1 and E_2 exhibited excellent internal consistency as predictors of ψ, with a resulting Cronbach's alpha of .975. This result indicates that the quadrifactorial exploration index contains highly correlated components and, as such, reliably measures the same underlying constructs of gaze distribution in an attentional heatmap image.

The normalised quadrifactorial exploration index N_ψ enables complex analyses of spatial- and temporal complexity in relation to gaze distribution of 360-degree videos in VR. The index offers a



(a) $N_\psi = .272$ (b) $N_\psi = .117$

Figure 14: N_ψ values of heatmaps with relative high (a) and low levels (b) of exploration.

comprehensive yet nuanced understanding of gaze distribution, utilising gaze distribution heatmap imagery.

As the name suggests, the quadrifactorial exploration index incorporates a total of four factors. Firstly, the relative proportion of the 360-degree video frame that the user has explored indicates the extent of exploration. Secondly, the average intensity of the heatmap pixels denote the degree of gaze concentration within the explored areas. Thirdly, by using MS-SSIM, the structural dissimilarity signifies the diversity of exploration. Lastly, the entropy encapsulates the spread and complexity of gaze patterns by assessing the randomness or complexity of the gaze distribution signals. The factors were elegantly combined into a single weighted metric, utilising a data-driven approach based on the explained variance in the dataset. Furthermore, an internal-consistency assessment confirms the reliability of the novel metric in capturing and quantifying the exploratory gaze behaviour based on gaze distribution heatmap image signals.

In conclusion, the index encompasses not only the spatial extent and gaze concentration with which is explored, but also utilises the structural dissimilarity on multiple scales and entropy values to include the degree of concentration and diversity of gaze distribution. Therefore, the quadrifactorial exploration index N_ψ accounts for complexity and patterns which may not be evident through the analysis of traditional gaze metrics alone. Consequently the quadrifactorial exploration index reduces saliency-bias and is more indicative of the user's attentional and exploration patterns, regardless of gaze distribution position within the image frame. Lastly, the model's use of gaze distribution heatmap signals as input enables a higher degree of accessibility and user adoption, as heatmap imagery is more readily available.

Chapter 4

Diegetic Assessment δ

The unique format of 360-degree video in VR fundamentally shifts the paradigm of utilising cinematography to guide viewer attention. The immersive nature of VR and associated output modalities offer highly immersive viewing experiences, in which traditional use of camera angles, post-production processes and cinematographic principles are rendered sub-optimal [193, 375]. While traditional video content relies on camera angles to guide viewer attention and achieve high levels of attentional synchrony, 360-degree video enables complete freedom to explore the content. Despite the arbitrary methods of consuming omnidirectional content, Zink et al. (2019) established behavioural coherences across viewers and modalities. Therefore, the principle of common behavioural responses during 360-degree video interaction can be exploited by analysing the implemented cinematographic principles to generate predictions and estimates of the user's gaze behaviour.

The techniques applied to guide viewer attention in 360-degree environments heavily rely on the implementation of attentional guidance mechanisms and diegesis [23, 63, 375]. Diegesis, as a mechanism, has been proven extremely effective in guiding user attention [263, 284, 306, 341]. Derived from film theory, the implementation of diegetic mechanisms leverage non-verbal behaviour, using the internal story- / scene elements, to support the narrative in order to guide attention. These diegetic artefacts serve as a crucial tool for maintaining attentional continuity and guide gaze behaviour in 360-degree video.

Similar to edits in traditional cinematography, utilising character motion and non-verbal behaviour evokes a natural viewing orientation towards AOIs [100, 299]. Moreover, findings from current literature suggest that the inclusion of diegetic mechanisms increase levels of presence, user preference and user experience [48, 220, 262, 306]. Consequently, this positive enhancement of the viewing experience and high levels of presence elicits significant behavioural responses due to its correlation with user engagement [205, 230].

Despite the careful selection of 360-degree content, as detailed in § 2.2.3, inherent factors within the content itself remain influential to the viewer's sense of presence. In aims of achieving a positively enhanced user experience, the content adheres to Hall's model of proxemics by maintaining a cohesive camera height [119, 269]. Moreover, the selected content displays user-acknowledging behaviour by specific diegetic artefacts. As such, inherent factors of the selected content reduce risk of eliciting the Swayze effect, consequently enhancing the user's sensation of feeling a tangible relationship with the virtual environment [164, 224, 284, 338, 352]. This sense of presence is further emphasised by the heightened levels of place and plausibility illusion, predominantly driven by the sensorimotor contingencies of using head-mounted displays [227, 295, 296, 297].

Consequently, the cinematographic principles within a 360-degree video itself pose as a potential confounding factor in eliciting user behaviour. As such, in the study on how spatiotemporal complexity of a 360-degree video sequence influences gaze behaviour, the varying influence of visual artefacts was taken into consideration. Despite the systematic selection of the utilised 360-degree video sequences (see § 2.2.3), the presence of diegetic artefacts and attentional guidance mechanisms remain inextricable elements of the visual information and, as such, are inherent to the narrative world of the 360-degree video.

To account for the potential influence of these diegetic attention guiding mechanisms and visual artefacts on gaze behaviour, the diegetic assessment – as presented in this chapter – was performed. It aims to, partially, approach and answer the devised sub-question from § 1.7:

> To what degree do cinematographic principles impose a confounding effect on the user's behavioural response?

A systematic approach was employed, enabling for the coded identification, categorisation and quantification of such attributes within each of the six selected 360-degree videos. This chapter details the diegetic assessment process. A devised coding scheme was utilised to code the diegetic artefacts and elements (i.e., objects, persons and landmarks), enabling the quantification of how much attention-guiding content each 360-degree video contains, discussed in section 4.1.

As such, this diegetic assessment discusses the inextricable qualities of each of the selected 360-degree videos. A brief content description of each 360-degree video is presented in section 4.1

as well. Moreover, the identified attention-guiding content and visual artefacts are highlighted for each 360-degree video, as well as the resulting Diegetic Artefact Score δ. In section 4.2, an initial data exploration was conducted based on ratio- and reciprocate-values, to explore the complex relationship between the δ point-values and degree of gaze distribution N_ψ. A series of non-linear regression analyses were performed to model and assess the complex association of δ and N_ψ, as presented in section 4.3. Lastly, a nuanced interpretation of these findings is presented in section 4.4.

4.1 Coding Scheme

The devised coding scheme was central to the diegetic assessment, as it systematically enabled quantification of the diegetic artefacts present in each video using predefined and operationalised criteria. The influence of diegetic artefacts on gaze behaviour was quantified using a coding scheme, operating on two dimensions: the relative visual size of the artefacts and the duration of its presence in the 360-degree video. The motivation behind the use of these two dimensions was as follows: smaller artefacts are less likely to grasp the user's attention and consequently guide it, while larger artefacts are more likely to be noticed by the viewer. Similarly, artefacts that are present longer throughout the video are more likely to have a more significant effect on guiding attention, as compared to shorter presences. While more classes and criteria could be devised, the size and duration were binned in only three distinct sizes and intervals to ensure simplicity and interpretability of the score.

`Visual Size:`

- `Small:` an artefact that occupies less than 5% of the equirectangular frame
- `Medium:` an artefact that occupies between 5% and 15% of the equirectangular frame
- `Large:` an artefact that occupies more than 15% of the equirectangular frame

`Duration:`

- `Short:` an artefact that is present in the video between 1 and 2 seconds
- `Medium:` an artefact that is present in the video between 2 and 10 seconds
- `Long:` an artefact that is present in the video for more than 10 seconds

Each identified visual artefact in the 360-degree video was assigned to one class in each of the dimensions, based on size and duration. The combination of the two dimensions for each artefact resulted in a Diegetic Artefact Score δ, which is a composite measure of the visual size and presence duration of each artefact across the 360-degree video. The δ-value was computed by assigning point-values to each class in both dimensions:

- `Size: Small = 1 point, Medium = 2 points, Large = 3 points`
- `Duration: Short = 1 point, Medium = 2 points, Long = 3 points`

The individual δ-value of each artefact is the product of the size and duration point-values of said artefact, as detailed in Table 9. The total δ for each 360-degree video was acquired by summing the individual δ-values of all present artefacts. A manual annotation approach was used focusing solely on diegetic and visual artefacts that are identifiable from the default POV and which movement within the virtual space diverges from the default camera trajectory. Consequently, each artefact was manually annotated using Python, of which the script can be found in Appendix B9.

Artefact ID	Artefact Label	Artefact Start Time	Artefact End Time	Artefact Size	Duration Score	δ
A1_1	grazing_lion	1.0	60.0	l	3	9
A1_2	drinking_lion	1.0	60.0	s	3	3
A1_3	resting_lion	1.0	60.0	s	3	3
A1_4	bird_flying	27.0	60.0	s	3	3
A2_1	passing_ski1	3.0	6.0	s	2	2
A2_2	passing_ski2	3.0	6.0	s	2	2
A2_3	slalom_ski	18.0	26.0	s	2	2
A2_4	falling_ski	21.0	28.0	s	2	2
A2_5	trees_scenery	29.0	56.0	l	3	9
A2_6	passing_ski3	35.0	44.0	m	2	4
B1_1	block_green	1.0	2.5	s	1	1
B1_2	block_darkblue	4.0	8.5	m	2	4
B1_3	block_yellow	10.0	13.0	m	2	4
B1_4	block_white	18.0	20.0	s	1	1
B1_5	block_red	23.5	27.0	s	2	2
B1_6	block_turqouise	29.0	33.0	s	2	2
B1_7	block_gray	35.0	36.0	s	1	1
B1_8	block_purple	39.0	43.0	m	2	4
B1_9	block_violet	46.5	48.0	s	1	1
B1_10	block_darkred	49.0	52.0	l	2	6
B1_11	block_white2	57.0	59.0	m	1	2
B2_1	police_officer	1.0	60.0	s	3	3
B2_2	stationary_wagons	1.0	60.0	m	3	6
B2_3	second_player	41.5	53.0	s	3	3
B2_4	passing_scenery	1.0	60.0	m	3	6
C1_1	staff_member	1.0	6.5	m	2	4
C1_2	rock_scenery	1.0	47.0	l	3	9
C1_3	trees_scenery	15.0	29.0	m	3	6
C1_4	bridge_scenery	36.0	41.0	s	2	2
C1_5	animatronic1	54.0	57.0	m	2	4
C1_6	animatronic2	58.0	60.0	m	1	2
C2_1	ferris_wheel	0.0	10.5	m	3	6
C2_2	spinning_attraction	1.0	8.0	s	2	2
C2_3	scaffolding_loop	10.0	13.0	l	2	6
C2_4	scaffolding_white	39.0	45.0	l	2	6
C2_5	decor1	48.0	49.0	s	1	1
C2_6	decor2	16.0	18.0	s	1	1
C2_7	ferris_wheel	52.0	60.0	m	2	4

Table 9: Coded diegetic attention guiding artefacts.

A1 Content

The 360-degree video A1 displays a scenic sunny landscape, set in a savanna grassland. Presumably, it contains several elements that appear to be in the continent of Africa. Throughout the video, several Lions can be seen grazing, drinking and walking around. The most notable animal in the scene is a curious lion, which slowly approaches the camera from afar and ends up inspecting the camera up close. In the distance, a lion can be seen resting while another one grazes. Lastly, a bird flies into frame through the blue sky. The camera is positioned at eye-level of the observer and remains stationary throughout the video. The implementation of the established coding scheme on video A1 resulted in a total point-value of $\delta = 18$. A set of three diegetic artefacts present in video A1 are displayed in Figure 15.

(a) grazing_lion (b) resting_lion (c) bird_flying

Figure 15: Set of three diegetic artefacts identified in video A1.

A2 Content

The 360-degree video A2 presents a scenic landscape, set in a snowy mountain range. As such, the observer is a skier that is skiing of the mountain. Along the way, you are passing other skiers. One of which is slaloming in front of you, and another one falls right in front of you. At the beginning, the piste is very wide and open. As you ski down the piste, the trail becomes more narrow with dense trees along both sides. Lastly, another skier is passed. The camera is positioned at eye-level of the observer and consistently remains aimed in line with the movement of the camera. The camera is moving fast throughout the snowy environment, due to the speed of the skier. The implementation of the established coding scheme on video A2 resulted in a total point-value of $\delta = 21$. A set of three diegetic artefacts present in video A2 are displayed in Figure 16.

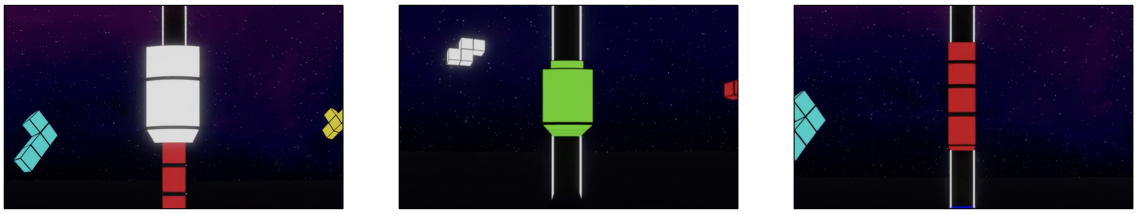
(a) slalom_ski (b) falling_ski (c) trees_scenery

Figure 16: Set of three diegetic artefacts identified in video A2.

B1 Content

In the 360-degree video B1, a digital game of Tetris is displayed. The environment was completely digitally rendered, and as such, the field of play is positioned in a digitally rendered black space. Notably, the observer is positioned within the field of play. Consequently, the brightly coloured Tetris-shaped blocks fall from above, on and around the observer. The camera moves around the x-axis of the field of play to dodge any falling blocks and on the y-axis to stay on top of the fallen blocks, but the camera remains stationary for most of the duration of the video. The

implementation of the established coding scheme on video B1 resulted in a total point-value of $\delta = 28$. A set of three diegetic artefacts present in video B1 are displayed in Figure 17.

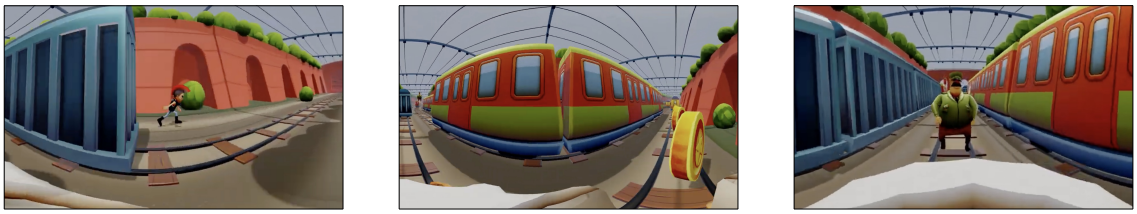


(a) block_white2 (b) block_green (c) block_red

Figure 17: Set of three diegetic artefacts identified in video B1.

B2 Content

The 360-degree video B2 is a digital rendering of the game Subway Surfers, developed by SYBO Games and Kiloo Studios. Similar to B2, the environment was completely digitally rendered. Throughout the video, the observer acts as the player in the game, who's main objective is to out-run the chasing police officer. The observer moves along three subway- and train tracks, while jumping over or dodging under a variety of obstacles. Coins and miscellaneous items are collected throughout, while a tertiary character can be seen also running away from the police officer halfway through. The scenery changes rapidly, as the observer passes many stationary and moving trains. The camera moves fast through the digital environment due to the speed of the running player. The implementation of the established coding scheme on video B2 resulted in a total point-value of $\delta = 18$. A set of three diegetic artefacts present in video B2 are displayed in Figure 18.



(a) second_player (b) stationary_wagons (c) police_officer

Figure 18: Set of three diegetic artefacts identified in video B2.

C1 Content

The 360-degree video C1 displays a scenic car ride through a landscape of canyons. The car is an attraction-specific vehicle, which contains several people. The attraction is located in a Californian theme park. The video starts of with the cast member starting the ride. Throughout the video, the observer stays seated in the vehicle as it moves slowly along the canyon and landscapes. Rock formations, trees and a waterfall are passed while riding the attraction. At last, a dark tunnel is entered in which moving animatronic vehicles light up and move around. The camera is positioned at eye-level and the observer moves with a slow-moving pace due to the speed of the vehicle. The implementation of the established coding scheme on video C1 resulted in a total point-value of $\delta = 27$. A set of three diegetic artefacts present in video C1 are displayed in Figure 19.

C2 Content

In the 360-degree video C2, the observer is seated in a roller coaster. The roller coaster is located in a Californian theme park. The observer is seated between two passengers. The roller coaster

(a) rock_scenery

(b) bridge_scenery

(c) animatronic1

Figure 19: Set of three diegetic artefacts identified in video C1.

moves with a fast pace through a variety of twists, turns, tunnels and a loop. Throughout the ride, miscellaneous objects and elements (i.e., scaffolding and palm trees) can be seen, placed outside the ride. The roller coaster passes a Ferris wheel, as well as other attractions in the area. The sky is vivid and prominent, as the video takes place during a sunset. The camera is positioned at eye-level. The observer moves in a fast pace throughout the video, due to the intensity of the roller coaster. The implementation of the established coding scheme on video C2 resulted in a point-value of $\delta = 26$. A set of three diegetic artefacts present in video C2 are displayed in Figure 20.

(a) ferris_wheel

(b) spinning_attraction

(c) decor2

Figure 20: Set of three diegetic artefacts identified in video C2.

4.2 Dynamics of δ and N_ψ

The resulting δ point-value for each of the utilised 360-degree video denotes the degree of visual artefacts that guide user attention. Specifically, the resulting δ was constructed based on size and duration of each of the diegetic and attentional guiding artefacts. A higher value suggests a higher density of visual elements prone to grasp user attention. An initial data exploration was done on the association between the resulting δ point-values and degree of exploration, to assess the complexity of the association of δ and N_ψ. This data exploration treats the δ point-value as the independent variable, and N_ψ as the dependent variable to explore how changes in N_ψ can be explained by unit increments of δ.

A dual-axis plot was constructed to provide an initial indication of the relationship between the presence of diegetic artefacts and the difference in degree of gaze exploration across the 360-degree videos. The histogram (primary y-axis) represents the total δ-values for each video, while the scatter plot (secondary y-axis) illustrates the variation in the quadrifactorial exploration index across the videos. To ensure simplicity and interpretability of the plot, only the N_ψ-values of the aggregate heatmaps from Figure A.1 (found in Appendix A2) were included. The aggregate heatmaps provide a sufficient initial interpretation of any association between δ- and N_ψ-values, as they provide an initial indication of the general degree of gaze exploration relative to height of the δ point-values across the videos. Utilising the heights within the histogram and N_ψ data points, an interesting pattern emerges. As can be seen in videos A1, B1, B2, C1 and C2, a positive association can be identified: a higher δ point-value is accompanied by a higher degree of gaze distribution N_ψ. However, video A2 deviates from this general trend. Despite the associated high

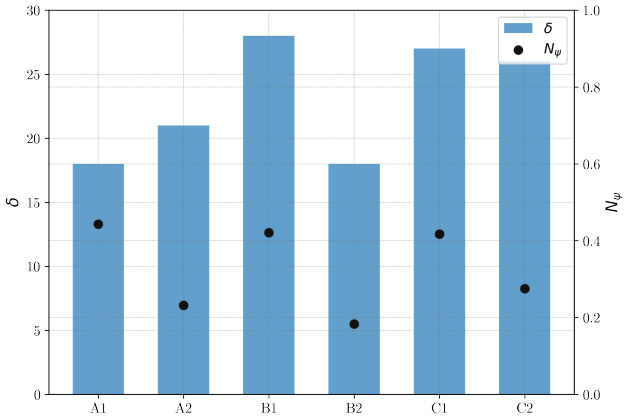


Figure 21: Dual-axis plot of the δ- and N_ψ-values across the 360-degree videos.

δ point-value, N_ψ remains significantly lower than expected. This interpretation is supported by calculating the ratios of δ to N_ψ:

$$\text{Ratio for A1} = 18/.443 \approx 40.6, \text{ for A2} = 21/.232 \approx 90.5.$$
$$\text{Ratio for B1} = 28/.421 \approx 66.5, \text{ for B2} = 18/.183 \approx 98.4.$$
$$\text{Ratio for C1} = 27/.417 \approx 64.7, \text{ for C2} = 26/.275 \approx 94.5.$$

The ratios provide an indication of the relative magnitude of the δ to N_ψ-values. For each of the videos A1, B1, B2, C1, and C2, the δ point-value is x times larger than the respective N_ψ-values, suggesting the general positive association between the δ- and N_ψ-values. For A1, the δ point-value is approximately 40.6 times greater than N_ψ. Similarly, the ratios increase to approximately 66.5, 98.4, 64,7 and 94.5 for B1, B2, C1 and C2, respectively. This positive increase further implies the general positive association between δ- and N_ψ-values. However, as evident in Figure 21, A2 contains a disproportionately low N_ψ-value, despite having a similar δ-score as A1. The extremely high ratio of A2, comparable to B2 and C2, is not accompanied by a similarly high δ point-value. For A2, despite a slight increase in δ, the corresponding increase in N_ψ remains disproportionate. Consequently, this discrepancy suggests that the general positive association between δ- and N_ψ-values is not strictly linear, implying a more complex relationship between the variables. This characteristic of A2 further indicates an underlying non-linear relationship between δ and N_ψ.

A correlation analysis was performed on the variables δ and N_ψ, utilising the resulting dataset from the experiment including $n = 52$ participants, as described in § 2.7.2. To ensure reliability of the models additional data points were not extrapolated. A Shapiro-Wilk test was conducted to assess the normality assumption prior to the correlation analysis. The resulting p-value .0593 is slightly above the significance threshold $\alpha = .05$. As such, the null hypothesis was not rejected. However, the proximity of the p-value to α doesn't strongly support that the data is normally distributed. Separate Shapiro-Wilk tests on the distribution of δ and N_ψ evince that both variables are significantly non-normal, with both p-values $< .01$. Considering the violation of normality, a non-parametric Spearman's rank-order correlation was conducted to examine the relationship between δ point-values and N_ψ. There was a weak, positive correlation between presence of diegetic attention guiding visual artefacts and degree of gaze exploration, which was statistically significant ($\rho(310) = .153$, p-value $= .007$). These findings suggest that while the presence of diegetic artefacts δ increases, the degree of gaze exploration N_ψ increases as well – though not necessarily at a constant or reliable rate. When violating the normality assumption, a Pearson's correlation further emphasises the unreliable linearity of the relationship ($r = .10$, p-value $> .05$).

The weak, positive relationship between δ and N_ψ and the violation of normality indicate that non-linear models could more accurately capture the complexity of the two variables. The non-linear nature of the relationship and behaviour of δ and N_ψ was examined by utilising the reciprocal-ratios, based on the previously calculated ratio-values:

$$\text{Reciprocal for A1} = 1/40.6 \approx .0246, \text{ for A2} = 1/90.5 \approx .0110.$$
$$\text{Reciprocal for B1} = 1/66.5 \approx .0150, \text{ for B2} = 1/98.4 \approx .0102.$$
$$\text{Reciprocal for C1} = 1/64.7 \approx .0155, \text{ for C2} = 1/94.5 \approx .0106.$$

The reciprocal ratio represents how N_ψ changes for a given increase in δ. For A2, the reciprocal value of .0110 indicates a mere 1.1% increase in N_ψ for a one unit increase in δ point-value. Essentially, in the case of A2, N_ψ is not increasing at a similar rate relative to the other videos. This anomaly suggest a different relationship between δ and N_ψ for A2 compared to the others, as previously implied.

Utilising the reciprocals of the other videos enabled a better understanding of the complex relationship of δ and N_ψ. For A1, utilising the ratio (40.6) and reciprocal (.0246), a unit increase in δ point-value results in an N_ψ increase of approximately 2.46%. For B1 and C1, containing higher ratios (66.5 and 64.7) than A1, N_ψ only increases by approximately 1.5%, implying a less proportionally growth. This is further evidenced by the reciprocals of B2 (.0102) and C2 (.0106) associated with significantly higher ratios of 98.4 and 94.5, respectively. As such, for B2 and C2, a unit increase in δ only results in approximate N_ψ increase of 1.0%, indicating an even less proportional growth in N_ψ compared to A1, B1 and C1.

The use of ratio- and reciprocal-values, as well as the findings from the Spearman's rank-order correlation on δ and N_ψ, indicate an intricate and complex relationship between the variables. The weak, positive linear relationship, as suggested by Spearman's rank-order correlation, was not strongly supported by the violation of normality and $\rho = .153$ (df = 310). Despite the p-value $< .05$, the reciprocal ratios indicate a non-linear relationship. As such, alternative non-linear regressions were performed to further explore the complex relationship between presence of diegetic attention guiding artefacts (δ) and degree of gaze exploration (N_ψ).

4.3 Non-Linear Regression Analyses of δ and N_ψ

The reciprocal-values indicate that for A1, B1, B2, C1 and C2, as δ point-value increases, the growth of in which N_ψ gradually decreases. As such, the δ- and N_ψ-values are positively associated, but contain a disproportionate growth. This general trend strongly resembles the behaviour of a logarithmic relationship in the form of:

$$f(x) = a \cdot \log_b(x) + c$$

By substituting $f(x)$ and x by N_ψ and δ, the logarithmic relationship can be modelled as:

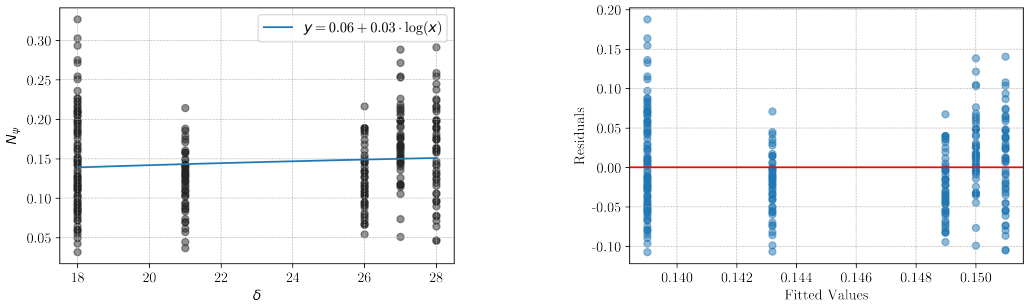
$$N_\psi = a \cdot \log_b(\delta) + c \tag{35}$$

Constants a and b determine the curvature and c adjusts the curve along the y-axis.

A logarithmic regression was run to determine the goodness of fit of a logarithmic relationship between the presence of diegetic attention guiding artefacts (δ) and the degree of gaze exploration (N_ψ) across the sampled $n = 52$ users. The resulting logarithmic regression model, expressed as

$$N_\psi = 0.0607 + 0.0271 \cdot \log(\delta) \tag{36}$$

accounts for approximately 0.9% of the variance in N_ψ ($R^2 = .009$). The F-statistic of 2.707 was not statistically significant (p-value = .101). Furthermore, the coefficient for $\log(\delta) = .0271$ was also not statistically significant (t(310) = 1.645, p = .101, 95% CI $[-.005, .060]$), suggesting



(a) Logarithmic Regression

(b) Fitted Values vs. Residuals Plot

Figure 22: Logarithmic model of δ and N_ψ.

Model Summary						
R^2	**Adj.** R^2	**F-statistic**	**Prob. (F-statistic)**	**MSE**	**AIC**	**BIC**
.009	.005	2.707	.101	.0029	-934.8	-927.3

Coefficients					
	Coefficient	**Std Error**	**t-Statistic**	**p-value**	**95% CI**
Constant	.0607	.051	1.180	.239	$[-.041, .162]$
x_1 (log(δ))	.0271	.016	1.645	.101	$[-.005, .060]$

Table 10: Logarithmic model statistics.

that the logarithmic model does not significantly explain the relationship between δ and N_ψ. The logarithmic model statistics are presented in Table 10.

As evident in Figure 22a, the data points generally adhere to the logarithmic curve. The shallow curve is consistent with the low R^2 of the model. No data points intercept with $f(x)$, suggesting that the logarithmic model may not adequately capture the complex relationship. Similarly, the fitted values vs. residuals plot in Figure 22b shows that the logarithmic model underestimates the N_ψ-values for lower δ point-values, evidenced by the spread of positive residuals between A2, B2 and C2. Furthermore, the residual plot seems to overestimate N_ψ-values for A1, B1 and C1, visible by the relatively even spread of negative residuals. While the data points generally follow the logarithmic curve, the regression and residual plots indicate that the relationship between δ point-value and N_ψ might not be purely logarithmic.

A polynomial regression, as a non-linear alternative, was employed to examine whether relationship between δ and N_ψ might be better approximated by a polynomial function rather than a logarithmic function:

$$p(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \ldots + \beta_n x^n + \epsilon$$

By substituting $p(x)$ and x by N_ψ and δ, the polynomial relationship can be modelled as:

$$N_\psi = \beta_0 + \beta_1 \cdot \delta + \beta_2 \cdot \delta^2 + \beta_3 \cdot \delta^3 + \ldots + \beta_n \cdot \delta^n + \epsilon \tag{37}$$

where the β-coefficients determine the parabola of the polynomial, and ϵ represents the error term.

By conducting multiple polynomial regressions, varying degrees of n could be compared. As evident in Figure 23, a quadratic polynomial regression ($n = 2$) captures the relationship between δ and N_ψ more accurately, as the data points are more closely aligned with the quadratic curve compared to the logarithmic model. However, no data points intercept with the single quadratic curve. The quadratic polynomial is expressed as:

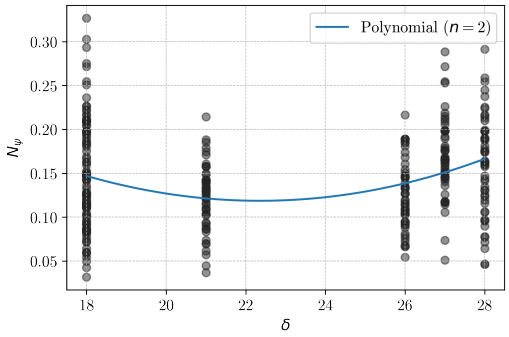


Figure 23: Quadratic Polynomial

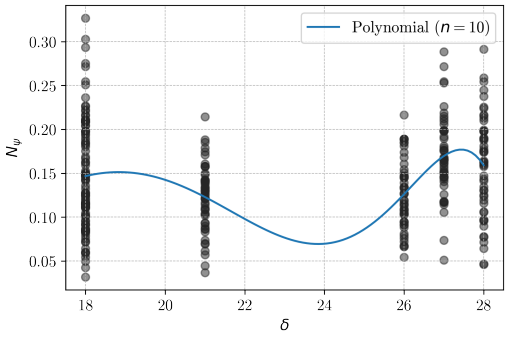


Figure 24: 10th-Degree Polynomial

$$N_\psi = \beta_0 + \beta_1 \cdot \delta + \beta_2 \cdot \delta^2 + \epsilon \tag{38}$$

A quadratic polynomial regression was run to assess the goodness of fit of a 2nd-degree polynomial model on the relationship between δ and N_ψ. The resulting quadratic polynomial model, expressed as

$$N_\psi = 0.87 - 0.07 \cdot \delta + 0.0015 \cdot \delta^2 \tag{39}$$

accounts for approximately 6.3% of the variance in N_ψ ($R^2 = .063$). The linear and quadratic parameters are statistically significant with coefficients -.0669 (t(310) = -4.069, p < .001, 95% CI [-.099, -.035]) and .0015 (t(310) = 4.155, p < .001, 95% CI [.001, .002]). The F-statistic (10.39) was statistically significant (p-value < .001), suggesting that the quadratic polynomial model significantly explains the relationship between δ and N_ψ. The quadratic polynomial model statistics are presented in Table 11.

Model Summary						
R^2	**Adj.** R^2	**F-statistic**	**Prob. (F-statistic)**	**MSE**	**AIC**	**BIC**
.063	.057	10.39^{***}	4.31e-05	.0027	-950.3	-939.1
Coefficients						
	Coefficient	**Std Error**	**t-Statistic**	**p-value**	**95% CI**	
Constant	$.8665^{***}$	.182	4.766	.000	[.509, 1.224]	
x_1 (δ)	$-.0669^{***}$	.016	-4.069	.000	[−.099, −.035]	
x_2 (δ^2)	$.0015^{***}$	.000	4.155	.000	[.001, .002]	

(*) p-value < .05, (**) p-value < .01, (***) p-value < .001

Table 11: Quadratic polynomial model statistics.

Contrary to the quadratic polynomial model, a higher-degree polynomial model ($n = 10$) contains multiple data points intercepting the curve. As presented in Figure 24, a 10th-degree polynomial model contains several data points intercepting with the curve, capturing the relationship between δ and N_ψ even more accurately. While a 10th-degree polynomial better fits the data points, the inclusion of three curves indicates that the model is overfitting, capturing noise and reducing performance on new data points. The 10th-degree polynomial is expressed as:

$$N_\psi = \beta_0 + \beta_1 \cdot \delta + \beta_2 \cdot \delta^2 + \beta_3 \cdot \delta^3 + \ldots + \beta_{10} \cdot \delta^{10} + \epsilon \tag{40}$$

A 10th-degree polynomial regression was run to assess the goodness of fit of a higher-degree polynomial model on the relationship between δ and N_ψ. The resulting 10th-degree polynomial model, expressed as

$$N_\psi = -1.78 \times 10^{-13} - (2.43 \times 10^{-12}) \cdot \delta - (3.08 \times 10^{-11}) \cdot \delta^2 + \ldots - (9.21 \times 10^{-13}) \cdot \delta^{10} \quad (41)$$

accounts for approximately 9.5% of the variance in N_ψ ($R^2 = .095$). The F-statistic (8.049) was statistically significant (p-value $< .001$), indicating that the 10th-degree polynomial model as a whole significantly explains the relationship between δ and N_ψ. However, none of the individual coefficients were statistically significant at $\alpha = .05$. Notably, the higher-degree polynomial parameters approach significance with coefficients for δ^9 being < 0.01 (t(307) = 1.738, p = .083, 95% CI [$< .01$, $< .01$]) and for δ^{10} being $< .01$ (t(307) = -1.912, p = .057, 95% CI [$< .01$, $< .01$]). The 10th-degree polynomial model statistics are presented in Table 12.

Model Summary						
R^2	**Adj. R^2**	**F-statistic**	**Prob. (F-statistic)**	**MSE**	**AIC**	**BIC**
.095	.083	8.049***	3.50e-06	.0026	-957.2	-938.4
Coefficients						
	Coefficient	**Std Error**	**t-Statistic**	**p-value**	**95% CI**	
Constant	-1.784e-13	1.94e-13	-.920	.359	[-5.6e-13, 2.03e-13]	
x_1 (δ)	-2.427e-12	2.64e-12	-.919	.359	[-7.62e-12, 2.77e-12]	
⋮	⋮	⋮	⋮	⋮	⋮	
x_9 (δ^9)	7.848e-11	4.52e-11	1.738	.083	[-1.04e-11, 1.67e-10]	
x_{10} (δ^{10})	-9.21e-13	4.82e-13	-1.912	.057	[-1.87e-12, 2.7e-14]	

(*) p-value $< .05$, (**) p-value $< .01$, (***) p-value $< .001$

Table 12: 10th-degree polynomial regression model statistics.

The initial data exploration in § 4.2 identified A2 as an anomaly, deviating from the general trend. Therefore, it was decided to run the logarithmic and polynomial regressions on the full dataset as well as the dataset excluding A2-data. However, excluding A2 yielded only marginal differences. For the logarithmic and both polynomial regressions, excluding A2 from the dataset resulted in a decrease in R^2, indicating an even weaker fit to the data: $R^2 = .004$, $R^2 = .033$, $R^2 = .069$, respectively. In the 10th-degree polynomial regression, more significant polynomial coefficients were present, which could potentially be attributed to overfitting. Moreover, the majority of coefficients and statistical significance levels remained largely consistent with those in the models that included A2. These findings suggest that excluding A2 from the data does not significantly improve the models and as such, the regressions were run utilising the dataset in its entirety.

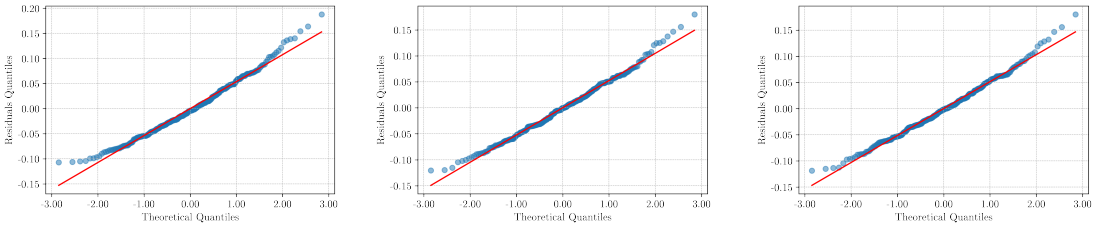
Assumptions The assumptions of normality, independence of errors, and homoscedasticity for the logarithmic, quadratic, and 10th-degree polynomial regression models were assessed.

The assumption of normality was assessed using Q-Q plots. Evident in Figure 25, the Q-Q plots of all three models align for the majority with a straight line. Slight deviation from the reference line can be seen in both tail-ends across the Q-Q plots. This observation suggests that the residuals for all three models are approximately normally distributed. A Shapiro-Wilk test was used to confirm the normality in both polynomial regressions. The Shapiro-Wilk test statistic .991 (p-value = .064) for the quadratic polynomial model suggests that the residuals are not significantly deviated from a normal distribution. The 10th-degree polynomial produced similar results (p-value = .059)

Durbin-Watson tests were performed to assess the independence of errors in the residuals. The logarithmic, quadratic and 10th-degree polynomial regressions resulted in $1.702, 1.418$ and 1.270, respectively. The proximity of these results to 2 indicates that the assumption of independence of error was satisfied.

Homoscedasticity was assessed by conducting the Breusch-Pagan test. For all three regressions, the assumption of homoscedasticity was violated (p-value $< .05$), indicating the presence of heteroscedasticity in the regressions.

The logarithmic, quadratic and 10th-degree polynomial regressions meet the assumptions of normality and independence of errors. However, the presence of heteroscedasticity suggests caution in the interpretation of the results.



(a) Logarithmic (b) Polynomial ($n = 2$) (c) Polynomial ($n = 10$)

Figure 25: Q-Q Plots.

4.4 Interpretation of Regression Models

This diegetic assessment provides an initial indication on the nature of the relationship between presence of diegetic artefacts in 360-degree videos and degree of gaze distribution. Utilising logarithmic, quadratic polynomial and 10th-degree polynomial models, as well as reciprocate ratios, enabled a better understanding of the underlying data-structure. A set of regression analyses were conducted to further assess the proposed non-linear models goodness of fit, statistical significance and examining the complexity of the relationship between δ and N_ψ. The utilisation of non-linear regression models was motivated by the preliminary results from Spearman's rank-order correlation $\rho = .153$ (df $= 310$). Coupled with the violation of assumptions initiated the use of non-linear models to capture the relationship between δ and N_ψ.

A total of three non-linear regressions were conducted, using δ as a predictor. The logarithmic, quadratic polynomial and 10th-degree polynomial were assessed in their effectiveness in predicting N_ψ using R-squared and the significance of coefficients. The logarithmic model displayed non-significant coefficients and a relatively low $R^2 = .009$ (p-value $= .101$). These findings suggest that the logarithmic model does not explain much of the variation in N_ψ and indicates a much more complex association between δ and N_ψ. Consequently, polynomial models of varying degrees were employed to better capture the complex relationship. Firstly, the quadratic polynomial model resulted in a much higher $R^2 = .063$ (p-value $< .001$), explaining approximately 6.3% of the variance in N_ψ. Secondly, a 10th-degree polynomial model was utilised. The model showed the highest R-squared ($R^2 = .095$, p-value $< .001$) across the three non-linear models. However, none of the coefficients were statistically significant. Notably, the higher-order coefficients x^9 and x^{10} approach statistical significance with p-value $= .083$ and p-value $= .057$, respectively. This indicates that higher-order complexities may capture more of the variance in N_ψ, but it elicits the risk of overfitting, emphasising the complexity of the relationship. As such, the quadratic polynomial regression provides a more accurate representation of the relationship as compared to the logarithmic model without overfitting as the 10th-degree polynomial model. This is further supported by the lower Mean Squared Error ($= .0027$), as compared to the logarithmic model (MSE $= .0029$), suggesting that the predicted values are on average closer to the actual values in the quadratic polynomial model. Despite A2 deviating further from the general data trend, exclusion of subset A2 in the regressions did not substantially impact the model's performances, implying that A2 is not an outlier in the dataset. This observation that a higher δ does not invariably correspond to a higher degree of gaze exploration, emphasises the complex interaction dynamics between user and 360-degree videos. As such, changes in N_ψ can be attributed to factors

not accounted for, such as cognitive perceptions and usability context. Despite the violation of homoscedasticity, the diegetic assessment was exploratory in nature, aiming to gain a better understanding into the underlying association between presence of diegetic artefacts and degree of gaze exploration. While it is imperative to approach these results critically, the results can be considered as a preliminary step in understanding the complex relationship between the presence of diegetic artefacts δ and the degree of gaze distribution N_ψ in 360-degree video interactions.

The implications from the linear and non-linear models suggest a complex relationship between δ and N_ψ. As such, the presence of diegetic artefacts within the 360-degree video content (δ) was considered a confounding variable in the primary analyses of spatiotemporal image complexity on gaze distribution N_ψ, presented in Chapter 5.

Chapter 5

Results

The research methodology, as detailed in Chapter 2, employed an eye-tracking study and subsequent user evaluation to acquire both quantitative and qualitative data. A systematic analytical framework was devised as a mechanism to elucidate the main research objective and related sub-questions, defined in section 1.7. The analytical framework in section 2.7.3 established the following analytical objectives:

- I: Relationship between the perceptual attributes and gaze exploration;
- II: Determining the influence of spatiotemporal 360-degree video complexity on gaze exploration;
- III: Assessing the interaction effect of usability context;
- IV: Defining the general consensus on the experiential statements;
- V: Identifying the key trends in the user's self-perception of gaze behaviour.

The results from the quantitative and qualitative data analyses are presented in this section.

Firstly, the descriptive statistics are presented in section 5.1. The descriptive statistics provide an overview of the central tendencies and distributions of the independent and dependent variables used, as well as enables insight into the data-characteristics. Section 5.2.1 details the use of a linear mixed-effects analysis to approach objective I: assessment of the effect of each of perceptual attributes on degree of gaze exploration. The perceptual attributes with a resulting significant effect were considered as confounding variables in subsequent analyses. In section 5.2.2, a mixed-effects multiple regression was employed to approach objective II: examine how the varying degrees of spatial- and temporal image complexity in 360-degree videos impact the degree of gaze exploration. The mixed-effects model accounts for the presence of diegetic artefacts in each of the videos (δ) and controls for the significant perceptual attributes from section 5.2.1, as well as for the repeated measures design by including random effects. The model includes the confounding variables to further isolate the behavioural effect as induced by changes in spatiotemporal image complexity. Subsequently, section 5.2.3 focuses on objective III, in which a subgroup and moderation analysis were performed to assesses how the impact of spatiotemporal image complexity on the degree of gaze distribution might vary across different usability contexts. The analyses examine whether the effect of spatiotemporal image complexity on gaze exploration depends on seating type, providing insights into how seating type interacts with the spatiotemporal image complexity to influence gaze behaviour. Furthermore, section 5.2.4 encompasses objective IV by assessing the general consensus on the experiential statements from the user evaluation. A combination of both descriptive statistics as well as a non-parametric comparative analyses were performed. The descriptive statistics were used to assess the central tendencies and data distribution of ratings for each of the statements, while a series of Mann-Whitney U tests were performed to assess the difference in ratings between groups. Lastly, analytical objective V was approached utilising a grounded theory analysis containing various emergent coding procedures, employed to analyse the qualitative data for patterns in the conscious gaze behaviour of the users. The qualitative analysis is presented in section 5.3.

5.1 Descriptive Statistics

This section presents the resulting descriptive statistics for the utilised independent and dependent variables across all users, as well as for the two distinct groups R and F. The acquired dataset was coded in long-format as the results include both single-measure variables as well as multiple-measure variables per user depending on the variables. The perceptual attributes of engagement (x_1), attentional focus (x_2), spatial awareness (x_3), fear of missed content (x_4) were measured once across all users, similar to the usability factors of comfort (x_5) and enjoyment (x_6). These variables, due to the length of their respective terminology, were denoted by variables of x. The descriptive statistics of the single-measure variables are presented in Table 13. The repeated measures of mean opinion score (MOS) and degree of gaze exploration (N_ψ) were measured six times for each 360-degree video, across all users. The descriptive statistics of repeated measures

variables are presented in Table 14. Firstly, the descriptive statistics of the perceptual attributes are presented. Subsequently, the values of the usability factors are discussed. Lastly, the degree of gaze exploration across the six videos and between the groups are detailed.

The distribution of data across the attributes of perception resulted in diverse patterns across the two groups. Most notably, group R demonstrated a higher level of engagement ($\mu = 5.011, \sigma = .305$) as compared to group F ($\mu = 4.462, \sigma = .331$). This trend can be seen for the perceptual attribute of spatial awareness (x_3) as well, observed in group R with $\mu = 5.796, \sigma = .638$ and in group F as $\mu = 4.910, \sigma = .636$. However, group F achieved higher mean values for attentional focus (x_2) and fear of missed content (x_4). The attentional focus (x_2) in group R had averaged at $\mu = 5.885, \sigma = .993$, compared to group F ($\mu = 5.923, \sigma = .845$). Fear of missed content (x_4) produced a higher mean $\mu = 4.923, \sigma = 1.017$, as compared to group R ($\mu = 3.192, \sigma = 1.297$). Notably, fear of missed content (x_4) contained the widest range of data, from 1.000 to 7.000.

Similar to the level of engagement and spatial awareness, the mean scores for quality of experience were higher in group R. Despite the per-video measurements, the overall mean across the users in each group was higher in group R ($\mu = 4.548, \sigma = 1.688$) compared to group F ($\mu = 4.413, \sigma = 1.380$). Moreover, and except for video B1 and B2, the mean per-video MOS were higher in group R than in group F. Video A2 resulted in the highest mean MOS of $\mu = 6.048, \sigma = .788$ across all $n = 52$ users, while video B1 had the lowest mean MOS of $\mu = 2.067, \sigma = .805$. This pattern was also observed among both groups R and F. In conclusion, the data exhibits a dichotomous pattern in the distribution of perceptual attributes among the two groups R and F. While group R produced higher mean values for levels of engagement, spatial awareness, and quality of experience, group F displayed higher mean values for attentional focus and fear of missed content. Attentional focus and level of comfort were relatively high among both groups.

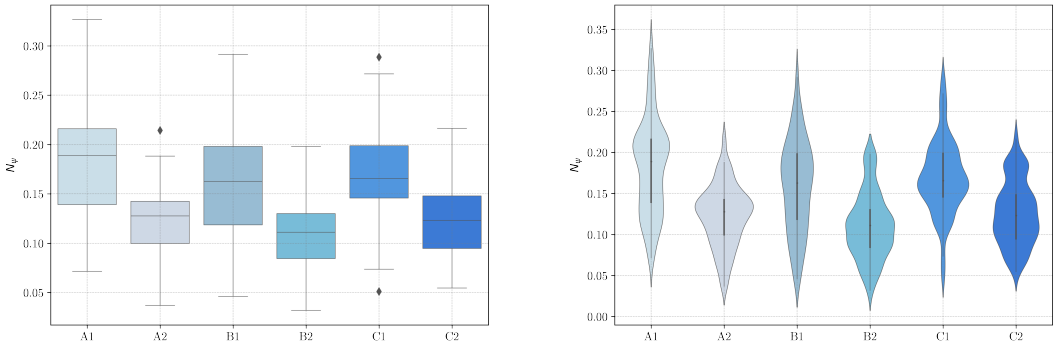
All Users				
	Mean	**Standard Deviation**	**Minimum**	**Maximum**
x_1	4.736	.420	3.714	5.571
x_2	5.904	.913	4.000	7.000
x_3	5.340	.766	3.000	7.000
x_4	4.058	1.447	1.000	7.000
x_5	5.615	1.013	4.000	7.000
x_6	5.000	1.085	3.000	7.000
Group R				
	Mean	**Standard Deviation**	**Minimum**	**Maximum**
x_1	5.011	.305	4.286	5.571
x_2	5.885	.993	4.000	7.000
x_3	5.769	.638	4.667	7.000
x_4	3.192	1.297	1.000	5.000
x_5	5.769	0.992	4.000	7.000
x_6	4.731	0.874	3.000	6.000
Group F				
	Mean	**Standard Deviation**	**Minimum**	**Maximum**
x_1	4.462	.331	3.714	5.000
x_2	5.923	.845	4.000	7.000
x_3	4.910	.636	3.000	6.000
x_4	4.923	1.017	3.000	7.000
x_5	5.462	1.029	4.000	7.000
x_6	5.269	1.218	3.000	7.000

Note: x_1 = engagement, x_2 = attentional focus, x_3 = spatial awareness, x_4 = fear of missed content, x_5 = comfort of seating type, x_6 = enjoyment of seating type

Table 13: Descriptive statistics of the perceptual attributes and usability factors.

The usability factors of assessing the user experience of the specific seating type associated with each group was assessed using the level of comfort (x_5) and enjoyment (x_6) of the seating type. Across all users, the level of comfort (x_5) and enjoyment (x_6) scored high on average, with

$\mu = 5.615, \sigma = 1.013$ and $\mu = 5.000, \sigma = 1.085$, respectively. Group R averaged higher in terms of comfort (x_5) with mean $\mu = 5.769, \sigma = .992$, compared to group F ($\mu = 5.462, \sigma = 1.029$). However, group R averaged higher in terms of enjoyment (x_6) with mean $\mu = 5.269, \sigma = 1.218$, compared to group R ($\mu = 4.731, \sigma = .874$).



(a) Box Plots

(b) Violin Plots

Figure 26: Distributions of N_ψ.

All Users									
	MOS					N_ψ			
Video ID	count	mean	std	min	max	mean	std	min	max
All	312	4.481	1.540	1.000	7.000	.145	.054	.032	.327
A1	52	4.952	1.189	2.500	7.000	.183	.059	.071	.327
A2	52	6.048	.788	4.000	7.000	.123	.038	.037	.214
B1	52	2.067	.805	1.000	4.000	.159	.057	.046	.291
B2	52	4.952	1.072	1.000	7.000	.111	.041	.032	.198
C1	52	4.144	.904	2.500	7.000	.170	.046	.051	.289
C2	52	4.721	.866	2.500	7.000	.126	.039	.054	.216
Group R									
	MOS					N_ψ			
Video ID	count	mean	std	min	max	mean	std	min	max
Group R	156	4.548	1.688	1.000	7.000	.165	.057	.042	.327
A1	26	5.038	1.392	2.500	7.000	.225	.043	.144	.327
A2	26	6.019	.842	4.000	7.000	.132	.033	.070	.214
B1	26	1.865	.867	1.000	4.000	.178	.057	.046	.291
B2	26	5.096	1.312	1.000	7.000	.116	.047	.042	.198
C1	26	4.288	1.124	2.500	7.000	.192	.043	.117	.289
C2	26	4.981	.842	4.000	7.000	.149	.034	.076	.216
Group F									
	MOS					N_ψ			
Video ID	count	mean	std	min	max	mean	std	min	max
Group F	156	4.413	1.380	1.000	7.000	.125	.043	.032	.238
A1	26	4.865	.965	4.000	7.000	.141	.039	.071	.216
A2	26	6.077	.744	5.500	7.000	.114	.041	.037	.188
B1	26	2.269	.696	1.000	4.000	.140	.053	.047	.238
B2	26	4.808	.763	4.000	5.500	.105	.033	.032	.163
C1	26	4.000	.600	2.500	5.500	.147	.036	.051	.199
C2	26	4.462	.824	2.500	5.500	.103	.029	.054	.185

Table 14: Descriptive statistics of MOS- and N_ψ-values.

The degree of gaze distribution was quantified using the quadrifactorial exploration index N_ψ. For video A1, the mean was $\mu = .183$ with a standard deviation of $\sigma = .059$. Video A2 had a lower overall mean $\mu = .123$ and standard deviation $\sigma = .038$. The observations for B1 and B2 were $\mu = .159, \sigma = .057$ and $\mu = .111, \sigma = .041$, respectively. The observations for videos C1 and C2 averaged at $\mu = .170, \sigma = .046$ and $\mu = .126, \sigma = .039$, respectively. Across all users and within group R, video A1 produced the highest degree of gaze exploration while video B2 elicited the lowest degree of gaze exploration across the 360-degree content set. Notably, group F produced contrasting results. Among the users in group F, video C1 and C2 resulted in the highest and lowest degrees of gaze exploration, with $\mu = .147, \sigma = .036$ and $\mu = .103, \sigma = .029$, respectively. However, as evident in the resulting data distributions, group R maintained overall higher mean degrees of gaze exploration across all six videos compared to group F. Most notable when comparing the data distributions of Tables 13 and 14, higher mean-values for each of the perceptual attributes or usability factors did not inherently result in a higher mean degrees of gaze exploration.

The data distribution of degree of gaze exploration N_ψ across each of the 360-degree videos are visualised utilising box- and violin plots, presented in Figure 26. Figure 26 provides a comprehensive visualisation of the overall patterns in the data, utilising a similar colour palette found within Figure 5. The box plots in Figure 26a represent the general distribution, while the violin plots in Figure 26b visualise the data distributions, as well as data clusters and the range of variability across different N_ψ values for each 360-degree video.

5.2 Quantitative Results

A series of analyses were performed on the quantitative set of data, acquired during the study. This subsection presents the quantitative results of the conducted analyses, based on the beforementioned analytical objectives.

5.2.1 Linear Mixed-Effects Model Analysis

A linear mixed-effects model (LMM) was employed to assess the significant associations between each of the perceptual attributes and degree of gaze distribution N_ψ. Utilising fixed and random effects, the model accounts for both single-measure and repeated-measure perceptual attributes. The model showed a significant, positive association between x_3 (spatial awareness) and N_ψ with coefficient = .019 ($z = 3.265$, p-value = .001). A significant, negative association was found between x_5 (quality of experience) and N_ψ with coefficient = -.006 ($z = -3.568$, p-value $< .001$).

Model Summary					
	Coefficient	**Std Error**	**z-Statistic**	**p-value**	**95% CI**
Intercept	.050	.058	.865	.387	$[-.064, .165]$
x_1	.009	.011	.787	.431	$[-.013, .030]$
x_2	-.000	.001	-.326	.745	$[-.002, .001]$
x_3	.019***	.006	3.265	.001	$[.007, .030]$
x_4	-.004	.003	-1.382	.167	$[-.010, .002]$
x_5	-.006***	.002	-3.568	.000	$[-.010, -.003]$

(*) p-value $< .05$, (**) p-value $< .01$, (***) p-value $< .001$

Table 15: Linear mixed-effects model statistics.

With respective p-values of .001 and .000, there was strong evidence to reject the null hypotheses for x_3 and x_5 and accept the following alternative hypotheses:

- H_1: There is a significant relationship between spatial awareness (x_3) and degree of gaze distribution (N_ψ).
- H_1: There is a significant relationship between quality of experience (x_5) and degree of gaze distribution (N_ψ).

In contrast, the coefficients for x_2 (attentional focus) and x_4 (fear of missed content) were $-.000$ and $-.004$, respectively, but the associations with N_ψ were not statistically significant (p-value $> .05$). Furthermore, the coefficient $= .009$ for x_1 (engagement) was statistically non-significant as well (p-value $= .431$). With respective p-values $> .05$, there was no sufficient evidence to reject the null hypotheses for x_1, x_2 and x_4. As such, there was no significant association between these perceptual attributes and degree of gaze distribution N_ψ. The results suggest that among the perceptual attributes, spatial awareness and quality of experience are significant predictors of N_ψ. The attributes of engagement, attentional focus and fear of missed content did not produce significant effects on N_ψ, according to the LMM. Figure 27 presents a visualisation of the coefficients, highlighting the magnitude and directionality of relationship between the perceptual attributes and degree of gaze distribution. The linear mixed-effects model statistics are presented in Table 15.

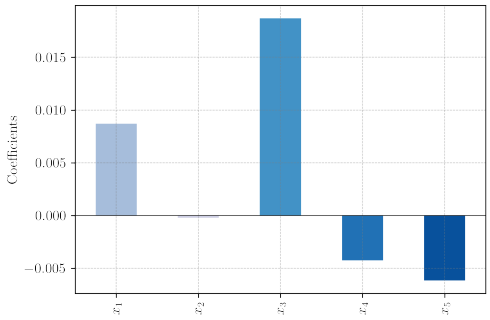


Figure 27: Coefficient plot of perceptual attributes as predictors.

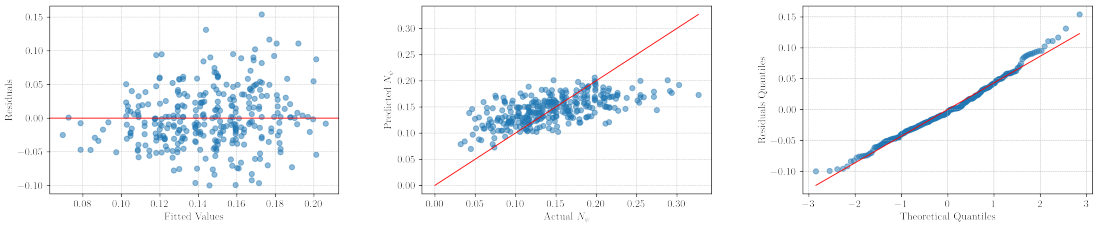
Assumptions The assumptions of linearity, homoscedasticity, normality, independence, and multicollinearity for the linear mixed-effects model were assessed.

The assumptions of linearity and homoscedasticity were assessed using the fitted values vs. residuals plot, presented in Figure 28a. The scatter plot displays a somewhat asymmetrically distributed data pattern. As no distinct patterns could be identified, it was assumed that the relationship between SI, TI and N_ψ is approximately linear. As such, assessment of homoscedasticity was also done utilising the residuals plot. While no distinct patterns can be identified in the scatter plot, its slight asymmetry and subtle funnel-like pattern do not strongly evidence that the assumption of homoscedasticity was met. Homoscedasticity could not be assessed using a Breusch-Pagan test due to the hierarchical structure of the mixed model.

Normality of the residuals was assessed using the Q-Q plot from Figure 28c. Despite a slight deviation from the reference line can be seen in both tail-ends, the observation suggests that the residuals of the linear mixed-effects model are approximately normally distributed. A Durbin-Watson test was conducted to assess the independence of errors in the residuals. The linear mixed-effects model resulted in a value of 2.414, and as such indicates that the assumption of independence was satisfied.

Lastly, the assumption of no multicollinearity was assessed by calculating the Variance Inflation Factors (VIF). A strong intercorrelation between engagement and spatial awareness was found, with VIF-values of 69.49 and 57.84, respectively. The model displayed moderate VIF-values for quality of experience and fear of missed content, suggesting a potential multicollinearity between these predictors as well. These results posit strong evidence that the assumption of multicollinearity was violated.

The linear mixed-effects model's predictions seem to consistently underestimate the actual values, as evident in Figure 28b. This suggest that the associations between the predictors and dependent variable N_ψ might not be adequately modelled by a linear function. Coupled with the slight asymmetry of the scatter plot, suggests cautious interpretation of the results.



(a) Fitted Values vs. Residuals (b) Actual vs. Predicted Values (c) Q-Q

Figure 28: Linear mixed-effects model plots.

5.2.2 Mixed-Effects Multiple Regression Analysis

A mixed-effects multiple regression was run to assess the significance of spatial complexity (SI) and temporal complexity (TI) on the degree of gaze distribution N_ψ, while controlling for δ (presence of diegetic artefacts), x_3 (spatial awareness), x_5 (quality of experience) and $C[T.1]$ (seating type R/F). The model showed that temporal complexity (TI) was a statistically significant predictor of N_ψ, with a p-value $< .001$ ($z = -5.615$). The TI-coefficient $-.003$ indicates a negative association with N_ψ, in which a unit increase in TI decreases N_ψ by .003 units. Consequently, the null hypothesis for the significance of temporal complexity on N_ψ was rejected, accepting the alternative hypothesis:

- H_1: There is a significant relationship between temporal image complexity and degree of gaze distribution (N_ψ).

Spatial awareness (x_3) also exhibited a statistically significant association with N_ψ (coefficient .012, $z = 2.027$, p-value = .043), validating its inclusion as a control variable. Furthermore, the model showed a statistically significant relationship between seating type and degree of gaze exploration N_ψ ($z = 3.418$, p-value = .001). Specifically, seating type R (T.1) increased N_ψ by .030 units compared to seating type F. In § 5.2.3, this mixed-effects multiple regression model is extended with interaction terms to evaluate the association between usability context (i.e., seating type) and degree of gaze exploration across both groups.

In contrast, spatial complexity (SI) did not exhibit a statistically significant association with N_ψ in the mixed-effects multiple regression model (coefficient .000, $z = .006$, p-value = .995). As such, there was no sufficient evidence to reject the null hypothesis for the significance of spatial image complexity on degree of gaze exploration. Moreover, despite the results from § 4.3 and § 5.2.1, δ (presence of diegetic artefacts) and x_5 (quality of experience) did not show a statistically significant association with N_ψ (p-values $> .05$).

Model Summary					
	Coefficient	**Std Error**	**z-Statistic**	**p-value**	**95% CI**
Intercept	.120	.071	1.698	.090	$[-.019, .258]$
$C[T.1]$	.030***	.009	3.418	.001	$[.013, .048]$
SI	.000	.000	.006	.995	$[-.001, .001]$
TI	-.003***	.001	-5.615	.000	$[-.005, -.002]$
x_3	.012*	.006	2.027	.043	$[.000, .023]$
x_5	-.000	.002	-.054	.957	$[-.004, .004]$
δ	-.000	.001	-.249	.803	$[-.003, .002]$

(*) p-value $< .05$, (**) p-value $< .01$, (***) p-value $< .001$

Table 16: Mixed-effects multiple regression model statistics.

The results of the mixed-effects multiple regression model suggest that there is not sufficient evidence to reject the null hypothesis for spatial complexity, despite its theoretical significance. Under the current parameters of the model, which controls for x_3 (spatial awareness), x_5 (quality of

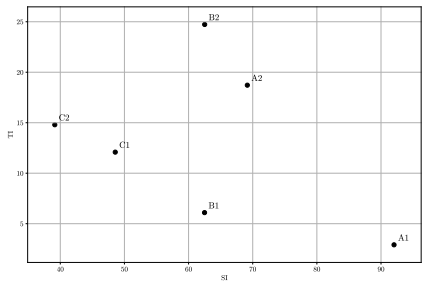


Figure 29: Spatiotemporal Matrix

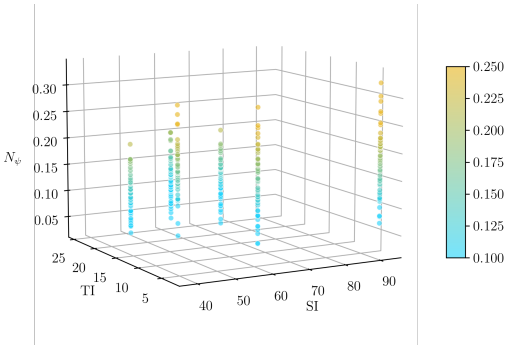
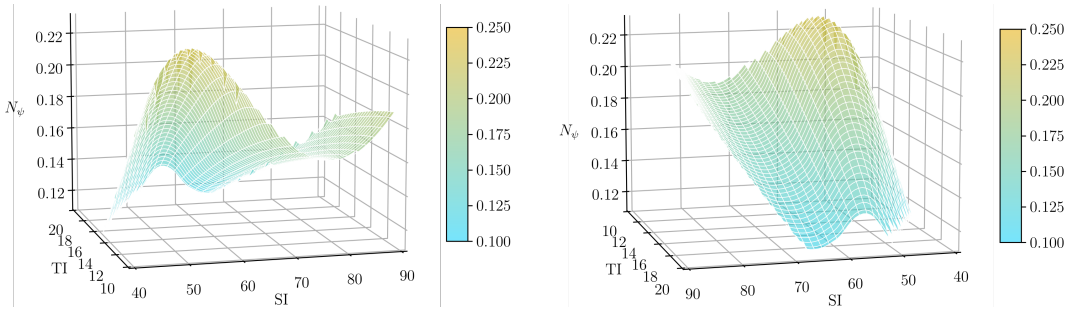


Figure 30: 3D Scatter Plot

experience), δ (presence of diegetic artefacts) and seating type, only temporal complexity exhibited a significant relationship with the degree of gaze distribution. The mixed-effects multiple regression model statistics are presented in Table 16.

The N_ψ-values were visualised as an additional z-dimension to the computed spatiotemporal matrix in § 2.2.4 (see Figure 29), providing an enhanced understanding of the underlying data structure. As such, a three-dimensional scatter plot was constructed where the N_ψ-values were mapped within a three-dimensional space, enabling a spatial perspective on the degree of gaze distribution and respective spatiotemporal complexity coordinates. The 3D scatter plot is presented in Figure 30. The viewing angle of the plot is configured using a 10° elevation and a rotation around the vertical axis by an azimuthal angle of 241°. Due to the dependency of N_ψ on known spatiotemporal image complexities, only six vertical data patterns can be observed. Each data pattern corresponds to each utilised 360-degree video and the respective position on the spatiotemporal matrix.

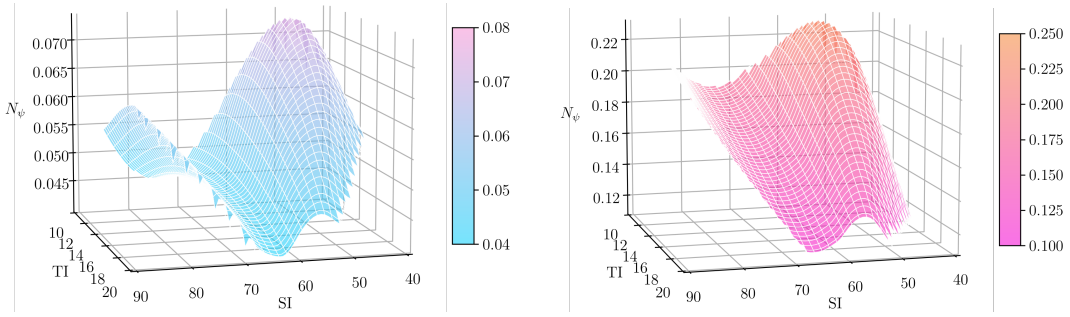
However, the six original data patterns were not sufficient to provide an in-depth representation of the relationships among the three variables. An interpolation was performed on the original data-frame, which interpolates the N_ψ-values at all points on the defined grid. The grid was defined utilising the spatiotemporal matrix containing the same min-max value range for the respective SI- and TI-values. By utilising a cubic interpolation process, a finer mesh was constructed containing additional data points across the grid. The mesh was used to construct a three-dimensional surface plot, representative of how the degree of gaze distribution varies with respect to changes in both spatial- and temporal image complexity simultaneously. The resulting surface plot in Figure 31a is configured with a rotation around the vertical axis by an azimuthal angle of 255°. Figure 31b is rotated by -180° (= azimuthal angle of 75°) along the vertical axis, providing an additional perspective on the surface plot. Figure 31 presents a dual-perspective view, enabling a more detailed visualisation of the multivariate relationship between SI, TI and N_ψ. Similar to the three-dimensional scatter plot, the colour of the surface is determined by the respective N_ψ-values.



(a) 255°

(b) 75°

Figure 31: Surface plots of SI, TI and N_ψ.



(a) Group F

(b) Group R

Figure 32: Surface plots of SI, TI and N_ψ (per group).

The surface plot illustrates a complex association between SI, TI and N_ψ. The curvature of the surface varies across the spatiotemporal plane, indicative of higher-degree non-linear interactions between SI, TI and N_ψ. The surface appears to be sloped at a downwards angle across the temporal plane. The surface displays a combination of both convex and concave shapes across the spatial plane, strongly indicating the intricate and complex relationship between SI, TI and N_ψ. Coupled with the presence of steep slopes, the surface plot suggests that the impact of SI- and TI-values on N_ψ is not consistent throughout the observed data-range. A notable feature includes a peak in N_ψ at a relative low spatiotemporal complexity.

The mixed-effects multiple regression model exhibited a statistically significant relationship between seating type and degree of gaze distribution ($z = 3.418$, p-value = .001). As such, two additional surface plots were constructed to examine the complex multivariate relationship between spatiotemporal complexity on N_ψ across the two usability contexts. The surface plots in Figure 32 were similarly configured, utilising a 10° elevation and a rotation around the vertical axis by an azimuthal angle of 75°. Notably, the subset of data from group R (rotating chair) resulted in a very similar surface plot as the primary surface plot in Figure 31b. This indicates that behaviours and patterns inherent to group R (rotating chair) are, to a large extent, mirrored in group F (fixed chair). This also suggests that the overall data pattern is less significantly influenced by the data from group F. It is important to note that, despite structural similarity in both surface plots, the overall degree of gaze distribution across group F is notably smaller (as indicated by the z-axis). Furthermore, as apparent by the appearance of parallel ridges, key behavioural trends found within group R seem to be true for group F as well. However, within group F, higher levels of spatial- and lower levels of temporal image complexity appear to be curving the surface plot more convex compared to group R.

Furthermore, a series of parallel ridges can be identified in the surface plots. The parallel ridges indicate that for certain configurations of both spatial- and temporal image complexity, repeated patterns in gaze exploration occur. As such, an increase in either spatial- or temporal complexity, while keeping the other constant, could lead to periodic increases or decreases in gaze distribution. This observation implies a certain commonality in the degrees of gaze distribution across the different seating types, despite the structural difference between the two surface plots.

While the cubic interpolation process provides additional data points, it is still limited by the current parameters of the utilised dataset. As such, the surface plot serves as an indicator of the multidimensional correlation between spatiotemporal image complexity and degree of gaze distribution. Additional N_ψ-measurements for alternative 360-degree videos with varying spatiotemporal complexities are vital to the enhancement of the interpolation process.

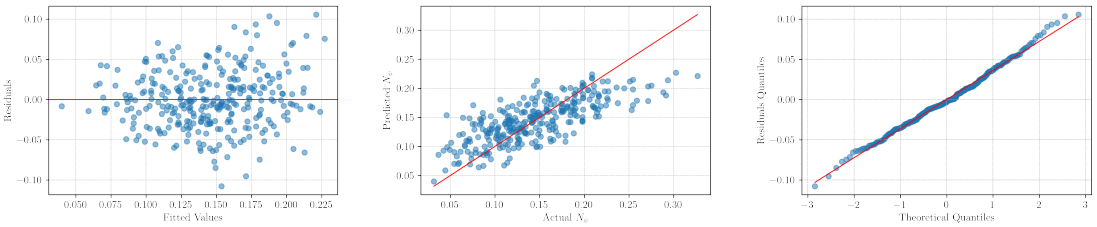
Assumptions The assumptions of linearity, homoscedasticity, normality, independence, and multicollinearity for the mixed-effects multiple regression model were assessed.

By utilising the fitted values vs. residuals plot from Figure 33a, the assumptions of linearity and homoscedasticity were assessed. Despite the absence of clear data pattern, the scatter plot

does display an asymmetrical distribution of data points. The actual vs. predicted values plot (see Figure 33b) further emphasises the cautious assumption of linearity, as the model appears to be slightly underestimating the actual values in its predictions. As such, it was assumed that the model approximates linearity. Moreover, a subtle funnel-shape can be identified in the distribution of data points, indicating the presence of heteroscedasticity. This pattern becomes more apparent when compared to Figure 28a. These findings do not provide sufficient evidence to assume that the homoscedasticity assumption was met. A connection can be made between the approximate linearity of the model and the findings of the diegetic assessment (see § 4.3), in which a complex, non-linear polynomial relationship was identified between δ and N_ψ. As such, the insufficient evidence of linearity could partially be attributed to the inclusion of δ within the mixed-effects model.

The Q-Q plot from Figure 33c was utilised to assess normality of the residuals. The distribution sufficiently aligns with the reference line, strongly evidencing a normal distribution. A Durbin-Watson test was run to assess the independence of errors in the residuals. The resulting value of 2.154 strongly indicates the observation that the assumption was met. Lastly, the presence of multicollinearity was examined using Variance Inflation Factors (VIF). Both SI and δ had VIFs of 7.38 and 7.61, respectively, indicating a moderate degree of multicollinearity. However, the other variables exhibited VIFs below 5, suggesting acceptable levels of multicollinearity.

The insufficient evidence to support homoscedasticity, as well as the slight presence of multicollinearity, suggests caution in the interpretation of these findings. However, the results suggest that further examination in the underlying data structure could be advantageous.



(a) Fitted Values vs. Residuals (b) Actual vs. Predicted Values (c) Q-Q

Figure 33: Mixed-effects multiple regression model plots.

5.2.3 Usability Group Moderation Analysis

The mixed-effects multiple regression model from § 5.2.2 exhibited a statistically significant relationship between the degree of gaze distribution N_ψ and seating type. The underlying mechanisms and interaction of different usability contexts (i.e., seating type) on this relationship was assessed using a combination of subgroup and interaction analyses. A separate subgroup analysis was conducted for each group F and R, which included similar confounding variables and parameters of the mixed-effects model of § 5.2.2. Each model was fitted to the different subgroups.

For group F, a mixed linear model regression was run to assess the significance of spatial complexity (SI) and temporal complexity (TI) on the degree of gaze distribution N_ψ, while controlling for δ (presence of diegetic artefacts), x_3 (spatial awareness), x_5 (quality of experience). The model (scale = .009) exhibited a statistically significant relationship between spatial complexity (SI) and degree of gaze distribution N_ψ, with coefficient = .001 (z = 1.963, p-value = .05). Despite the slightly positive effect of SI, no statistical significance was found among the other predictor variables (p-values > .05). The model statistics of group F are presented in Table 17.

Similarly for group R, a mixed linear model regression was run to assess the significance of spatial complexity (SI) and temporal complexity (TI) on the degree of gaze distribution N_ψ, while also controlling for δ (presence of diegetic artefacts), x_3 (spatial awareness), x_5 (quality of experience). A statistically significant relationship was found between temporal complexity (TI)

Model Summary for Group F					
	Coefficient	**Std Error**	**z-Statistic**	**p-value**	**95% CI**
Intercept	-.022	.083	-.267	.789	$[-.184, .140]$
SI	.001*	.000	1.963	.050	$[.000, .002]$
TI	-.001	.001	-.873	.383	$[-.002, .001]$
x_3	.012	.009	1.351	.177	$[-.006, .030]$
x_5	-.003	.002	-1.220	.222	$[-.008, .002]$
δ	.003	.002	1.640	.101	$[-.001, .006]$

(*) p-value < .05, (**) p-value < .01, (***) p-value < .001

Table 17: Mixed linear model regression results for group F.

and degree of gaze distribution N_ψ, with coefficient = $-.006$ ($z = -6.585$, p-value < .001).Notably, the intercept exhibited a positive effect (coefficient = .293, $z = 2.635$ and p-value = .008). The intercept coefficient .293 indicates that, when all variables are set to zero within group R, the model's predicted value of N_ψ will be .293. It's important to emphasise that this scenario is mostly theoretical, as a spatiotemporal image complexity of zero and a complete absence of diegetic artefacts are highly unlikely. The model predictions deviate slightly more from the actual outcomes N_ψ (scale = .0017) compared to group F. No statistically significant results were found among the other predictor variables. The model statistics of group R are presented in Table 18.

Model Summary for Group R					
	Coefficient	**Std Error**	**z-Statistic**	**p-value**	**95% CI**
Intercept	.293**	.111	2.635	.008	$[.075, .512]$
SI	-.001	.001	-1.306	.192	$[-.002, .000]$
TI	-.006***	.001	-6.585	.000	$[-.008, -.004]$
x_3	.012	.008	1.567	.117	$[-.003, .027]$
x_5	.001	.003	.500	.617	$[-.004, .007]$
δ	-.003	.002	-1.495	.135	$[-.008, .001]$

(*) p-value < .05, (**) p-value < .01, (***) p-value < .001

Table 18: Mixed linear model regression results for group R.

Both models exhibit contradictory results. For group F (fixed-position chair), the effect of spatial complexity on degree of gaze distribution was statistically significant, while the effect of temporal complexity was not. Contrary, for group R, the effect of temporal complexity on gaze exploration was statistically significant while the effect of spatial complexity was not. These findings suggest that the relationship between SI and TI with N_ψ depends on the seating type. As such, there was sufficient evidence to reject the null hypothesis and accept the alternative hypothesis:

- H_1: There is a significant difference in the relationship between spatial- and temporal image complexity and degree of gaze distribution (N_ψ) across different seating types.

It should be mentioned that the significant differences are specific to either spatial- or temporal complexity. For spatial complexity, the difference was significant in group F. For temporal complexity, the difference was significant in group R. For the other predictor variables x_3 (spatial awareness), x_5 (quality of experience) and δ (presence of diegetic artefacts), there was insufficient evidence to reject the null hypothesis.

To further examine the interaction dynamics that cause the significant difference in the relationship across groups, an interaction analysis was conducted. Additional parameters, interaction terms, were included in the same mixed-effect multiple regression model from § 5.2.2 to evaluate the interaction terms between both spatial- and temporal complexity and seating type. The extended mixed-effects multiple regression model includes both respective interaction terms $SI : C[T.1]$ and $TI : C[T.1]$, while controlling for the presence of diegetic artefacts (δ), spatial awareness (x_3),

and quality of experience (x_5). The mixed-effects multiple regression model statistics, including interaction terms $SI : C[T.1]$ and $TI : C[T.1]$ are presented in Table 19.

Interaction Model Summary					
	Coefficient	**Std Error**	**z-Statistic**	**p-value**	**95% CI**
Intercept	.098	.069	1.419	.156	$[-.037, .233]$
$C[T.1]$	.061*	.023	2.589	.010	$[.015, .106]$
SI	-.000	.000	-0.063	.949	$[-.001, .001]$
$SI : C[T.1]$	.000	.000	0.315	.753	$[-.000, .001]$
TI	-.002**	.001	-3.069	.002	$[-.003, -.001]$
$TI : C[T.1]$	-.003***	.001	-4.370	.000	$[-.004, -.001]$
x_3	.012*	.006	2.035	.042	$[.000, .023]$
x_5	.000	.002	0.201	.841	$[-.003, .004]$
δ	-.000	.001	-0.159	.873	$[-.003, .003]$

(*) p-value < .05, (**) p-value < .01, (***) p-value < .001

Table 19: Mixed linear model regression results for the interaction model.

The model exhibited a significant interaction effect between seating type and temporal complexity (TI), which was statistically significant (coefficient $-.003$, $z = -4.370$ and p-value $< .001$), signifying that the effect of temporal complexity on gaze distribution N_ψ significantly differs between seating types. As such, a unit increase in TI results in an additional decrease of .003 units in N_ψ. Furthermore, the independent effect size of temporal complexity in N_ψ was also statistically significant, with coefficient $= -.002$ ($z = -3.069$, p-value .002). As such, a unit increase in temporal complexity results in a .002 unit decrease in N_ψ, regardless of seating type.

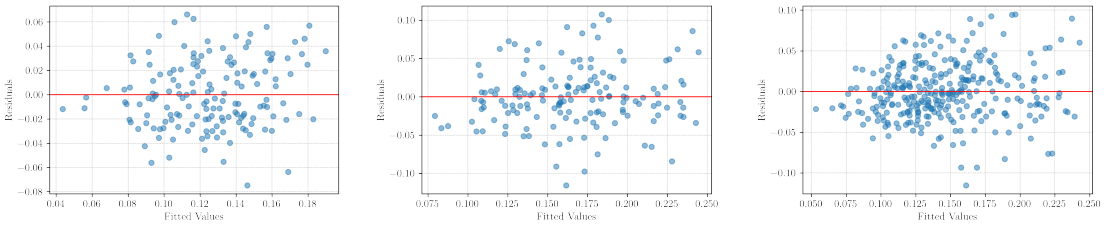
The model did not show a significant relationship between spatial complexity (SI) and N_ψ, nor did it significantly interact with seating type (coefficients = .000, p-values > .05). The relationship between x_3 (spatial awareness) and N_ψ was also statistically significant (coefficient = .012, $z = 2.035$, p-value .042). Additionally, the mixed-effects model exhibited a significant main effect of seating type ($C[T.1]$) on degree of gaze distribution N_ψ ($z = 2.589$, p-value = .010). When keeping all the predictor variables constant, the model a .061 unit increase in N_ψ when switching from a fixed-position chair to a rotating chair. No significant effects were observed for x_5 (quality of experience) and δ (presence of diegetic artefacts) (p-values > .05).

Notably, for most of the predictor variables, the inclusion of interaction terms within the mixed-effects multiple model did not significantly change the relationship with the degree of gaze distribution (N_ψ), When compared to the mixed-effects model without interaction terms (see Table 16), the inclusion of interaction terms within the mixed-effects multiple model did not significantly change the relationship with the degree of gaze distribution (N_ψ) for most of the predictor variables. However, a notable increase was observed in the coefficient of seating type ($C[T.1]$), from .030 to .061.

The findings from the interaction analysis further support the results from the subgroup analyses, demonstrating a complex interaction and moderation between spatiotemporal image complexity and seating type in its effect on gaze distribution.

Assumptions The assumptions of linearity, homoscedasticity, normality, independence, and multicollinearity for both subgroup analyses as well as the interaction analysis using the mixed-effects multiple regression model were assessed.

The fitted value vs. residuals plots of the three models are presented in Figure 34, which were used in the assessment of linearity and homoscedasticity. None of the three plots demonstrate a clear data pattern. As such, it was assumed that the relationships between the IVs and DVs were approximately linear. However, the fitted values vs. residuals plots in Figures 34b and 34c display a slight funnel-pattern in the distribution, indicating the presence of heteroscedasticity. This pattern is less evident in the scatter plot of group F (see Figure 34a). This observation further supports the previously established notion that group F exhibits a less significant influence on the overall data pattern, evident in Figures 31 and 32. Furthermore, a slight asymmetry can be observed across the three plots.

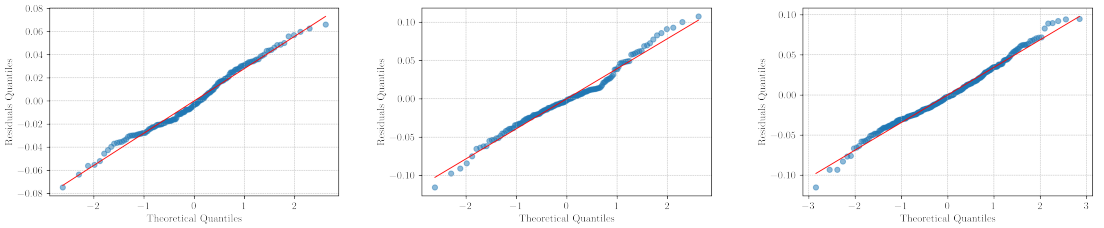


(a) Subgroup F (b) Subgroup R (c) Mixed-Effects Model (Ext.)

Figure 34: Fitted Values vs. Residuals Plots

The normality of the residuals was assessed utilising the Q-Q plot presented in Figure 35. The Q-Q plots imply that the residuals of each of the models were approximately normally distributed, as despite a slight deviation in the tail-ends, the Q-Q distributions align sufficiently with the reference lines across all three models. Additionally, a Durbin-Watson test was performed to examine the independence of errors of these residuals. For group F, group R, and the extended mixed-effects multiple regression model, the resulting values were 2.489, 1.958 and 2.124, respectively. The values strongly indicate the absence of autocorrelation, thereby meeting the independence assumption.

For both group F and group R, the VIF values of x_3 (spatial awareness) and δ (presence of diegetic artefacts) were > 5, suggesting the presence of multicollinearity within both models. The other VIF values were < 5, indicating a lower level of multicollinearity. The interaction terms in the extended mixed-effects multiple regression model suggest a higher level of multicollinearity. Specifically, $C[T.1]$ and $SI : C[T.1]$ were above threshold > 10, and $TI : C[T.1]$ was > 5. These levels of multicollinearity between multiple predictor variables strongly imply that the assumption of no multicollinearity was violated.



(a) Subgroup F (b) Subgroup R (c) Mixed-Effects Model (Ext.)

Figure 35: Q-Q Plots

Specifically the presence of multicollinearity is indicative that the model's performance is dependent on the levels of the independent variables and therefore requires cautious interpretation of the results.

5.2.4 Non-Parametric Comparative Rating Analysis

The ratings for each of the experiential statements were assessed using descriptive statistics, as presented in Table 20. For reference, each of the statements is presented in Table 21. A series of non-parametric Mann-Whitney U tests were conducted on statements S_1, S_2, S_3 and S_4, which were rated by all $n = 52$, to examine the differences between groups.

As evident, users agreed moderately with statements S_1 ($\mu = 4.442, \sigma = 1.259$) and S_3 ($\mu = 4.942, \sigma = 1.110$), indicating that users found themselves to be distracted by background elements due to the static camera. In contrast, users did perceive an increase in spatial understanding due to

All users				
	Mean	**Standard Deviation**	**Minimum**	**Maximum**
S_1	4.442	1.259	1.000	6.000
S_2	5.731	1.223	3.000	7.000
S_3	4.942	1.110	2.000	7.000
S_4	3.404	1.287	1.000	6.000
Group R				
	Mean	**Standard Deviation**	**Minimum**	**Maximum**
S_1	4.615	1.359	1.000	6.000
S_2	5.769	1.306	3.000	7.000
S_3	4.808	1.167	2.000	7.000
S_4	3.269	1.430	1.000	6.000
SR_1	4.500	1.393	2.000	7.000
SR_2	5.962	.871	4.000	7.000
Group F				
	Mean	**Standard Deviation**	**Minimum**	**Maximum**
S_1	4.269	1.151	1.000	6.000
S_2	5.692	1.158	4.000	7.000
S_3	5.077	1.055	3.000	7.000
S_4	3.538	1.140	1.000	5.000
SF_1	6.077	.796	5.000	7.000
SF_2	4.577	1.027	3.000	6.000

Table 20: Descriptive statistics of the experiential statements.

the static camera. A slight disagreement was observed with statement S_4 ($\mu = 3.404, \sigma = 1.287$), suggesting that the moving camera did not necessarily impact the spatial orientation of users. users agreed strongest with statement S_2 ($\mu = 5.731, \sigma = 1.223$), implying that the static camera evoked a more focused gaze behaviour. Similar distributions were found across both groups. Regarding the chair-specific statements, users in group R – strongly – agreed on both statements SR_1 ($\mu = 4.500, \sigma = 1.393$) and SR_2 ($\mu = 5.962, \sigma = .871$). The results suggest that the rotating chair facilitated more exploratory behaviour and tracking of camera movements. In group F, users – strongly – agreed on both statements SF_1 ($\mu = 6.077, \sigma = .796$) and SF_2 ($\mu = 4.577, \sigma = 1.027$). As such, the users agreed on the limiting effects of a fixed-position of the chair on their ability to explore the virtual environment as well as keeping track of camera movements.

Variable	**Statement**
S_1	"I found myself getting distracted by the background elements when watching the videos with a static camera."
S_2	"I found myself more focused on the details of the scene when the camera was moving slowly."
S_3	"I had a better understanding of the layout of the environment when watching the 360-degree content with a static camera."
S_4	"I found it difficult to orient myself and understand the layout of the environment when watching the 360-degree video with a moving camera."
SR_1	"I felt more encouraged to look around because of the rotating chair."
SR_2	"The rotating chair made it easier for me to follow the camera movements."
SF_1	"I felt limited in the amount of exploring I could do due to the fixed chair."
SF_2	"I found it harder to keep track of the camera movements because of the fixed chair."

Table 21: Statement Variables

A series of Mann-Whitney U tests were performed to assess the significant differences between both usability context groups for each of the experiential statements S_1, S_2, S_3 and S_4. With

respective p-values: .181, .711, .404 and .404, all four experiential statements failed to reach statistical significance. Despite observing mean differences between the two groups, the differences were not statistically significant at threshold $\alpha = .05$. As such, the observed differences could potentially be due to random variation. The Mann-Whitney U tests did not provide sufficient evidence to reject the null hypotheses:

- H_0: There is no significant difference in the median rating for statements S_1, S_2, S_3 and S_4 between the two seating types.
- H_1: There is a significant difference in the median rating for a given statement X between the two seating types.

Assumptions The assumptions of independence of observations, ordinal data and equal distributions for the Mann-Whitney U tests were assessed. All observations (i.e., ratings) were independent of each other across all $n = 52$ users. The ratings were measured on an ordinal 7-point Likert scale. Despite slightly deviations in the distributions, there was enough evidence to support that the assumptions of the Mann-Whitney U tests were met.

5.3 Qualitative Results

The qualitative data, as acquired during the user evaluation, was analysed using a systematic grounded theory approach, in which a series of emergent coding procedures and graphical representations were used to identify key trends in the user's self-perception of conscious gaze behaviour. Important to highlight is the subjective component of the qualitative data. The resulting key trends of subjective gaze behaviour were based on the, as specifically mentioned by the users, behavioural responses users experienced when reflecting on their gaze behaviour across varying 360-degree videos.

5.3.1 Grounded Theory Analysis

The transcript data of the user's self-perception of active gaze behaviour was systematically assessed using the Straussian Grounded Theory methodology [318, 340]. Iterative open, axial and selective coding procedures were employed to identify and assign codes to discrete strings of text (i.e., concepts) that encompassed the user experience and perception of gaze behaviour. The code packages were categorised and combined in clusters, which were assigned to higher-level categories. Frequent notions of the same behavioural concept were placed in higher-level categories. The multi-level categories and codes were selectively linked based on causal relationship, context, consequences and conditions. The resulting inter-connected theoretical concepts and relationships were utilised to construct a framework of recurrent themes and patterns identified from the responses. An acyclic forest graph was constructed to visualise the inter-connected framework of codes and categories, as presented in Figure 36. The forest graph contains a series of nodes (codes) and branches, linking the multi-level categories and corresponding codes. The hierarchical organisation of theoretical concepts and observations enables an increased understanding of the self-perception of active gaze behaviour.

However, to ensure comprehensibility during the user evaluation, the questions were oriented towards changes in genre and camera motion, both of which inherently represent variations in spatiotemporal complexities. Consequently, the resulting framework (see Figure 36) from the grounded theory analysis and emergent coding procedures predominantly focuses on the genre and motion-related distinctions rather than the underlying connections with spatial and temporal complexities.

To identify key trends of active gaze behaviour in relation to varying levels of spatial and temporal image complexities, it is imperative to interpret the observations as such. The forest graph from Figure 36 was modified to better reflect the user's self-perception in relation to varying levels

Figure 36: Acyclic forest graph of gaze behaviour perceptions.

of spatiotemporal image complexity, see Figure 37. The revised forest graph visualises the interconnected framework of relationships and observed patterns within the context of spatiotemporal image complexity.

5.3.2 Trends in User Self-Perception

General overarching themes were derived from the qualitative transcript data. Elements of engagement, focus, pace, interest, immersion, relatability and anticipation were prominent and frequent notions within the data. As such, users expressed a variety of both unique as well as common behavioural responses. Causal relationships between the observations enabled the identification and definition of a set of key trends found within the active perception of a user's gaze behaviour across the sample size $n = 52$.

Figure 37: Adapted forest graph of gaze behaviour perceptions.

The following set of key behavioural trends were defined, utilising the forest graph containing the users' self-reflected behavioural trends in relation to spatiotemporal image complexity:

1. Adaptive Behaviour and Interest-Driven Engagement;
2. Objective-Focused Behaviour and Anticipation-Driven Engagement;
3. Active Disengagement in Less Complex Content;
4. Augmented Sense of Presence.

It is important to emphasise that the trends represent frequent notions and patterns within the active gaze behaviour of the users. As such, these are patterns and behavioural choices that users actively recalled and were aware of. Thereby, these behavioural choices and patterns are representative of the overall self-perception of the user's viewing behaviour and how changes in presented 360-degree videos may have caused active behavioural changes in gaze behaviour. The trends ought to be interpreted as they were derived from the self-perception of users, and remain rather subjective notions on self-reflected gaze behaviour.

Adaptive Behaviour and Interest-Driven Engagement Most notably, user mentioned a certain degree of adaptability in their viewing behaviour depending on the video content. The users found their focus was impacted by the wide variety (i.e., spatiotemporal complexities) of the presented content. For instance, users felt more engaged and focused (mentally) during videos pertaining a higher temporal complexity due to the dynamic camera motion. This resulted in a decreased tendency to explore, as the users felt more drawn into the action. On the contrary, static and visually rich videos with a lot of detail resulted in increased exploration, as user's felt a higher

need to observe and appreciate the virtual environment and scenery. Overall, users mentioned that they were more engaged with content that better aligned with their personal interest.

Objective-Focused Behaviour and Anticipation-Driven Engagement The sense of "feeling drawn" into the action furthermore diminished the user's tendency to explore. A trend was observed in users mentioning an enhanced level of focus when viewing 360-degree videos with a clear objective. The sense of anticipation and adrenaline led to users being concentrated on the objective in front (e.g. the roller coaster track or following a path). As such, there was a notable focus on the objectives and ongoing action, accompanied by a lower tendency to explore, in dynamic videos.

Augmented Sense of Presence The dynamic videos, due to their higher temporal complexities, further amplified the users' sense of presence. The ongoing action, sense of anticipation and strong focus on the objective resulted in an intense sense of involvement among the users. This increased sense of immersion was experienced in spatially complex videos as well, as the virtual environment began to mirror the visual richness of real-life experiences, blurring the clear distinction. The intricate details found in spatially complex videos was perceived to be more inviting.

Active Disengagement in Less Complex Content The before-mentioned trends in behavioural tendencies focused on the impact of spatially or temporally complex videos on active gaze behaviour, as perceived by the users. However, a lack of complexity in either dimension also elicited a behavioural response. Users mentioned an active form of disengagement when viewing 360-degree videos that were perceived as uninteresting, or which failed to grasp user attention. A higher sense of distraction was found, which resulted in a higher tendency to explore as the users sought out more engaging visual attributes.

Configuration	Subjective Behavioural Response
SI_{high}, TI_{high}	Augmented sense of immersion coupled with an adaptive gaze behaviour as exploration tendency was highly dependent on user interest and video objective.
SI_{high}, TI_{low}	Increased tendency to explore or focus on details, due to rich spatial details and fewer changes happening over time.
SI_{low}, TI_{high}	Increased focus and limiting extent of exploration, due to the sense of anticipation and objective-oriented gaze.
SI_{low}, TI_{low}	Increased tendency to explore and desire for more engaging visual artefacts due to active disengagement, caused by lack of complexity.

Table 22: Spatiotemporal configurations and subjective behavioural responses.

Notably, users expressed different behavioural tendencies depending on the level of complexity of either spatial- or temporal dimension. Users experienced a lower tendency to explore during temporally complex videos, while in contrast, more spatially complex videos encouraged exploration. Despite users feeling actively disengaged with the content, lower complexities (both spatial- and temporal) still resulted in exploratory tendencies. However, these tendencies were primarily motivated by the prominent search for visually engaging artefacts within the 360-degree environment. Users frequently mentioned the pivotal role of user interest and video objective, as their gaze behaviour depends more those components during highly spatial and highly temporal videos. Users expressed different subjective behavioural responses depending on the specific spatial- and temporal configuration of each 360-degree video. Table 22 presents an overview of the subjective behavioural response, as reflected and self-perceived by the users, depending on the specific configuration of spatiotemporal video complexity.

Chapter 6

Discussion

Motivated by the existing body of literature and research, this thesis presents a systematic approach to bridge the dichotomous state of research, predominantly focused on the technical limitations rather than content-aware approaches, by examining the behavioural consequences inherently imposed by 360-degree video content in VR. As such, the 360-degree imagery acts as an autonomous factor in the dynamic interaction process. The resulting research objective was defined as follows:

> To discern the extent of which spatiotemporal image complexity of a 360-degree video sequence in VR influences gaze behaviour within the multifaceted interaction model, while factoring in the complex dynamics of cognitive perceptions and usability context.

Central to the research, as proposed in this thesis, was the provision, exploration and enabling of a robust understanding of how 360-degree video sequences, as an independent component, impact the user's gaze behaviour. Specifically, the spatiotemporal image complexity of a 360-degree video sequence was used to quantify the intricacies of the content in terms of space and time, providing an accurate representation of the human visual system. In the context of a 360-degree video sequence, spatial complexity (SI) refers to the amount of detail and variation within each frame, whereas the temporal complexity (TI) measures the amount of change or motion in consecutive frames [144]. A mixed-methods design was applied to measure user gaze, utilising measured fixations throughout various 360-degree video sequences in VR and across different seating types. This work holistically approaches the research objective by employing computer vision techniques and oculesics to analyse and quantify 360-degree image structures as well as consequent gaze behaviour in a comprehensive and multi-dimensional manner.

This chapter presents the discussion, in which the main findings are presented and interpreted against the backdrop of the cognitive perceptions, film theory, and usability context by incorporating relevant theoretical frameworks from the related fields. This discussion presents the main findings, as derived from the employed analytical framework (see § 2.7.3), adhering to a similar structure of the devised sub-questions from § 1.7. Firstly, the systematic employment of computer vision techniques and eye-tracking data to quantify gaze patterns is discussed. The resulting quadrifactorial exploration index measure was used to the examine the complex dynamics and identify the confounding effects of cinematographic principles and cognitive influences, prior to the primary analysis on the effect of spatial- and temporal image complexity on gaze behaviour. Following this, the findings of the primary spatiotemporal analysis on the effect of both spatial- and temporal image complexity on gaze distribution are presented, providing insight in the significance of 360-degree image complexity on directing user behaviour in VR. Furthermore, the degree to which this main effect is moderated by interactions across different usability contexts is discussed, highlighting the influence of seating type on the behavioural effects. Additionally, the association between subjective self-perception of gaze behaviour across users and objective gaze distribution is examined, providing insights into the discrepancies between self-awareness and actual gaze behaviour. As such, the comparison sheds light on the balance between unconscious cognitive responses and active gaze. Lastly, the main findings are briefly reiterated and succinctly presented.

This chapter furthermore discusses the presented findings in relation to higher-level implications for theory and practice, presented in sections 6.1 and 6.2, respectively. The chapter concludes by discussing the limitations of this thesis in section 6.3, while the future prospects of potential research continuation is discussed in section 6.4.

The Quantification of Gaze Patterns Through Computer Vision Techniques and Oculesics

The implementation of oculesics to acquire sensor-based physiological eye-tracking data, enabled highly detailed insight of the user's gaze and behavioural responses to various 360-degree stimuli. Despite its effectiveness, traditional eye-tracking metrics only present a one-dimensional view, solely encapsulating the location and duration of fixations. A comprehensive understanding of user

behaviour and gaze patterns, in the context of 360-degree video interaction in VR, necessitates a methodological approach to adequately quantify the intricacy of gaze distribution in 360-degree environments, reducing the need for a separate post-test subjective visual analysis to interpret the user's gaze and attention [129]. This work introduced the quadrifactorial exploration index N_ψ, a novel metric formulated to quantify gaze distribution based on heatmap imagery extracted from advanced eye-tracking software. The heatmap image was constructed by combining both fixation and duration data to visualise the extent of gaze distribution. By utilising image segmentation techniques, derived from the field of computer vision, the metric segments four components from the gaze distribution heatmap image: pixel area coverage ratio, average pixel intensity, structural image dissimilarity and entropy. A binary map, superimposed on the heatmap image signal, enabled the identification and computation of each individual pixel as part of the heatmap. As such, the surface area of all heatmap pixels and the average pixel intensity were used to compute the extent of gaze exploration. Despite providing a sufficient indication of exploratory extent, those two factors alone did not take into account the complex patterns of gaze distribution. Consequently, the index was enhanced by computing the structural dissimilarity between a blank reference image (representative of zero gaze exploration) and the superimposed heatmap image. The approach combines a modification to the Multi-Scale Structural Similarity Index Measure from Wang et al. (2004) and integrates Shannon entropy to account for the variability of the gaze patterns, enhancing the index with comparable factors of the human visual system [131, 345, 357]. A data-driven approach was applied to mathematically formulate and derive the relative contribution of each of the four factors to the overall degree of gaze exploration, resulting in a nuanced index that denotes the degree of gaze exploration. The internal consistency was assessed and achieved, validating the inclusion of highly correlated components which consistently reflect the spatial extent, concentration, diversity and randomness of gaze distribution patterns.

The Confounding Cinematographic and Cognitive Influences

The quadrifactorial exploration index N_ψ was furthermore employed to ensure internal validity of the primary spatiotemporal analyses. The index was utilised to identify and examine potential influences from other influences, as implied by research insights from cognitive science and film theory [9, 23, 42, 63, 263, 284, 356, 375]. As such, the versatility of the quadrifactorial exploration index N_ψ extended beyond the primary analysis of spatiotemporal image complexity on gaze behaviour, as it was furthermore employed to identify potential confounding effects on gaze distribution. Specifically, the index served as a crucial tool in the assessment of cinematographic principles and influence thereof on guiding user attention. One of the key challenges in achieving external validity was the unique nature of the content, as no two 360-degree video sequences are identical. Therefore, it was decided to assess the confounding effect of the internal story- and scene elements that support the narrative structure (i.e., diegesis) in 360-degree environments. The diegetic assessment takes into account the internal presence of diegetic artefacts within the utilised 360-degree video sequences, as it was proven to guide user attention in traditional video format [23, 63, 263, 375]. The conducted diegetic assessment enabled the identification and assessment of the presence of diegetic artefacts in the employed 360-degree video sequences. The diegetic assessment introduced another novel metric, δ, which utilised a devised coding scheme to quantify the presence of such visual artefacts based on size and duration for each of the 360-degree video sequences. The metric δ denotes the presence of visual artefacts, representing the extent to which inherent narrative story- or scene elements guide user attention and impact user gaze. The assessment was driven by this exact understanding: that a higher presence of diegetic artefacts δ influences exploratory tendencies. The data exhibited non-linear behaviour, as an increase in diegetic artefacts did not invariably correspond to a higher degree of gaze distribution, underpinning the complexity of the relationship between δ and N_ψ. It was found, through a series of non-linear regression models, that a significant variation in gaze distribution could be explained utilising a quadratic polynomial model. These findings evince the notion that internal story- or scene elements within the 360-degree video sequence act as a confounding variable, and as such, was included in the primary analytical models. However, despite exhibiting significant results,

a substantial proportion of the variation remained unaccounted for, as the presence of diegetic artefacts only accounted for a marginal proportion of the observed variation in gaze distribution. These observations emphasised the necessity of incorporating influential factors from other related fields, such as cognition and usability.

As such, the quadrifactorial exploration index N_ψ furthermore enabled the examination of independent influences for each of the devised attributes of perception on gaze behaviour. Founded by research insights from literary works in the field of cognitive science, the perceptual attributes included the levels of QoE, engagement, attentional focus, spatial awareness and fear of missed content [81, 90, 205, 223, 230, 376]. The independent effect of each attribute on gaze behaviour was individually examined using a linear mixed-effects model. Most notably, only two attributes exhibited a significant effect. Firstly, a positive association was observed between spatial awareness and extent of gaze exploration, suggesting that users increasingly navigated the 360-degree environment when more aware of their spatial surroundings. Secondly, a slightly negative association between quality of experience and gaze distribution was found, suggesting that a higher quality of experience may lead to users focusing more on specific AOIs and reducing the overall gaze extent. The attributes of engagement, attentional focus and fear of missed content did not demonstrate a significant independent effect on gaze distribution, suggesting that their influence on gaze behaviour may be less direct, despite their theoretical significance. Consequently, the presence of diegetic artefacts, as well as both attributes of cognitive perception and usability context were validated as confounding effects. The specific influence of usability context is later discussed in this chapter.

The formulation and employment of the quadrifactorial exploration index N_ψ was evidently pivotal in understanding the complexity of gaze behaviour in virtual 360-degree environments. The index utilised computer vision techniques and oculesics to encapsulate the multidimensional nature of gaze behaviour in which it reflects not just the extent and intensity of user gaze, but also its variability and unpredictability. As such, the index was further employed to detect confounding effects from various domains, reiterating the versatility of the multifaceted measure and ensuring a higher degree of internal validity. This approach resulted in a reliable and nuanced measure of gaze patterns, enabling more intricate exploration of user interactions in VR.

The Dynamics of Spatial- and Temporal Image Complexities on Gaze Distribution

Consequently, the quadrifactorial exploration index N_ψ was utilised in the primary spatiotemporal analysis, adhering to the main research objective of discerning how gaze distribution is influenced by spatiotemporal 360-degree image complexity in VR. The primary spatiotemporal analyses controlled for factors of diegesis, cognitive perception and usability context, as well as respective spatial- or temporal dimensions.

The primary spatiotemporal analyses revealed that spatial image complexity (SI) did not significantly impact the user's gaze when watching the 360-degree video sequence. This suggest that the users' gaze exploration remained relatively unaffected, irrespective of the visual richness within each frame of the video. This finding challenges the theoretical significance of spatial complexity in traditional videos [28, 202], which notes that more complex imagery would inherently elicit a more distributed gaze as users aim to capture as much visual information. In contrast, these findings imply that the spatial richness of a 360-degree video sequence might not be a critical factor in eliciting gaze behaviour, and emphasise that other factors such as usability context or cognitive load might override the significance of spatial complexity on gaze behaviour.

In contrast, temporal image complexity (TI) was found to be a significant predictor of gaze distribution. The results indicate a negative association between TI and N_ψ, where for every unit increase in temporal complexity, a decrease in the extent of gaze distribution was observed. This suggests a behavioural tendency among users to concentrate their focus on specific AOIs when prompted with rapidly changing visual information. These finding highlights the pivotal role of time and temporal dynamics in guiding attention. It emphasises that time and change of visual

information over time are important dimensions in the understanding of gaze behaviour in virtual 360-degree environments.

The relationship between spatiotemporal image complexity in 360-degree video sequences and gaze distribution was found to be complex and nonlinear. While temporal complexity significantly impacts gaze behaviour, the independent effects of spatial complexity were found to be non-significant. However, the lack of a significant independent effect of SI does not diminish the impact of spatial image complexity. In fact, when spatial complexity is coupled with temporal complexity, intricate gaze distribution patterns emerge, emphasising the complex interaction between the two dimensions.

The generated three-dimensional surface plot visualise the complex interaction between SI, TI and N_ψ (see Figures 31 and 32). The surface plots reveal intricate patterns, as the curvature of the surface varies across the spatiotemporal plane. As such, the plots indicate higher-degree non-linear relationships between spatial complexity, temporal complexity and degree of gaze distribution. Parallel ridges in the surface plots were observed for specific configurations of spatiotemporal image complexity, strongly indicating repeated patterns in gaze exploration. Furthermore, the surface plot demonstrated that different configurations in the spatiotemporal matrix can elicit different behavioural patterns. Most notably, a significant peak in gaze distribution was found at relatively low levels of spatiotemporal complexity, predominantly impacted by the negative association with temporal image complexity. The surface plot exhibited a downward slope across the temporal plane, indicative of the significant negative association, while periodic curvature across the spatial plane can be observed. The periodic increase and decrease in gaze distribution, as indicated by the slopes and parallel ridges in the surface plots, further support the interaction effects between spatial- and temporal image complexity in virtual 360-degree environments. As such, the influence of spatial- and temporal complexity on gaze behaviour is not simply additive, but interacts in more complex ways. Therefore, despite the insignificant independent effect of spatial complexity, it is vital to consider both spatial- and temporal image complexity as a whole, as opposed to separately. The importance of this notion is discussed in more detailed in the subsequent section.

Notably, the presence of spatiotemporal image complexity was found to dominate the significant independent effects of presence of diegetic artefacts δ and quality of experience. In isolation, these confounding factors were observed to be significant predictors of gaze distribution. However, the independent effects became less pronounced when coupled with spatiotemporal image complexity, emphasising the significance of spatiotemporal image complexity and its interactions in impacting user behaviour. As hypothesised, both spatial awareness and usability context (i.e., seating type) remained significant predictors of N_ψ in the presence of spatiotemporal image complexity.

The Effects of Spatiotemporal Image Complexity on Gaze Behaviour Across Usability Contexts

The use of a different usability contexts (i.e., seating types) was found to significantly impact gaze behaviour during the 360-degree video interactions, which supports the findings of Ebrahimi et al. (2009) and Brunnström et al. (2013) [42, 90]. Two seating types were utilised in the study: a fixed-position chair and a rotating chair, representative of two common contexts in which the user interacts with 360-degree video in VR. The users perceived both seating types as sufficiently comfortable and enjoyable, ensuring that usability factors did not introduce additional confounding usability effects or influence the results due to discomfort. This approach enabled the examination of to what extent the effects of spatiotemporal image complexity on gaze behaviour are impacted across usability contexts, adding an extra dimension to the thesis. The overall gaze distribution across users was significantly smaller when utilising a fixed chair compared to a rotating chair.

Evidently, the before-mentioned overall spatiotemporal effects on gaze behaviour across all $n = 52$ users were primarily reflected in the gaze behaviour of users utilising the rotating chair. This notion is further supported by the extreme similarities between the two respective surface plots (Figures 31b and 32b). Furthermore, the group-specific surface plots from Figure 32 reveal

that the behavioural patterns inherent to using the rotating chair were to a large extent mirrored in the behavioural patterns of using a fixed chair.

Due to the significant impact of seating type on the overall effect of spatiotemporal image complexity on gaze behaviour, it was decided to examine this effect separately between groups. Consequently, a subgroup and interaction analysis examined the mediating effect of seating type. Most notably, depending on the seating type, the spatiotemporal effects were significantly different. For instance, when using a fixed-position chair, spatial image complexity (SI) exhibited a statistically significant positive relationship with gaze distribution N_ψ, while temporal image complexity (TI), spatial awareness, quality of experience and presence of diegetic artefacts δ did not. In contrast, when using a rotating chair, only temporal image complexity (TI) demonstrated a significantly negative association with gaze distribution and the other predictors did not demonstrate any significant effects. These findings suggest a different effect of either spatial- or temporal image complexity depending on the seating type. When using a fixed-position chair, spatial image complexity significantly affects gaze distribution, while in a rotating chair, temporal image complexity does.

To further examine the significance of this usability context group-dependent variation, additional interaction terms between both spatial- and temporal complexity and seating type were included in the mixed-effects multiple regression model of the primary spatiotemporal analysis. A significant interaction effect was found between seating type and temporal image complexity (TI), indicating that the effect of temporal complexity on gaze distribution N_ψ significantly varies depending on the seating type. Furthermore, the independent effect of temporal complexity remained significant, suggesting that an increase in temporal complexity leads to a decrease in gaze distribution, irrespective of seating type. However, the interaction and independent effects of spatial image complexity (SI) were not significant.

The primary spatiotemporal analysis already revealed that the significance of spatial complexity on gaze behaviour might be overriden by other factors such as usability context. The interaction analysis further supports this notion, as it revealed that users adopt different approaches to navigate the spatially complex environments i.e., by using a rotating chair. The ability to rotate more easily and consequently change viewing direction more conveniently enables the users to better handle a spatially complex 360-degree environment, mitigating the independent effect of spatial complexity on gaze behaviour. This finding indirectly points to the earlier notion of relevance of spatial image complexity.

Despite these findings, it is important to emphasise that the interaction effect between spatial complexity and seating type might only be significant within specific groups. This is strongly suggested by the findings from the subgroup analysis of using a fixed chair, in which spatial complexity was found to be a significant predictor. Therefore, while spatial image complexity might not exhibit a significant interaction effect with seating type on gaze behaviour across the entire sample, the interaction analysis strongly indicates the possibility that it still could within specific usability contexts. Notably, the interaction analysis also revealed a significant main effect of seating type on the degree of gaze distribution N_ψ, where a switch from a fixed chair to a rotating chair was associated with a .061 unit increase in N_ψ when all other predictor variables were kept constant. This is a substantial increase when taking into account the value range [0, 1] of the quadrifactorial exploration index N_ψ and is double the effect size found in the model without interaction terms. The inclusion of interaction terms in the model did not significantly change the relationships between the predictor variables and degree of gaze distribution N_ψ when compared to the model without interaction terms.

These findings highlight the complex relationship between spatiotemporal image complexity, gaze distribution and usability context. Both the fixed and rotating seating types, representative of various usability contexts, appear to significantly moderate the effect of 360-degree spatiotemporal image complexity on gaze behaviour, emphasising the pivotal role of usability context in the interaction process.

The Dichotomy Between Objective and Subjective Gaze Behaviour

A comprehensive understanding of the intricate 360-degree video user interaction and imposed behavioural responses due to variations in spatiotemporal image complexity necessitates a holistic perspective which integrates both objective and subjective angles. Motivated by the works of Holmqvist et al. (2011) and Egan et al. (2016), this thesis employed a combination of both quantitative and qualitative methodologies [91, 130]. This approach was further driven by the notion that the enhanced perceptual load, inherent to 360-degree user interactions in VR, could trigger unconscious behavioural responses [196, 223, 262]. As such, the employed research methodology exceeds the limited perspective, offered by purely focusing on objective gaze data, integrating the underlying subjective motivations and perceptual experiences that guide gaze behaviour. This holistic and empirical approach, which includes the integration of experiential statement results and qualitative interview transcript data on self-reflected gaze behaviour, not only highlights the dichotomy between objective gaze behaviour and subjective perception of gaze behaviour but also enables a deeper understanding in the underlying motivation behind the observed gaze patterns.

The qualitative analysis resulted in a set of four key behavioural trends that delineate the subjective experiences and perceptions of the users. Derived from the perceived senses of engagement, interest, focus, pace, relatability, anticipation and immersion, the key behavioural trends provide insights in the behavioural responses on a subjective and cognitive level, providing a holistic perspective in conjunction with the objective findings. To reiterate, the objective findings derived from the primary spatiotemporal analyses revealed that temporal image complexity significantly affects gaze behaviour, which were negatively correlated. However, despite the theoretical implications, no significant independent main effect was found between spatial image complexity on gaze distribution.

The subjective findings, indicative of the user's self-perceived gaze behaviour, provided further insight into this observation as they revealed an interesting discrepancy: users still experienced an increased tendency to explore the more spatially complex environments. This notable discrepancy could be attributed to the increased cognitive efforts involved in processing spatially complex 360-degree scenes, leading users to believe that they are exploring more than they objectively are. Moreover, users also experienced an increased tendency to focus on details, directly contradicting the previous observation. This dual observation of both increased exploration and focus was found during low temporally complex videos, as users reported that the static behaviour of the camera enabled them to both explore the background elements more extensively, as well as allowed for them to focus on the visually rich details of a specific area. These findings suggest that the user navigates the spatially complex virtual 360-degree environment in nuanced ways. The enhanced level of detail could lead to more exploratory behaviour for some users, while others might adopt a more focused gaze. A similar contradiction was found in less spatially complex environments, as the users' active disengagement (i.e., sense of boredom) elicited an increased search for more engaging visual attributes, counterbalancing the hypothesised decrease in gaze distribution due to low spatial image complexity. This adaptive gaze behaviour could result in an insignificant average effect and strongly supports the before-mentioned notion that cognitive load overrides the independent effect of spatial complexity on gaze behaviour. These findings furthermore highlight the relevance of taking into account spatial image complexity, as it, in conjunction with temporal image complexity, results in various distinct behavioural tendencies.

On the other hand, subjective perceptions of the impact of temporal image complexity were in line with the objective findings as users experienced an increased difficulty orienting and understanding the virtual layout during temporally complex 360-degree videos. The subjective findings reflect the significant negative relationship found between temporal image complexity and degree of gaze distribution, suggesting that the rapid visual changes in adjacent frames forces users to focus on the movement to maintain spatial orientation, limiting gaze exploration and reducing risk of cybersickness [2, 121, 207, 267]. This finding also strongly validates the statistically significant confounding influence of spatial awareness on the degree of gaze distribution. Notably, despite the controllable POV during the 360-degree video interaction, the temporal complexity in the 360-degree video still resulted in a narrower focus, seemingly rendering the additional degree

of freedom obsolete. The combined risk of potentially missing out content with rapidly changing visual information significantly enhances cognitive load, resulting in less demanding behavioural consequences in order regulate cognitive processing [262, 281, 313]. This observation can also be attributed to the theoretical underpinnings of event segmentation theory, the cognitive process of breaking information into meaningful and comprehensive events [28, 175, 197, 272]. In addition, the higher temporal complexity also evoked a sense of anticipation which made users feel more absorbed in the action, especially in objective-based videos.

The subjective influence of seating type also aligns with the significant effect of usability context on gaze behaviour as users found the rotating chair to facilitate more exploratory behaviour and camera-tracking, while the fixed-position chair limited this ability. This subjective perception explains the higher degree of gaze distribution found among users utilising the rotating chair compared to the users utilising the fixed chair. However, it is important to note that users might not be fully aware of the extent to which their physical context (i.e., their ability to move and look around) impacts their gaze behaviour.

The subjective user experiences provided deeper insights into the users' self-perception of gaze behaviour, distinctly highlighting the various behavioural tendencies and patterns depending on the level of complexity of spatiotemporal dimensions. The behavioural tendencies support the previous observation that specific configurations of spatial- and temporal image complexity in 360-degree videos elicit specific patterns in gaze distribution, as evinced by the three-dimensional surface plots in Figures 31 and 32. The surface plots display repeated patterns of periodic increase and decrease in gaze distribution, indicated by parallel slopes and ridges in the surface plots. The insights from the subjective user experiences suggest that the repeated and periodic patterns in gaze behaviour could be attributed to the perceptual implications of specific spatiotemporal configurations, emphasising the pivotal role of underlying subjective perceptions in facilitating adaptive gaze behaviour.

The integration of both quantitative and qualitative research methodologies enabled the identification of a substantial dichotomy between objective gaze data and subjective user experiences, often found between computational saliency models and actual gaze behaviours [86, 99, 116, 237]. The subjective findings enabled the identification of key behavioural trends, such as adaptive gaze behaviour, interest-and anticipation-driven engagement, objective-focused behaviour, active disengagement and augmented sense of presence, providing valuable insights on the emergence of specific gaze patterns. In particular, as they elucidate the complex cognitive processes involved in the experienced sense of immersion in VR environments. Furthermore, the subjective experiences aid in contextualising the objective findings as they reveal the nuances in behavioural responses not immediately evident from the objective gaze data. In addition, the subjective findings also facilitating a better understanding of the insignificant independent effect of spatial complexity and significant negative impact of temporal complexity on gaze behaviour on the level of cognitive perception. The insights gathered from the subjective experiences also enabled a better understanding in the intricate and complex patterns found in the three-dimensional surface plots, as the users unconsciously adapt their gaze depending on the specific configurations of spatial- and temporal complexity in the 360-degree video sequence as well as their subjective experiences and perceptions. These findings provide a holistic understanding of the user interaction, emphasising on how perception contributes to the complex gaze dynamics elicited by spatiotemporal complexities in 360-degree video sequences in VR.

Retrospective: Understanding Gaze Dynamics in VR through Spatiotemporal Image Complexity, Cognition and Usability Context

In retrospect, the research presented in this thesis discerns the extent to which spatiotemporal image complexity of a 360-degree video sequence in VR influences gaze behaviour within the multifaceted interaction model, while factoring in the complex dynamics of cognitive perceptions and usability context. This work predominantly focused on the behavioural consequences imposed by spatiotemporal image complexity, as representative dimensions of space and time, of 360-degree video sequences in VR. As such, this content-aware approach bridges the dichotomous

state of current research. It shifts the predominant focus on the technical limitations of 360-degree video interaction in VR, and instead signifies 360-degree video content as an autonomous and independent factor in the interaction process.

The findings revealed an intricate and complex relationship between spatial- and temporal image complexities and gaze behaviour. Temporal image complexity exhibited a significant negative effect on gaze behaviour, indicating that rapidly changing visual information over time leads to a more narrow and concentrated focus of gaze. However, spatial image complexity was found to not significantly affect gaze behaviour, implying that the visual richness and level of detail of the 360-degree video alone does not inherently increase the user's extent of gaze exploration. In addition, the findings attributed this lack of independent effect of spatial complexity to confounding factors of usability context and cognitive load.

This implication was supported by the qualitative analysis, as the increased cognitive efforts involved in processing the richness of a more spatially complex environment resulted in users subjectively reporting higher levels of exploration than was objectively measured. The integration of both quantitative and qualitative methodologies enabled the identification of this substantial dichotomy between objective gaze data and subjective user experiences, highlighting the substantial cognitive influence involved in the sense of immersion in VR environments and 360-degree video interaction. While spatial image complexity did not significantly impact gaze behaviour independently, it did impose a perceptual impact on the users' gaze behaviour. Its interaction with temporal complexity generates intricate patterns in gaze behaviour, emphasising the importance of considering both dimensions of spatiotemporal image complexity in 360-degree video user interaction. Furthermore, the subjective findings provided additional insights into the negative correlation between temporal complexity and gaze behaviour, aligning with the objective findings. The users' difficulty in maintaining spatial orientation during the rapid visual changes associated with a temporally complex 360-degree video reflects the observed significant negative effect of temporal image complexity, as well as the observed significant independent effect of spatial awareness.

As such, the deployed research methodology enabled the exploration of behavioural tendencies that not only validate the objective observations, but also elucidated the underlying cognitive processes that shape gaze behaviour. Most notably, the users' adaptive gaze behaviour, which is guided by their interest, sense of presence, anticipation, and immersion, highlights how users navigate the spatiotemporally complex virtual environments on a cognitive level.

Furthermore, the findings reveal that the complex and non-linear interactions between spatial- and temporal image complexities elicit discrepancies in the gaze patterns, further moderated by usability context such as seating types. As such, it was found that the specific usability context (i.e., seating types), in which the user watches the 360-degree video, significantly impacts gaze behaviour. Additionally, a more extensive gaze distribution was observed among users utilising a rotating chair as opposed to users utilising fixed-position chairs. Moreover, a variability in the effect of spatiotemporal complexity on gaze behaviour was observed across the seating types, as usability context moderates the independent effects of either spatial- and temporal image complexity. Specifically, when utilising a fixed chair, spatial image complexity exhibits a significant independent effect, whereas temporal image complexity does not. In contrast, when utilising the rotating chair, only temporal image complexity significantly impacts gaze behaviour. These findings suggest that the influence of spatial image complexity is inherently dependent on the users ability to navigate the spatially complex environment and underlying cognitive influences, mitigating the overall impact thereof.

The employment of oculesics and image segmentation techniques, derived from the field of computer vision, played an integral part in the quantification of gaze patterns and execution of the analyses. The formulation of the quadrifactorial exploration index N_{ψ}, which captures not only the extent and intensity of gaze but also the variability and randomness of gaze patterns, provided a reliable approach to quantify the degree of gaze distribution found in attentional heatmap imagery. Moreover, the versatility of the novel metric facilitated the detection and assessment of confounding influences from the domains of cognitive science and film theory, ensuring a higher degree of internal validity. The index facilitated a comprehensive diegetic assessment, which introduced

another novel metric δ, in which the confounding influence of diegetic artefacts on gaze behaviour was assessed. In addition, the versatility of the index furthermore detected the significant impact of specific perceptual attributes such as spatial awareness and quality of experience (QoE) on gaze distribution.

In conclusion, this thesis provides a significant and comprehensive understanding of gaze dynamics during 360-degree video interactions in VR. By quantifying the 360-degree video content in terms of its spatiotemporal image complexity, the properties of the content were represented in distinct dimensions of space and time, ensuring a higher external validity. The employment of computer vision techniques, oculesics and both quantitative and qualitative research methodologies revealed the nuanced interplay of confounding cognitive perceptions, usability context and cinematography in shaping gaze behaviour. The findings emphasise the importance of the 360-degree content's autonomous role in the user interaction process, and reinvigorates the significant contribution of content-aware approaches to modern research. The insights derived from this work holds significant implications for advancing not only advancing theoretical research within the field, but also the practical applications related to the development of immersive 360-degree environments in VR.

6.1 Implications for Theory

One of the most profound theoretical implications of this work is related to the significance of spatiotemporal image complexity on gaze behaviour. The work highlights the importance of a content-aware, interdisciplinary approach and reveals the complex influences of spatiotemporal image complexity, usability, cinematography and cognitive perceptions. The significant effect of content, as an autonomous and independent factor in the interaction process, signifies the importance of content-specific theoretical frameworks that encompass the intricacies of 360-degree video interactions [9, 90, 132, 151]. In addition, this work also emphasises the significance of contextual factors, such as different seating types, in influencing gaze behaviour during 360-degree video content in VR. It demonstrates significant role of physical and environmental contexts within the multifaceted interaction process. Moreover, the demonstrated contrast between objective gaze behaviour and subjective perception thereof adds to the understanding of the complex relationship between cognition and behaviour. This reciprocal relationship, where cognition shapes user behaviour and user behaviour informs cognition – through the use of different physical and environmental contexts – emphasises the relevance of comprehensive theoretical frameworks that encompass both behavioural responses and cognitive processes. This not only reduces saliency bias and optimises visual attention modelling, but furthermore provides a multifaceted understanding of the observed behavioural responses and gaze dynamics in VR. In addition, the insights gathered on the significant influence of temporally complex 360-degree video carries significant implications for the challenges of bit-rate variability in temporally complex videos, as implied by Afzal et al. (2017) [2].

As before-mentioned, the implications for the use of the index transcends the scope of this thesis as it is a universally adaptable tool for gaze behaviour analysis, applicable to any heatmap image signal. Specifically in the field of human-computer interaction and cognitive science, the index reduces saliency-bias by providing a more accurate representation of a user's visual attention and fixation patterns [217, 358, 359]. As such, the index can contribute significantly to the enhancement of foveated rendering techniques and could be instrumental in addressing the challenges of viewport prediction in VR research [39, 101, 145, 237, 360]. Furthermore, the technical advancements that could be made in foveated rendering techniques and viewport prediction ensure higher sensorimotor contingencies, which aids in restoring any potential disruption of place and plausibility illusion [108, 227, 294, 297]. As such, the implications extend to practical applications as well to ensure more more immersive and engaging virtual 360-degree environments. The practical implications of this thesis are discussed in § 6.2.

Furthermore, the quadrifactorial exploration index N_ψ could be employed within the field of machine learning and utilised for artificial intelligence applications that focus on image recognition

and computer vision, as the index can be utilised to improve model prediction to better capture human visual attention. As such, this thesis demonstrates the efficacy of integrating various related domains, such as computer vision, human-computer interaction, cognitive science and film studies, into a singular holistic approach.

This work also contributes to instrumentation within the field of human-computer interaction. A key contribution from this thesis is the novel formulation of the quadrifactorial exploration index N_ψ. The index enables a more advanced and comprehensive understanding of gaze behaviour in virtual 360-degree environments. As opposed to traditional gaze metrics, predominantly focused on the duration and location of fixations, the quadrifactorial exploration index utilises the heatmap imagery as extracted from advanced eye-tracking software. The provision and development of a singular Python script (see Appendix B10), which integrates and implements various image segmentation techniques, enhances accessibility to conducting research within the field of oculesics. The index enhances the interpretation and application of heatmaps in eye-tracking studies and provides a quantitative interpretive framework, reducing the technical barrier and elevating the field.

Moreover, this thesis introduced an initial quantification methodology to measure the influence of diegesis on user attention. The novel metric δ represents the presence of visual artefacts that are inherent to the internal story- or scene structure of the 360-degree video sequence, based on visual size and duration. Coupled with the quadrifactorial exploration index N_ψ, this work provides an initial understanding of how the narrative mechanisms in 360-degree video sequences guide user attention in VR. The diegetic assessment revealed the confounding influence of cinematographic principles on shaping gaze behaviour. The initial insights on the behavioural impact of diegesis could motivate further research into the complex dynamics of cinematography and human-computer interaction. It informs cinematographers and researchers within the field of film studies on the significant behavioural impact of internal story- or scene elements on users in VR. Moreover, the use of manual annotation in the current methodology necessitates the development of an apt computer vision-based framework capable of modelling the specific cinematographic visual influences more accurately. As a result, a real-time diegetic assessment could be realised and the framework could provide more nuanced assessments, not solely focusing on the influence of diegesis but take other cinematographic principles into account. Such a framework could encourage more interdisciplinary collaboration between the fields of human-computer interaction, computer science, cognitive science and film theory.

Ultimately, this work emphasises the importance of content-aware and interdisciplinary approaches in VR- and 360-degree research, and introduces instrumental tools such as the quadrifactorial exploration index N_ψ and the diegetic metric δ. This thesis provides a comprehensive understanding of the interplay of content, usability and cognition to elicit behavioural gaze responses. As such, it carries a multitude of implications that could set new theoretical directions in VR research. Interesting possibilities for future research, based on this thesis, are discussed in § 6.4.

6.2 Implications for Practice

The findings and insights derived from this thesis carry significant implications for practical applications in the various related domains. One of the most prominent contributions of this work lies in the development and creation of 360-degree video content in VR, as the use of spatiotemporal image complexity to quantify 360-degree content is not limited to the specific videos used in this research, but is applicable to any 360-degree video sequence. The demonstrated significant role of spatiotemporal image complexity on gaze dynamics can be utilised to improve user attention guidance. The discerned negative correlation of temporal image complexity with gaze behaviour can be employed in the design of videos that effectively direct attention, as the users tend to concentrate their focus when viewing rapidly changing visual information such as fast camera motions. Omnidirectional video content developers ought to take into account this information during development, as users tend to be less exploratory in temporally complex environments.

Moreover, this work demonstrated that spatial image complexity does not significantly impact gaze behaviour as an independent factor, implying the cautious consideration of spatially complex environments since an increase in exploration is not guaranteed. Importantly, employing more spatially complex environments could potentially be detrimental to the overall user experience, due to the increased cognitive load. As such, this finding could potentially reshape the design and composition of spatially complex virtual environments. Another significant contribution of this work was the demonstration of the significance and important role of usability context, as the findings illustrate the interaction effects between using a different seating type and the impact of spatiotemporal image complexity. These findings were utilised to devise a set design principles for 360-degree content development and creation, which are presented in § 6.2.1.

Moreover, the practical implications transcend beyond the scope of this thesis as professionals can utilise these insights into how users interact with the virtual 360-degree environment and use it to optimise the design of virtual environments aimed at impacting the users sense of presence and immersion. As such, this work carries substantial implications for other domain-specific applications aside from 360-degree video content creation and development as well. Despite entertainment and gaming being the most popular domains, the use of 360-degree video also becomes more prominent in other domains, such as education, telepresence and infotainment [343, 351]. Omnidirectional video is currently utilised within a range of different domains. As identified by Pirker et al. (2021), the vast majority of 360-degree video is applied in medicine healthcare (28.1%), 20.3% is used in STEM subject, 7.8% is in engineering and 4.7% is in computer science [244, 362]. Geology, history and social studies and general education adds up to 18.8% of 360-degree video application. Additionally, the use of 360-degree video in career training and teacher education accumulates to approximately 9.3%.

Specifically, in the domain of education, 360-degree videos could enhance the learning process, leading to increased levels of performance, motivation and knowledge retention [160, 244, 259]. As such, the findings of this thesis could be utilised the tailor the presentation of information and guide the user's attention to specific information. For instance, these insights could lead to a more optimised design of virtual classrooms, which present the information in the most optimal way. Similarly, when employed in the field of healthcare, could lead to more effective and beneficial therapy sessions by designing more immersive experiences. In particular, the findings could be used to control where patients look and help manage the exposure to triggering content in exposure therapy for PTSD or phobias.

Moreover, this work adds to the overall understanding how users interact with virtual 360-degree environments, which could inform the development of more immersive and engaging virtual reality experiences in the field of gaming and entertainment. As such, it could lead to enhancement of existing instructional guides for storytelling in virtual reality such as "The Storyteller's Guide to the Virtual Reality Audience" or other comparable (frame)works [43, 66, 84, 89, 139, 353].

6.2.1 Set of Design Principles

Utilising the findings of this thesis, a set of design principles was devised to provide actionable insights for the creation of 360-degree content development and creation. The design principles emphasise the role of both temporal- and spatial complexity, as well as usability context, spatial awareness, quality of experience and cognitive load. Most importantly, the principles can be utilised to adhere to the director's intent: ensuring the visual information is conveyed to the users as intended [223, 262]. The design principles provide a comprehensive understanding on how the director can regulate the visual elements of a 360-degree scene and adjust the visual properties to facilitate a more salient area in the preferred direction [99, 118].

- Guide user attention through temporal complexity; the negative correlation between temporal complexity and distribution of gaze behaviour can be exploited to both encourage exploration of the 360-degree environment, as well increase the user's focus on specific areas. The exploration of 360-degree of content can be encouraged by limiting the amount of visual change happening over time, such as camera motion and scene changes. On the other hand,

by increasing the temporal complexity of a 360-degree video sequence, the user's gaze will be more concentrated and centred. This principle can be utilised to increase focus on specific area's of visual information.

- Regulate spatial complexity to mitigate (cognitive) overload; the non-significant independent impact of spatial image complexity on gaze behaviour emphasises the need to cautiously regulate and balance 360-degree video scenes, as gaze exploration is not merely encouraged by enhancing the scenes with more visual details. The identified dichotomy between objective and subjective gaze behaviour also emphasises the cautious regulation of spatial image complexity, as users already feel that they explore more than they objectively do in spatially complex virtual 360-degree environments. However, it's important not to underestimate the importance of spatial complexity, as in conjunction with usability, physical, environmental and cognitive factors, spatial complexity does leverage a certain effect on gaze distribution. For instance, in case of viewing both scenery and video game 360-degree video sequences: despite the less spatially complex environment of the computer-generated video game graphics, underlying preference and user interest may still result in an unexpected increase in gaze exploration. Therefore, it is important to focus on the interplay between spatial complexity, usability context and cognitive load to optimise user engagement.

- Tailoring usability context to accommodate user behaviour; the significant impact of a specific seating type on gaze behaviour can be utilised to encourage or guide user attention. Specifically, rotating chairs can be utilised to enable and encourage users to explore more of the virtual 360-degree environment. In contrast, a fixed-position chair could be utilised to limit the users range of motion and to realise a higher degree of focus on specific AOIs. As such, different usability contexts can be utilised to evoke specific user behaviour. Moreover, the independent influence of spatial image complexity is moderated by different usability contexts. For example, when utilising a fixed chair, increasing spatial complexity could encourage more gaze exploration.

- Facilitating higher spatial awareness and QoE; the significant effect of spatial awareness and quality of experience on gaze behaviour could be exploited to encourage more exploratory behaviour. For instance, the use of visual cues to guide users through the virtual 360-degree environment could be used to create content that enhances spatial awareness. Similarly, user-friendly interfaces and high-quality graphics could generate a higher quality of experience among users. As such, by taking into account the factors of spatial awareness and QoE, these techniques could aid in encouraging more exploratory gaze behaviour.

The research, as presented in this thesis, provides insights into optimising 360-degree content in VR, emphasising the intricate impact of spatiotemporal image complexity, cognition, and usability context on shaping gaze behaviour. The devised set of design principles can be utilised to guide the development of 360-degree videos and in the design of more immersive and engaging virtual 360-degree environments. As such, the practical implications apply significantly within the field of human-computer interaction and resonate to a range of application domains such as education, health care, entertainment, engineering, marketing, cognition and cinematography.

6.3 Limitations

The multifaceted research methodology introduced aided in discerning the extent of which spatiotemporal image complexity of a 360-degree video sequence in VR influences gaze dynamics. However, in the context of interpreting the results, it is important to acknowledge the limitations of the employed methodologies and novel metrics of this thesis.

One of the most notable limitations of this thesis is regarding the novel methodological challenges. The quadrifactorial exploration index N_ψ, as introduced in this work, remains a novel instrument to quantify gaze patterns based on heatmap image signals from advanced eye-tracking software. Despite the assessment of its reliability in § 3.4, the overall adaptability and external

validity of the metric is still irresolute due to its novel nature. Moreover, its novel nature means its sensitive to biases and oversights. In addition, the use of MS-SSIM depends on high quality image signals, which should be considered in the adoption of the index in future studies. Similarly, the dimensions chosen for the diegetic assessment coding scheme were arbitrary. While the examination of the association between diegetic artefacts and gaze distribution was exploratory in nature, the use of visual size and duration to quantify the presence of diegetic artefacts might not be sufficient for larger scale and in-depth diegetic analyses, as it does not take into account the nuances of VR cinematography.

Furthermore, the subjective evaluations were conducted after the viewing session of all six 360-degree video sequences to omit the need for users to switch between virtual reality and real-life. This methodology heavily relies on the user's ability to recall experiences, and also prevented any video-specific subjective evaluations. As implied by McCarthy et al. (2004), this stimulated recall presents limitations to the interpretation, specifically in situations of intense cognitive activity as it might not reflect the exact nature of the experience [93, 205]. In addition, the employment of the M-ACR methodology to measure the QoE for each participant across all six 360-degree videos introduces potential bias. Given that the user was exposed to a short fragment of each utilised 360-degree video beforehand, there was a risk of priming that may have influenced their subsequent interaction and evaluation. Furthermore, cautious interpretation is advised in terms of model performance and interpretation of the results, as some of the statistical assumptions were not met. The statistical model's performance did also rely on the level of independent variables, indicating a need for additional non-linear models to better assess the goodness of fit.

Another limitation of this thesis relates to the database of the selected 360-degree videos. While the content was carefully filtered and selected, as detailed in § 2.2.3, the use of self-produced 360-degree content would have enabled a more controlled examination of the influence of specific spatiotemporal image complexities and configurations. Moreover, despite their efficiency in providing a sufficient coverage of the spatiotemporal matrix (Figure 4), it is important to emphasise that genre and camera motion should be considered basic indicators. This method might be oversimplified for spatiotemporal indication and filtering of 360-degree video databases based on spatiotemporal image complexity.

Lastly, the use of the purposive sampling method introduces potential selection bias and could limit the external validity of the results. Additionally, the research design did not take into account any cultural or demographic differences, remaining predominantly focused on the level of experience with VR. Moreover, the individual differences, such as different experiences with VR and susceptibility to motion sickness could result in varied gaze behaviour. Despite the efforts made to mitigate the risk of cybersickness or physical discomfort, by the implementation of a separate pre-test parameter study (see § 2.1.1) and selective sampling, the potential of such risks occurring during the experimental procedure still remains.

In conclusion, the discussed limitations emphasise the areas of consideration and provides a context in which the findings of this thesis can be interpreted. The detailed limitations also present opportunities for refinement and potential directions for future research.

6.4 Future Research

The research findings and implications from this thesis contribute various insights to the field, as well as a comprehensive understanding of the impact of 360-degree content as an independent factor in the VR user interaction process. This section discusses the potential directions for future research, building on the findings and presented limitations of this work. In particular, this section discusses future research in the variability of spatiotemporal image complexities, the diversification of usability contexts, employment of longitudinal studies, in-depth examinations of the cinematographic and cognitive influences and the enhancement of the quadrifactorial exploration index.

This research utilised a specific set of six 360-degree videos, which varied in spatiotemporal image complexity, representing sufficient coverage of the spatiotemporal matrix. However, fu-

ture research could expand the analysis by examining the gaze dynamics by leveraging additional spatiotemporal image complexities. By increasing the variability of spatiotemporal image complexities, the gaze dynamics could be more accurately analysed, mapping the nuances of gaze behaviour across more spatiotemporal configurations and exceeding the specific spatial- and temporal complexity value-ranges of this work. In addition, building upon the work by Cui et al. (2021) which illustrated correlations between variations in spatiotemporal complexity and specific genre-characteristics, future research could exploit the spatiotemporal matrix to devise a comprehensive framework which systematically catalogues particular spatiotemporal configurations with respective genre-characteristics [72, 172, 290].

The significant influence of usability context on the effect of spatial- and temporal image complexity on gaze behaviour also presents interesting possibilities for future research. In particular, the nuanced influence of spatially complex virtual 360-degree environments could be further examined by focusing on elucidating the intricacies between various spatially complex environments such as natural landscapes or computer-generated graphics. By examining how differences in spatial complexities influence gaze patterns, a more comprehensive understanding on the variable significant influence of spatial image complexity on gaze behaviour could be provided. Furthermore, the different usability contexts in this research were represented by two seating types: rotating and fixed-position. Expanding the usability context to involve physical movement, such as walking, can provide valuable insights. This presents research opportunities in diversifying the current study parameters, such as taking varying age groups and multi-modal output devices into consideration.

Since the research conducted in this thesis was a cross-sectional study, it could be very interesting for future research to conduct longitudinal studies. Specifically, longitudinal studies could focus on the learning and technology adaptation processes in repeated 360-degree viewing sessions in VR, examining how gaze dynamics vary over multiple sessions [133, 146, 229, 335]. Adding to this, future research could study in-depth the effects of cognitive fatigue and deploy different strategies to mitigate this fatigue and maintain engagement in longer-length VR viewing sessions. The influence of cognitive perception in this research also presents other potential research opportunities, as future research could try to better understand the underlying cognitive influences on gaze patterns. In particular, this could help elucidate the observed non-significant impact of the level of engagement as well as the negative association between QoE and gaze behaviour, which are in stark contrast with theoretical indications [205, 230, 376]. In addition, future studies could focus on how gaze dynamics are impacted by user fatigue when exposed to longer-length 360-degree videos, extending beyond the maximum 60 seconds duration of each 360-degree video in this research. Another interesting research direction could be to implement rewatchability in the study, similar to the work by Singla et al. (2017) [290]. Consequently, the longevity of the effect size of specific spatiotemporal complexities on gaze dynamics over repeated viewing sessions could be examined. In particular, how the effect size increases or decreases when users are more familiarised with the content.

Lastly, this work introduced the novel metric N_ψ, the quadrifactorial exploration index, which was instrumental in the quantification of gaze patterns in VR. By utilising image segmentation techniques and data-driven approaches, the index proved a reliable and elegant metric to represent the degree of gaze distribution based on heatmap image signals extracted from advanced eye-tracking software. Future research can exploit this metric with further refinement and by studying its adaptability to different research contexts, as it has already demonstrated its potential in quantifying gaze behaviour in VR. The novel nature of the metric enables future research to further iterate its formulation, providing a range of potential future research direction. For instance, the metric could be augmented by utilising it in conjunction with advanced computer vision techniques for automatic 360-degree scene analysis. By automatically extracting features from the 360-degree video frames, such as colour, texture, edges, or other visually salient elements, the index could be utilised to examine the gaze dynamics more in-depth and take into account even more complex narrative- and scene structures. Additionally, the index could be integrated with complementary physiological measures such as EEG signals and heart rate. Furthermore, neural networks could be utilised to analyse and improve model predictions on gaze behaviour based

on the respective spatiotemporal image complexities of 360-degree video content. Aside from the multi-modal integration and deep learning possibilities, future research should predominantly focus on the improvement of the metric due to its relatively novel nature. By studying the variability of the index more in-depth, across a larger sample size and additional respective spatiotemporal image complexities, the quadrifactorial exploration index N_{ψ} could become even more comprehensive and versatile in capturing the complexity of gaze behaviour patterns during 360-degree video interactions.

Conclusion

Over recent decades, virtual reality technology has significantly transformed the domain of interactive virtual experiences, shifting the paradigms of human-computer interaction. The emergence and ongoing commercialisation of 360-degree video technology, coupled with the substantial advances of multi-modal interaction methods, has presented an intriguingly diverse field of study. Motivated by its relative under-representation within the existing body of literature, in which research on the 360-degree interaction process resides within the boundaries of technical limitations, this work elucidated the pivotal role of 360-degree content as an independent factor within the interaction model. This thesis studied the 360-degree video sequence in terms of its spatiotemporal image complexity, as quantifiable representations of the content in dimensions of space and time, where the spatial complexity denotes the visual richness of the 360-degree video sequence and the temporal complexity reflects the change in visual information over time.

The primary research objective was to discern the influence of spatiotemporal image complexity on gaze behaviour in VR, while factoring in the complex dynamics of cognitive perceptions and usability context. A content-aware and user-centric approach was adopted, further enhanced with theoretical insights from the related fields of cognitive science, computer vision and film studies. This holistic approach, bridging both theoretical foundations and practical applications, integrated advanced methodologies from the field of computer vision, oculesics, and human-computer interaction to examine the autonomous and independent role of 360-degree video sequences within the multifaceted interaction model.

As such, this thesis employed physiological HMD eye-tracking, objective computer vision techniques, and subjective evaluations to reveal the significance of spatiotemporal complexity on gaze dynamics in VR. Despite theoretical implications, spatial image complexity demonstrated a nuanced interaction effect, primarily governed by underlying cognitive influences as highlighted by the observed dichotomy between objective gaze data and subjective experiences. The temporal complexity of a 360-degree video sequence emerged as a significant factor, negatively impacting the extent of user gaze. The usability context, as assessed through different seating types, further modulated these effects.

Instrumental to this thesis was the formulation and introduction of the novel quadrifactorial exploration index N_ψ, a measure of the degree of gaze distribution during the 360-degree video interaction. The index elegantly integrates image segmentation techniques from the field of computer vision with gaze distribution heatmap imagery, acquired from advanced eye-tracking software, to quantify complex gaze patterns. The index encapsulates the spatial extent, concentration, diversity and randomness of gaze distribution patterns, transforming traditional eye-tracking heatmaps into a quantifiable, multi-dimensional perspective of gaze behaviour and alleviates the need for subjective interpretations. In addition, grounded by the theoretical significance of cinematographic influence, this work also provided an initial framework – symbolised by δ – for examining the confounding influence of diegetic artefacts in 360-degree video sequences on gaze behaviour.

In conclusion, this work contributes to the field of VR research by providing a comprehensive understanding of the pivotal role of 360-degree content within the multi-dimensional interaction process. By employing spatiotemporal image complexity, this thesis elucidates the autonomous and independent influence of 360-degree content on eliciting specific gaze patterns. Evaluated against the backdrop of cognitive and cinematographic influences, as well as across usability contexts, the findings not only highlight the complex interplay of content-specific attributes, cognition, usability

and gaze dynamics in VR, but also reveal the potentiality of integrating oculesics and computer vision.

The theoretical and practical implications from this work provide a substantiated framework for the development and optimisation of immersive 360-degree videos and virtual environments. Additionally, they offer actionable insights into tailored strategies that exploit the dynamics of spatiotemporal image complexity to effectively guide user attention. By emphasising the intricate dynamics of content-specific attributes, cognitive perceptions, and usability contexts, this work revealed the interdisciplinary research possibilities of content-aware approaches, as well as sets a precedent for exploring the unique properties of 360-degree video content within the domain.

Bibliography

[1] Edward H Adelson, Charles H Anderson, James R Bergen, Peter J Burt, and Joan M Ogden. Pyramid methods in image processing. *RCA engineer*, 29(6):33–41, 1984. 23

[2] Shahryar Afzal, Jiasi Chen, and KK Ramakrishnan. Characterization of 360-degree videos. In *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, pages 1–6, 2017. 1, 7, 8, 9, 29, 30, 40, 127, 130

[3] A Deniz Aladagli, Erhan Ekmekcioglu, Dmitri Jarnikov, and Ahmet Kondoz. Predicting head trajectories in 360 virtual reality videos. In *2017 International Conference on 3D Immersion (IC3D)*, pages 1–6. IEEE, 2017. 13

[4] Seyed Ali Amirshahi and M-C Larabi. Spatial-temporal video quality metric based on an estimation of qoe. In *2011 Third International Workshop on Quality of Multimedia Experience*, pages 84–89. IEEE, 2011. 19

[5] Peter A Andersen. Eye behavior. *The International Encyclopedia of Interpersonal Communication*, pages 1–7, 2015. 14

[6] John R Anderson, Dan Bothell, and Scott Douglass. Eye movements do not reflect retrieval processes: Limits of the eye-mind hypothesis. *Psychological Science*, 15(4):225–231, 2004. 14, 38

[7] Ioannis Arapakis, Mounia Lalmas, B Barla Cambazoglu, Mari-Carmen Marcos, and Joemon M Jose. User engagement in online news: Under the scope of sentiment, interest, affect, and gaze. *Journal of the Association for Information Science and Technology*, 65(10):1988–2005, 2014. 30, 31

[8] ARInsider. Will VR revenue exceed $22 billion by 2025? *AR Insider*, Dec 2021. 6, 40

[9] Pablo Arnau-González, Turke Althobaiti, Stamos Katsigiannis, and Naeem Ramzan. Perceptual video quality evaluation by means of physiological signals. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017. 12, 30, 34, 49, 61, 64, 123, 130

[10] Sebastian Arndt, Jenni Radun, Jan-Niklas Antons, and Sebastian Möller. Using eye-tracking and correlates of brain activity to predict quality scores. In *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 281–285. IEEE, 2014. 12, 34, 38

[11] Jay D Aronson. Computer vision and machine learning for human rights video analysis: Case studies, possibilities, concerns, and limitations. *Law & Social Inquiry*, 43(4):1188–1209, 2018. 18

[12] R Venkatesh Babu, Patrick Perez, and Patrick Bouthemy. Robust tracking with motion estimation and local kernel-based color modeling. *Image and Vision computing*, 25(8):1205–1216, 2007. 18

[13] Jeremy N Bailenson, Jim Blascovich, Andrew C Beall, and Jack M Loomis. Interpersonal distance in immersive virtual environments. *Personality and social psychology bulletin*, 29(7):819–833, 2003. 7

[14] Illya Bakurov, Marco Buzzelli, Raimondo Schettini, Mauro Castelli, and Leonardo Vanneschi. Structural similarity index (SSIM) revisited: A data-driven approach. *Expert Systems with Applications*, 189:116087, 2022. 21, 23, 24, 75

[15] Paulo Bala, Mara Dionisio, Valentina Nisi, and Nuno Nunes. Ivrux: A tool for analyzing immersive narratives in virtual reality. In *Interactive Storytelling: 9th International Conference on Interactive Digital Storytelling, ICIDS 2016, Los Angeles, CA, USA, November 15–18, 2016, Proceedings 9*, pages 3–11. Springer, 2016. 16

[16] Paulo Bala, Valentina Nisi, and Nuno Nunes. Evaluating user experience in 360º storytelling through analytics. In *Interactive Storytelling: 10th International Conference on Interactive Digital Storytelling, ICIDS 2017 Funchal, Madeira, Portugal, November 14–17, 2017, Proceedings 10*, pages 270–273. Springer, 2017. 16, 62

[17] Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, 26(11):5217–5231, 2017. 19, 30

[18] Xiaojuan Ban, Xiaolong Lv, and Jie Chen. Color image retrieval and classification using fuzzy similarity measure and fuzzy clustering method. In *Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with 2009 28th Chinese Control Conference*, pages 7777–7782. IEEE, 2009. 21, 75

[19] Rosa M Baños, Cristina Botella, Isabel Rubió, Soledad Quero, Azucena García-Palacios, and Mariano Alcañiz. Presence and emotions in virtual environments: The influence of stereoscopy. *CyberPsychology & Behavior*, 11(1):1–8, 2008. 29, 30

[20] Rosa María Baños, Cristina Botella, Mariano Alcañiz, Víctor Liaño, Belén Guerrero, and Beatriz Rey. Immersion and emotion: their impact on the sense of presence. *Cyberpsychology & behavior*, 7(6):734–741, 2004. 29, 30

[21] Yanan Bao, Huasen Wu, Albara Ah Ramli, Bradley Wang, and Xin Liu. Viewing 360 degree videos: Motion prediction and bandwidth optimization. In *2016 IEEE 24th International Conference on Network Protocols (ICNP)*, pages 1–2. IEEE, 2016. 1, 16

[22] Yanan Bao, Huasen Wu, Tianxiao Zhang, Albara Ah Ramli, and Xin Liu. Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 1161–1170. IEEE, 2016. 13

[23] Yanan Bao, Tianxiao Zhang, Amit Pande, Huasen Wu, and Xin Liu. Motion-prediction-based multicast for 360-degree video transmissions. In *2017 14th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*, pages 1–9. IEEE, 2017. 27, 86, 123

[24] Charles Bell. XV. On the motions of the eye, in illustration of the uses of the muscles and nerves of the orbit. *Philosophical Transactions of the Royal Society of London*, (113):166–186, 1823. 14

[25] Neha Bhargava and Fabio Cuzzolin. Challenges and opportunities for computer vision in real-life soccer analytics. *arXiv:2004.06180*, 2020. 19

[26] Simone Bianco, Luigi Celona, and Paolo Napoletano. Disentangling image distortions in deep feature space. *Pattern Recognition Letters*, 148:128–135, 2021. 21

[27] Simone Bianco, Luigi Celona, and Flavio Piccoli. Single image dehazing by predicting atmospheric scattering parameters. In *London Imaging Meeting*, volume 2020, pages 74–77. Society for Imaging Science and Technology, 2020. 21

[28] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 25, 124, 128

[29] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. State-of-the-art of visualization for eye tracking data. In *EuroVis (STARs)*, 2014. 16, 17

[30] Thomas Blass. Understanding behavior in the milgram obedience experiment: The role of personality, situations, and their interactions. *Journal of personality and social psychology*, 60(3):398, 1991. 28

[31] Lizzy Bleumers, Wendy Van den Broeck, Bram Lievens, and Jo Pierson. Seeing the bigger picture: a user perspective on 360 tv. In *Proceedings of the 10th European conference on Interactive TV and video*, pages 115–124, 2012. 29, 30, 31

[32] Agnieszka Bojko. Informative or misleading? heatmaps deconstructed. In *Human-Computer Interaction. New Trends: 13th International Conference, HCI International 2009, San Diego, CA, USA, July 19-24, 2009, Proceedings, Part I 13*, pages 30–39. Springer, 2009. 16

[33] Sorana D Bolboacă and Lorentz Jäntschi. Design of experiments: Useful orthogonal arrays for number of experiments from 4 to 16. *Entropy*, 9(4):198–232, 2007. 48

[34] David Bordwell, Kristin Thompson, and Jeff Smith. *Film art: An introduction*, volume 7. McGraw-Hill New York, 1993. 25

[35] James V Bradley. Complete counterbalancing of immediate sequential effects in a latin square design. *Journal of the American Statistical Association*, 53(282):525–528, 1958. 48

[36] Johan Braeken and Marcel ALM Van Assen. An empirical Kaiser criterion. *Psychological methods*, 22(3):450, 2017. 79

[37] Steve Branson, Grant Van Horn, Serge Belongie, and Pietro Perona. Bird species categorization using pose normalized deep convolutional nets. *arXiv:1406.2952*, 2014. 17

[38] Peter Broadwell, Tomoko Bialock, and Hiroyuki Ikuura. Macroscopic exploration of large text and image collections via similarity heatmaps. *JADH 2017*, page 1, 2017. 19

[39] Marc Van den Broeck, Fahim Kawsar, and Johannes Schöning. It's all around you: Exploring 360 video viewing experiences on mobile devices. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 762–768, 2017. 29, 30, 130

[40] Susan Bruck and Paul A Watters. Estimating cybersickness of simulated motion using the simulator sickness questionnaire (ssq): A controlled study. In *2009 sixth international conference on computer graphics, imaging and visualization*, pages 486–488. IEEE, 2009. 35

[41] Dominique Brunet, Edward R Vrscay, and Zhou Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2011. 21

[42] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, et al. Qualinet white paper on definitions of quality of experience. 2013. 10, 30, 35, 50, 52, 123, 125

[43] John Bucher. *Storytelling for virtual reality: Methods and principles for crafting immersive narratives*. Routledge, 2017. 132

[44] David C Burr, M Concetta Morrone, and John Ross. Selective suppression of the magnocellular visual pathway during saccadic eye movements. *Nature*, 371(6497):511–513, 1994. 15

[45] Guy Thomas Buswell. How people look at pictures: a study of the psychology and perception in art. 1935. 15

[46] Martin Cadik and Pavel Slavik. Evaluation of two principal approaches to objective image quality assessment. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pages 513–518. IEEE, 2004. 75

[47] Frank Candocia and Malek Adjouadi. A similarity measure for stereo feature matching. *IEEE transactions on Image Processing*, 6(10):1460–1464, 1997. 18

[48] Chong Cao, Zhaowei Shi, and Miao Yu. Automatic generation of diegetic guidance in cinematic virtual reality. In *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 600–607. IEEE, 2020. 27, 28, 86

[49] Benjamin T Carter and Steven G Luke. Best practices in eye tracking research. *International Journal of Psychophysiology*, 155:49–62, 2020. 15, 38, 59

[50] Eric Castet, Sébastien Jeanjean, and Guillaume S Masson. Motion perception of saccade-induced retinal translation. *Proceedings of the National Academy of Sciences*, 99(23):15159–15163, 2002. 15

[51] Kevin M Chambers, Bhaskar S Mandavilli, Nick J Dolman, and Michael S Janes. General staining and segmentation procedures for high content imaging and analysis. *High Content Screening: A Powerful Approach to Systems Cell Biology and Phenotypic Drug Discovery*, pages 21–31, 2018. 18

[52] Peter McFaul Chapman. *Models of engagement: Intrinsically motivated interaction with multimedia learning software*. PhD thesis, University of Waterloo, 1997. 29

[53] Manri Cheon and Jong-Seok Lee. Temporal resolution vs. visual saliency in videos: Analysis of gaze patterns and evaluation of saliency models. *Signal Processing: Image Communication*, 39:405–417, 2015. 14

[54] Federico Chiariotti. A survey on 360-degree video: Coding, quality of experience and streaming. *Computer Communications*, 177:133–155, 2021. 1, 6, 9, 14, 40, 62

[55] Jaeseob Choi, Donghyun Kim, Bumsub Ham, Sunghwan Choi, and Kwanghoon Sohn. Visual fatigue evaluation and enhancement for 2d-plus-depth video. In *2010 IEEE International Conference on Image Processing*, pages 2981–2984. IEEE, 2010. 19, 30

[56] Marc Christie, Rumesh Machap, Jean-Marie Normand, Patrick Olivier, and Jonathan Pickering. Virtual camera planning: A survey. In *Smart Graphics: 5th International Symposium, SG 2005, Frauenwörth Cloister, Germany, August 22-24, 2005. Proceedings 5*, pages 40–52. Springer, 2005. 25

[57] Sunaina K Chugani, Julie R Irwin, and Joseph P Redden. Happily ever after: The effect of identity-consistency on product satiation. *Journal of Consumer Research*, 42(4):564–577, 2015. 29

[58] Alasdair DF Clarke, Aoife Mahon, Alex Irvine, and Amelia R Hunt. People are unable to recognize or report on their own eye movements. *Quarterly Journal of Experimental Psychology*, 70(11):2251–2270, 2017. 15, 38

[59] Alex Clarke, Kirsten I Taylor, and Lorraine K Tyler. The evolution of meaning: spatio-temporal dynamics of visual object recognition. *Journal of cognitive neuroscience*, 23(8):1887–1899, 2011. 19, 30

[60] Norman Cliff. The eigenvalues-greater-than-one rule and the reliability of components. *Psychological bulletin*, 103(2):276, 1988. 79

[61] Simon Colton, Michel F Valstar, and Maja Pantic. Emotionally aware automated portrait painting. In *Proceedings of the 3rd international conference on Digital Interactive Media in Entertainment and Arts*, pages 304–311, 2008. 18

[62] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou. Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation. *IEEE Transactions on Image Processing*, 27(2):568–579, 2017. 18

[63] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 360-degree video head movement dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 199–204, 2017. 27, 86, 123

[64] Xavier Corbillon, Gwendal Simon, Alisa Devlic, and Jacob Chakareski. Viewport-adaptive navigable 360-degree video delivery. In *2017 IEEE international conference on communications (ICC)*, pages 1–7. IEEE, 2017. 1, 6, 7

[65] Valve Corporation. Steam VR, 2023. 55

[66] Ana Rita Jesus Costa. Storytelling for cinematic virtual reality: Audience's needs and expectations. 2018. 132

[67] Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951. 81

[68] Emily M Crowe, Iain D Gilchrist, and Christopher Kent. New approaches to the analysis of eye movement behaviour across expertise while viewing brain mris. *Cognitive research: principles and implications*, 3:1–14, 2018. 19

[69] James P Crutchfield. Spatio-temporal complexity in nonlinear image processing. *IEEE transactions on circuits and systems*, 35(7):770–780, 1988. 19, 30

[70] Mihaly Csikszentmihalyi. Flow. *The psychology of optimal experience*, pages 1–22, 1990. 11

[71] Mihaly Csikszentmihalyi. Flow and the psychology of discovery and invention. *HarperPerennial, New York*, 39:1–16, 1997. 11

[72] Ruifang Cui, Jinliang Jiang, Lu Zeng, Lijun Jiang, Zeling Xia, Li Dong, Diankun Gong, Guojian Yan, Weiyi Ma, and Dezhong Yao. Action video gaming experience related to altered resting-state eeg temporal and spatial complexity. *Frontiers in Human Neuroscience*, page 365, 2021. 19, 21, 30, 41, 75, 135

[73] James J Cummings and Jeremy N Bailenson. How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media psychology*, 19(2):272–309, 2016. 7

[74] Dragoş Datcu, Stephan Lukosch, and Frances Brazier. On the usability and effectiveness of different interaction types in augmented reality. *International Journal of Human-Computer Interaction*, 31(3):193–209, 2015. 29

[75] Fred D Davis. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly*, pages 319–340, 1989. 11

[76] Fred D Davis, Richard P Bagozzi, and Paul R Warshaw. User acceptance of computer technology: A comparison of two theoretical models. *Management science*, 35(8):982–1003, 1989. 10, 49, 52, 64

[77] Edmund B Delabarre. A method of recording eye-movements. *The American Journal of Psychology*, 9(4):572–574, 1898. 15

[78] Frank Dellaert, Sebastian Thrun, and Chuck Thorpe. Jacobian images of super-resolved texture maps for model-based motion estimation and tracking. In *Proceedings Fourth IEEE Workshop on Applications of Computer Vision. WACV'98 (Cat. No. 98EX201)*, pages 2–7. IEEE, 1998. 18

[79] K Ding and S Gunasekaran. Shape feature extraction and classification of food material using computer vision. *Transactions of the ASAE*, 37(5):1537–1545, 1994. 18

[80] C Distler and K-P Hoffmann. The optokinetic reflex. 2011. 15

[81] Florin Dobrian, Vyas Sekar, Asad Awan, Ion Stoica, Dilip Joseph, Aditya Ganjam, Jibin Zhan, and Hui Zhang. Understanding the impact of video quality on user engagement. *ACM SIGCOMM computer communication review*, 41(4):362–373, 2011. 29, 35, 52, 124

[82] Richard Dosselmann and Xue Dong Yang. A comprehensive assessment of the structural similarity index. *Signal, Image and Video Processing*, 5:81–91, 2011. 24, 75

[83] Edward Dougherty. *Mathematical morphology in image processing*, volume 1. CRC press, 2018. 18

[84] Stanford d.school. The storyteller's guide to the virtual reality audience, 2016. 132

[85] Fanyi Duanmu, Eymen Kurdoglu, S Amir Hosseini, Yong Liu, and Yao Wang. Prioritized buffer control in two-tier 360 video streaming. In *Proceedings of the Workshop on Virtual Reality and Augmented Reality Network*, pages 13–18, 2017. 13

[86] Andrew T Duchowski and Andrew T Duchowski. Diversity and types of eye tracking applications. *Eye Tracking Methodology: Theory and Practice*, pages 247–248, 2017. 15, 38, 128

[87] Andrew T Duchowski, Eric Medlin, Nathan Cournia, Anand Gramopadhye, Brian Melloy, and Santosh Nair. 3d eye movement analysis for VR visual inspection training. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 103–110, 2002. 14

[88] Andrew T Duchowski, Margaux M Price, Miriah Meyer, and Pilar Orero. Aggregate gaze visualization with real-time heatmaps. In *Proceedings of the symposium on eye tracking research and applications*, pages 13–20, 2012. 16, 17

[89] Richard Scott Dunham and Richard Scott Dunham. Artificial intelligence, virtual reality and computer-driven storytelling. *Multimedia Reporting: How Digital Tools Can Improve Journalism Storytelling*, pages 355–367, 2020. 132

[90] Touradj Ebrahimi. Quality of multimedia experience: past, present and future. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 3–4, 2009. 9, 30, 35, 38, 49, 50, 52, 55, 124, 125, 130

[91] Darragh Egan, Sean Brennan, John Barrett, Yuansong Qiao, Christian Timmerer, and Niall Murray. An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments. In *2016 eighth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2016. 12, 13, 34, 38, 49, 127

[92] Ulrich Engelke, Daniel P Darcy, Grant H Mulliken, Sebastian Bosse, Maria G Martini, Sebastian Arndt, Jan-Niklas Antons, Kit Yan Chan, Naeem Ramzan, and Kjell Brunnström. Psychophysiology-based qoe assessment: A survey. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):6–21, 2016. 12

[93] K Anders Ericcson and Herbert A Simon. Protocol analysis: Verbal reports as data (rev. ed.). *Cambridge, MA: Bradford*, 1993. 50, 134

[94] Laura Ermi and F Mayra. Fundamental components of the gameplay experience: Analyzing immersion, changing views: Worlds in play. selected papers of the 2005 digital games research association's second international conference. *Online], http://www.uta.fi/˜frans.mayra/gameplay_experience. pdf [11 August 2009]*, pages 15–27, 2005. 13

[95] Irfan A. Essa and Alex Paul Pentland. Coding, analysis, interpretation, and recognition of facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 19(7):757–763, 1997. 18

[96] Majid Fakheri, Tohid Sedghi, Mahrokh G Shayesteh, and Mehdi Chehel Amirani. Framework for image retrieval using machine learning and statistical similarity matching techniques. *IET Image Processing*, 7(1):1–11, 2013. 18, 19

[97] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. Fixation prediction for 360 video streaming in head-mounted virtual reality. In *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video*, pages 67–72, 2017. 13

[98] Yuyuan Fang, Jun Hu, Chuan Du, Zhibo Liu, and Lei Zhang. Sar-optical image matching by integrating siamese u-net with fft correlation. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021. 19

[99] Colm O Fearghail, Sebastian Knorr, and Aljosa Smolic. Analysis of intended viewing area vs estimated saliency on narrative plot structures in VR film. In *2019 International Conference on 3D Immersion (IC3D)*, pages 1–8. IEEE, 2019. 17, 25, 31, 128, 132

[100] Colm O Fearghail, Cagri Ozcinar, Sebastian Knorr, and Aljosa Smolic. Director's cut-analysis of aspects of interactive storytelling for VR films. In *Interactive Storytelling: 11th International Conference on Interactive Digital Storytelling, ICIDS 2018, Dublin, Ireland, December 5–8, 2018, Proceedings*, pages 308–322. Springer, 2018. 27, 86

[101] Xianglong Feng, Viswanathan Swaminathan, and Sheng Wei. Viewport prediction for live 360-degree mobile video streaming using user-content hybrid motion tracking. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(2):1–22, 2019. 14, 30, 130

[102] Diana Fonseca and Martin Kraus. A comparison of head-mounted and hand-held displays for 360 videos with focus on attitude and behavior change. In *Proceedings of the 20th International Academic Mindtrek Conference*, pages 287–296, 2016. 29, 30

[103] Daniel Freeman, Philippa A Garety, Paul Bebbington, Mel Slater, Elizabeth Kuipers, David Fowler, Catherine Green, Joel Jordan, Katarzyna Ray, and Graham Dunn. The psychology of persecutory ideation II: a virtual reality experimental study. *The Journal of nervous and mental disease*, 193(5):309–315, 2005. 28

[104] Daniel Freeman, Mel Slater, Paul E Bebbington, Philippa A Garety, Elizabeth Kuipers, David Fowler, Alican Met, Cristina M Read, Joel Jordan, and Vinoba Vinayagamoorthy. Can virtual reality be used to investigate persecutory ideation? *The Journal of nervous and mental disease*, 191(8):509–514, 2003. 28

[105] Stephan Fremerey, Frank Hofmeyer, Steve Göring, and Alexander Raake. Impact of various motion interpolation algorithms on 360° video qoe. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2019. 2, 12, 19, 30, 38, 39

[106] Fei Gao, Huibin Jia, Xiangci Wu, Dongchuan Yu, and Yi Feng. Altered resting-state eeg microstate parameters and enhanced spatial complexity in male adolescent patients with mild spastic diplegia. *Brain topography*, 30:233–244, 2017. 19

[107] Yang Gao, Abdul Rehman, and Zhou Wang. CW-SSIM based image classification. In *2011 18th IEEE International Conference on Image Processing*, pages 1249–1252. IEEE, 2011. 23, 75

[108] Maia Garau, Doron Friedman, Hila Ritter Widenfeld, Angus Antley, Andrea Brogni, and Mel Slater. Temporal and spatial variations in presence: Qualitative analysis of interviews from an experiment on breaks in presence. *Presence: Teleoperators and Virtual Environments*, 17(3):293–309, 2008. 28, 130

[109] M-N Garcia, Francesca De Simone, Samira Tavakoli, Nicolas Staelens, Sebastian Egger, Kjell Brunnström, and Alexander Raake. Quality of experience and http adaptive streaming: A review of subjective studies. In *2014 sixth international workshop on quality of multimedia experience (qomex)*, pages 141–146. IEEE, 2014. 9, 30

[110] Vahid Ghodrati, Jiaxin Shao, Mark Bydder, Ziwu Zhou, Wotao Yin, Kim-Lien Nguyen, Yingli Yang, and Peng Hu. Mr image reconstruction using deep learning: evaluation of network structure and loss functions. *Quantitative imaging in medicine and surgery*, 9(9):1516, 2019. 23

[111] Arnob Ghosh, Vaneet Aggarwal, and Feng Qian. A rate adaptation algorithm for tile-based 360-degree video streaming. *arXiv:1704.08215*, 2017. 9

[112] Dion Hoe-Lian Goh, Chei Sian Lee, and Khasfariyati Razikin. Interfaces for accessing location-based information on mobile devices: An empirical evaluation. *Journal of the Association for Information Science and Technology*, 67(12):2882–2896, 2016. 29

[113] Jie Gong and Carlos H Caldas. Computer vision-based video interpretation model for automated productivity analysis of construction operations. *Journal of Computing in Civil Engineering*, 24(3):252–263, 2010. 18

[114] Claudia Gorbman. *Unheard melodies: Narrative film music*. Indiana University Press, 1987. 27

[115] Mario Graf, Christian Timmerer, and Christopher Mueller. Towards bandwidth efficient adaptive streaming of omnidirectional video over http: Design, implementation, and evaluation. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 261–271, 2017. 7

[116] Thomas J Grindinger, Vidya N Murali, Stephen Tetreault, Andrew T Duchowski, Stan T Birchfield, and Pilar Orero. Algorithm for discriminating aggregate gaze points: comparison with salient regions-of-interest. In *Computer Vision–ACCV 2010 Workshops: ACCV 2010 International Workshops, Queenstown, New Zealand, November 8-9, 2010, Revised Selected Papers, Part I 10*, pages 390–399. Springer, 2011. 17, 38, 128

[117] Kwangsung Ha and Munchurl Kim. A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video. In *2011 18th IEEE International Conference on Image Processing*, pages 2525–2528. IEEE, 2011. 19, 30

[118] Aiko Hagiwara, Akihiro Sugimoto, and Kazuhiko Kawamoto. Saliency-based image editing for guiding visual attention. In *Proceedings of the 1st international workshop on pervasive eye tracking & mobile eye-based interaction*, pages 43–48, 2011. 31, 132

[119] Edward T Hall. The silent language garden city. *NY: Doubleday*, 240, 1959. 28, 86

[120] Robert M Haralick, Stanley R Sternberg, and Xinhua Zhuang. Image analysis using mathematical morphology. *IEEE transactions on pattern analysis and machine intelligence*, (4):532–550, 1987. 18

[121] Gary M Hardee and Ryan P McMahan. FIJI: a framework for the immersion-journalism intersection. *Frontiers in ICT*, 4:21, 2017. 29, 127

[122] Mohammed Hassan and Chakravarthy Bhagvati. Structural similarity measure for color images. *International Journal of Computer Applications*, 43(14):7–12, 2012. 23, 75

[123] Marc Hassenzahl and Noam Tractinsky. User experience: A research agenda. *Behaviour & information technology*, 25(2):91–97, 2006. 29

[124] Jiale He, Gaobo Yang, Xin Liu, and Xiangling Ding. Spatio-temporal saliency-based motion vector refinement for frame rate up-conversion. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–18, 2020. 19

[125] Bernhard JM Hess. Vestibular response. 2011. 15

[126] Lawrence J Hettinger and Gary E Riccio. Visually induced motion sickness in virtual environments. *Presence: Teleoperators & Virtual Environments*, 1(3):306–310, 1992. 13, 35, 36

[127] CC Heyde. Central limit theorem. *Encyclopedia of actuarial science*, 1, 2006. 60

[128] Tobias Hobfeld, Raimund Schatz, Martin Varela, and Christian Timmerer. Challenges of qoe management for cloud applications. *IEEE Communications Magazine*, 50(4):28–36, 2012. 10

[129] Kenneth Holmqvist and R Andersson. Eye tracking: A comprehensive guide to methods. *paradigms and measures*, 2017. 16, 123

[130] Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost Van de Weijer. *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford, 2011. 49, 127

[131] Alain Hore and Djemel Ziou. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010. 21, 75, 123

[132] Tobias Hoßfeld, Michael Seufert, Christian Sieber, and Thomas Zinner. Assessing effect sizes of influence factors towards a qoe model for http adaptive streaming. In *2014 sixth international workshop on quality of multimedia experience (qomex)*, pages 111–116. IEEE, 2014. 9, 30, 130

[133] Chin-Lung Hsu and Hsi-Peng Lu. Why do people play online games? An extended TAM with social influences and flow experience. *Information & management*, 41(7):853–868, 2004. 11, 135

[134] Ching-Ting Hsu, Chia-Hung Yeh, Chao-Yu Chen, and Mei-Juan Chen. Arbitrary frame rate transcoding through temporal and spatial complexity. *IEEE Transactions on Broadcasting*, 55(4):767–775, 2009. 19

[135] HTC. Vive Eye Tracking SDK, 2023. 55

[136] Yuxiang Hu, Yu Liu, and Yumei Wang. Vas360: Qoe-driven viewport adaptive streaming for 360 video. In *2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 324–329. IEEE, 2019. 13

[137] Isabelle Hupont, Joaquin Gracia, Luis Sanagustin, and Miguel Angel Gracia. How do new visual immersive systems influence gaming qoe? a use case of serious gaming with oculus rift. In *2015 Seventh international workshop on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2015. 12, 13, 36, 39

[138] Jacopo Iannacci. Reliability of mems: A perspective on failure mechanisms, improvement solutions and best practices at development level. *Displays*, 37:62–71, 2015. 13

[139] Jesús Ibanez, Ruth Aylett, and Rocio Ruiz-Rodarte. Storytelling in virtual environments from a virtual guide perspective. *Virtual Reality*, 7:30–42, 2003. 132

[140] iMotions A/S. iMotions, 2023. 54, 55

[141] Poika Isokoski, Jari Kangas, and Päivi Majaranta. Useful approaches to exploratory analysis of gaze data: enhanced heatmaps, cluster maps, and transition maps. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–9, 2018. 16

[142] Kenji Itoh, Jean-Pierre Hansen, and FR Nielsen. Cognitive modelling of a ship navigator based on protocol and eye-movement analysis. *Le Travail Humain*, pages 99–127, 1998. 14

[143] Kenji Itoh, Hiromasa Tanaka, and Masaki Seki. Eye-movement analysis of track monitoring patterns of night train operators: Effects of geographic knowledge and fatigue. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 44, pages 360–363. SAGE Publications Sage CA: Los Angeles, CA, 2000. 14

[144] P ITU-T RECOMMENDATION. Subjective video quality assessment methods for multimedia applications. 1999. 11, 14, 19, 20, 39, 43, 44, 45, 122

[145] Susmija Jabbireddy, Xuetong Sun, Xiaoxu Meng, and Amitabh Varshney. Foveated rendering: Motivation, taxonomy, and research directions. *arXiv:2205.04529*, 2022. 14, 30, 130

[146] Cynthia M Jackson, Simeon Chow, and Robert A Leitch. Toward an understanding of the behavioral intention to use an information system. *Decision sciences*, 28(2):357–389, 1997. 11, 135

[147] Richard Jacques. Engagement as a design concept for multimedia. *Canadian Journal of Educational Communication*, 24(1):49–59, 1995. 29

[148] Richard David Jacques. *The nature of engagement and its role in hypermedia evaluation and design.* PhD thesis, South Bank University, 1996. 51

[149] Anil K Jain and Stan Z Li. *Handbook of face recognition*, volume 1. Springer, 2011. 18

[150] Eakta Jain, Yaser Sheikh, Ariel Shamir, and Jessica Hodgins. Gaze-driven video re-editing. *ACM Transactions on Graphics (TOG)*, 34(2):1–12, 2015. 25

[151] Ramesh Jain. Quality of experience, IEEE Multimedia, 2004. 9, 30, 130

[152] Charlene Jennett, Anna L Cox, Paul Cairns, Samira Dhoparee, Andrew Epps, Tim Tijs, and Alison Walton. Measuring and defining the experience of immersion in games. *International journal of human-computer studies*, 66(9):641–661, 2008. 13

[153] Oliver P John, Richard W Robins, and Lawrence A Pervin. *Handbook of personality: Theory and research.* Guilford Press, 2010. 10

[154] Cheryl I Johnson and Richard E Mayer. A testing effect with multimedia learning. *Journal of Educational Psychology*, 101(3):621, 2009. 10

[155] Maria Jorquera-Chavez, Sigfredo Fuentes, Frank R Dunshea, Ellen C Jongman, and Robyn D Warner. Computer vision and remote sensing to assess physiological responses of cattle to pre-slaughter stress, and its impact on beef quality: A review. *Meat science*, 156:11–22, 2019. 19

[156] Roy S Kalawsky et al. The validity of presence as a reliable human performance metric in immersive environments. In *3rd International Workshop on Presence*, pages 1–16, 2000. 13, 43

[157] Leon A Kappelman. Measuring user involvement: A diffusion of innovation perspective. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, 26(2-3):65–86, 1995. 29

[158] Shunichi Kasahara, Shohei Nagai, and Jun Rekimoto. First person omnidirectional video: System design and implications for immersive experience. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*, pages 33–42, 2015. 13, 36

[159] Hirokatsu Kataoka, Kenji Iwata, and Yutaka Satoh. Feature evaluation of deep convolutional neural networks for object recognition and detection. *arXiv:1509.07627*, 2015. 18

[160] Sam Kavanagh, Andrew Luxton-Reilly, Burkhard Wüensche, and Beryl Plimmer. Creating 360 educational video: A case study. In *Proceedings of the 28th Australian conference on computer-human interaction*, pages 34–39, 2016. 29, 132

[161] Michal Kawulok, Emre Celebi, and Bogdan Smolka. *Advances in face detection and facial image analysis*. Springer, 2016. 18

[162] A Donald Keedwell and József Dénes. *Latin squares and their applications*. Elsevier, 2015. 48

[163] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993. 37, 207

[164] Tuuli Keskinen, Ville Mäkelä, Pekka Kallioniemi, Jaakko Hakulinen, Jussi Karhu, Kimmo Ronkainen, John Mäkelä, and Markku Turunen. The effect of camera height, actor behavior, and viewer position on the user experience of 360 videos. In *2019 IEEE conference on virtual reality and 3D user interfaces (VR)*, pages 423–430. IEEE, 2019. 28, 86

[165] Cory D Kidd, Robert Orr, Gregory D Abowd, Christopher G Atkeson, Irfan A Essa, Blair MacIntyre, Elizabeth Mynatt, Thad E Starner, and Wendy Newstetter. The aware home: A living laboratory for ubiquitous computing research. In *Cooperative Buildings. Integrating Information, Organizations, and Architecture: Second International Workshop, CoBuild'99, Pittsburgh, PA, USA, October 1-2, 1999. Proceedings 2*, pages 191–198. Springer, 1999. 18

[166] Young Youn Kim, Hyun Ju Kim, Eun Nam Kim, Hee Dong Ko, and Hyun Taek Kim. Characteristic changes in the physiological components of cybersickness. *Psychophysiology*, 42(5):616–625, 2005. 13

[167] Kwang-Eun Ko and Kwee-Bo Sim. Development of a facial emotion recognition method based on combining aam with dbn. In *2010 International Conference on Cyberworlds*, pages 87–91. IEEE, 2010. 18

[168] Matthew J Koehler, Aman Yadav, Michael Phillips, and Sean Cavazos-Kottke. What is video good for? examining how media and story genre interact. *Journal of Educational Multimedia and Hypermedia*, 14(3):249–272, 2005. 29

[169] Ellen M Kok, Avi M Aizenman, Melissa L-H Võ, and Jeremy M Wolfe. Even if i showed you where you looked, remembering where you just looked is hard. *Journal of Vision*, 17(12):2–2, 2017. 15, 38

[170] H Kolb, E Fernandez, and R Nelson. Facts and figures concerning the human retina-webvision. *The Organization of the Retina and Visual System.[Google Scholar]*, 1995. 15

[171] Istvan Kondakor, Marton Toth, Jiri Wackermann, Csilla Gyimesi, Jozsef Czopf, and Bela Clemens. Distribution of spatial complexity of eeg in idiopathic generalized epilepsy and its change after chronic valproate therapy. *Brain topography*, 18:115–123, 2005. 19

[172] Baris Konuk, Emin Zerman, Gokce Nur, and Gozde Bozdagi Akar. A spatiotemporal no-reference video quality assessment model. In *2013 IEEE International Conference on Image Processing*, pages 54–58. Ieee, 2013. 19, 34, 56, 135

[173] Richard J Krauzlis, Laurent Goffart, and Ziad M Hafed. Neuronal control of fixation and fixational eye movements. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1718):20160205, 2017. 15

[174] S Shunmuga Krishnan and Ramesh K Sitaraman. Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs. In *Proceedings of the 2012 Internet Measurement Conference*, pages 211–224, 2012. 9

[175] Christopher A Kurby and Jeffrey M Zacks. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79, 2008. 25, 128

[176] Sang Gyu Kwak and Jong Hae Kim. Central limit theorem: the cornerstone of modern statistics. *Korean journal of anesthesiology*, 70(2):144–156, 2017. 60

[177] Dmitry Lagun and Mounia Lalmas. Understanding user attention and engagement in online news reading. In *Proceedings of the ninth ACM international conference on web search and data mining*, pages 113–122, 2016. 29, 30

[178] Nina Siu-Ngan Lam, Hong-lie Qiu, Dale A Quattrochi, and Charles W Emerson. An evaluation of fractal methods for characterizing image complexity. *Cartography and Geographic Information Science*, 29(1):25–35, 2002. 19

[179] Wei Guang Lan, Ming Keong Wong, Ni Chen, and Yoke Min Sin. Orthogonal array design as a chemometric method for the optimization of analytical procedures. part 1. two-level design and its application in microwave dissolution of biological samples. *Analyst-letchworth*, 119(8):1659–1668, 1994. 48

[180] Zhengzhong Lan, Ming Lin, Xuanchong Li, Alex G Hauptmann, and Bhiksha Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 204–212, 2015. 23

[181] Michelle A LaRue, Seth Stapleton, and Morgan Anderson. Feasibility of using high-resolution satellite imagery to assess vertebrate wildlife populations. *Conservation biology*, 31(1):213–220, 2017. 17

[182] Brenda Laurel. Computers as theatre reading. *Mas: Addison-Wesley Publishing Company*, 1991. 29

[183] Stephen Leary, Atul Bhaskar, and Andy Keane. Optimal orthogonal-array-based latin hypercubes. *Journal of Applied Statistics*, 30(5):585–598, 2003. 48

[184] Shen Li, Yong Jiang, Takeshi Ikenaga, and Satoshi Goto. Content-based motion estimation with extended temporal-spatial analysis. *IEICE transactions on information and systems*, 88(7):1561–1568, 2005. 19

[185] Yiming Li, Jizheng Xu, and Zhenzhong Chen. Spherical domain rate-distortion optimization for 360-degree video coding. In *2017 IEEE international conference on multimedia and expo (ICME)*, pages 709–714. IEEE, 2017. 9

[186] Yin Li, Alireza Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Proceedings of the IEEE international conference on computer vision*, pages 3216–3223, 2013. 44

[187] Yan Liang, An Kang, Tong Xie, Xiao Zheng, Chen Dai, Haiping Hao, A Jiye, Longsheng Sheng, Lin Xie, and Guang-ji Wang. Influence of segmental and selected ion monitoring on quantitation of multi-component using high-pressure liquid chromatography–quadrupole mass spectrometry: Simultaneous detection of 16 saponins in rat plasma as a case. *Journal of Chromatography A*, 1217(26):4501–4506, 2010. 23

[188] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2535–2545, 2017. 29, 30

[189] Yung-Ta Lin, Yi-Chi Liao, Shan-Yuan Teng, Yi-Ju Chung, Liwei Chan, and Bing-Yu Chen. Outside-in: Visualizing out-of-sight regions-of-interest in a 360 video using spatial picture-in-picture previews. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 255–265, 2017. 29

[190] Åsa Linder. Key factors for feeling present during a music experience in virtual reality using 360 video, 2017. 29

[191] CH Lo and Alan Chalmers. Stereo vision for computer graphics: the effect that stereo vision has on human judgments of visual realism. In *Proceedings of the 19th spring conference on Computer Graphics*, pages 109–117, 2003. 18

[192] Matthew Lombard and Matthew T Jones. Defining presence. *Immersed in media: Telepresence theory, measurement & technology*, pages 13–34, 2015. 13, 43

[193] Lester C Loschky, Adam M Larson, Joseph P Magliano, and Tim J Smith. What would jaws do? The tyranny of film and the relationship between gaze and higher-level narrative film comprehension. *PloS one*, 10(11):e0142474, 2015. 1, 24, 86

[194] Thomas Löwe, Michael Stengel, Emmy-Charlotte Förster, Steve Grogorick, and Marcus Magnor. Visualization and analysis of head movement and gaze data for immersive video in head-mounted displays. In *Proceedings of the workshop on eye tracking and visualization (ETVIS)*, volume 1, 2015. 16, 17, 30, 34, 35, 52, 67

[195] Jane F Mackworth and Norman H Mackworth. Eye fixations recorded on changing visual scenes by the television eye-marker. *JOSA*, 48(7):439–445, 1958. 16, 17

[196] Andrew MacQuarrie and Anthony Steed. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *2017 IEEE Virtual Reality (VR)*, pages 45–54. IEEE, 2017. 53, 127

[197] Joseph P Magliano and Jeffrey M Zacks. The impact of continuity editing in narrative film on event segmentation. *Cognitive science*, 35(8):1489–1517, 2011. 25, 128

[198] Ilias Maglogiannis, Demosthenes Vouyioukas, and Chris Aggelopoulos. Face detection and recognition of natural human emotion using markov random fields. *Personal and Ubiquitous Computing*, 13:95–101, 2009. 18

[199] T Mahalakshmi, R Muthaiah, and P Swaminathan. Image processing. *Research Journal of Applied Sciences, Engineering and Technology*, 4(24):5469–5473, 2012. 18

[200] Francesco Marchetti. Convergence rate in terms of the continuous SSIM (cSSIM) index in rbf interpolation. *Dolomites Research Notes on Approximation*, 14(1), 2021. 23

[201] Fran Marquis-Faulkes, Stephen J McKenna, Peter Gregor, and Alan Newell. Scenario-based drama as a tool for investigating user requirements with application to home monitoring for elderly people. In *Human-Centered Computing*, pages 512–516. CRC Press, 2019. 18

[202] Masterclass. Film 101: What is cinematography and what does a cinematographer do?, 2021. 25, 124

[203] Aditya Mavlankar and Bernd Girod. Video streaming with interactive pan/tilt/zoom. *High-Quality Visual Experience: Creation, Processing and Interactivity of High-Resolution and High-Dimensional Video Signals*, pages 431–455, 2010. 13

[204] Richard E Mayer. Incorporating motivation into multimedia learning. *Learning and instruction*, 29:171–173, 2014. 10

[205] John McCarthy and Peter Wright. Technology as experience. *interactions*, 11(5):42–43, 2004. 20, 29, 50, 51, 52, 86, 124, 134, 135

[206] Lars Meinel, Markus Hess, Michel Findeisen, and Gangolf Hirtz. Effective display resolution of 360 degree video footage in virtual reality. In *2017 IEEE International Conference on Consumer Electronics (ICCE)*, pages 21–24. IEEE, 2017. 7

[207] Miguel Melo, Sofia Sampaio, Luís Barbosa, José Vasconcelos-Raposo, and Maximino Bessa. The impact of different exposure times to 360 video experience on the sense of presence. In *2016 23rd Portuguese Meeting on Computer Graphics and Interaction (EPCGI)*, pages 1–5. IEEE, 2016. 29, 127

[208] Alex Mihailidis, Joseph C Barbenel, and Geoff Fernie. The efficacy of an intelligent cognitive orthosis to facilitate handwashing by persons with moderate to severe dementia. *Neuropsychological Rehabilitation*, 14(1-2):135–171, 2004. 18

[209] Alex Mihailidis, Brent Carmichael, and Jennifer Boger. The use of computer vision in an intelligent environment to support aging-in-place, safety, and independence in the home. *IEEE Transactions on information technology in biomedicine*, 8(3):238–247, 2004. 18

[210] Alex Mihailidis, Geoffrey R Fernie, and Joseph C Barbenel. The use of artificial intelligence in the design of an intelligent cognitive orthosis for people with dementia. *Assistive Technology*, 13(1):23–39, 2001. 18

[211] Ravina Mithe, Supriya Indalkar, and Nilam Divekar. Optical character recognition. *International journal of recent technology and engineering (IJRTE)*, 2(1):72–75, 2013. 18

[212] Sebastian Möller. *Quality Engineering: Qualität kommunikationstechnischer Systeme*. Springer-Verlag, 2017. 10

[213] Wayne S Murray, Martin H Fischer, and Benjamin W Tatler. Serial and parallel processes in eye movement control: Current controversies and future directions. *Quarterly Journal of Experimental Psychology*, 66(3):417–428, 2013. 14, 38

[214] Mohammad Nabil, Anne HH Ngu, and John Shepherd. Picture similarity retrieval using the 2d projection interval representation. *IEEE Transactions on Knowledge and Data Engineering*, 8(4):533–539, 1996. 21, 75

[215] Afshin Taghavi Nasrabadi, Anahita Mahzari, Joseph D Beshay, and Ravi Prakash. Adaptive 360-degree video streaming using layered video coding. In *2017 IEEE Virtual Reality (VR)*, pages 347–348. IEEE, 2017. 7

[216] Peter Ndajah, Hisakazu Kikuchi, Masahiro Yukawa, Hidenori Watanabe, and Shogo Muramatsu. SSIM image quality metric for denoised images. In *Proc. 3rd WSEAS Int. Conf. on Visualization, Imaging and Simulation*, pages 53–58, 2010. 21

[217] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1190–1198, 2018. 13, 130

[218] Tan Bao Nguyen and Djemel Ziou. Contextual and non-contextual performance evaluation of edge detectors. *Pattern Recognition Letters*, 21(9):805–816, 2000. 75

[219] Jakob Nielsen and Kara Pernice. *Eyetracking web usability*. New Riders, 2010. 15

[220] Lasse T Nielsen, Matias B Møller, Sune D Hartmeyer, Troels CM Ljung, Niels C Nilsson, Rolf Nordahl, and Stefania Serafin. Missing the point: an exploration of how to guide users' attention during cinematic virtual reality. In *Proceedings of the 22nd ACM conference on virtual reality software and technology*, pages 229–232, 2016. 27, 28, 86

[221] Jim Nilsson and Tomas Akenine-Möller. Understanding ssim. *arXiv:2006.13846*, 2020. 21, 23, 24, 75

[222] Mark Nixon and Alberto Aguado. *Feature extraction and image processing for computer vision*. Academic press, 2019. 17, 18

[223] Nahal Norouzi, Gerd Bruder, Austin Erickson, Kangsoo Kim, Jeremy Bailenson, Pamela Wisniewski, Charlie Hughes, and Greg Welch. Virtual animals as diegetic attention guidance mechanisms in 360-degree experiences. *IEEE Transactions on Visualization and Computer Graphics*, 27(11):4321–4331, 2021. 1, 27, 30, 31, 44, 51, 52, 53, 124, 127, 132

[224] Nahal Norouzi, Kangsoo Kim, Gerd Bruder, Austin Erickson, Zubin Choudhary, Yifan Li, and Greg Welch. A systematic literature review of embodied augmented reality agents in head-mounted display environments. In *Proceedings of the International Conference on Artificial Reality and Telexistence & Eurographics Symposium on Virtual Environments*, 2020. 27, 86

[225] David Noton and Lawrence Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, 1971. 16, 17

[226] Thomas P Novak, Donna L Hoffman, and Yiu-Fai Yung. Measuring the customer experience in online environments: A structural modeling approach. *Marketing science*, 19(1):22–42, 2000. 11, 51

[227] Alva Noë. Action in perception cambridge. MA MIT Press. 2004. 28, 86, 130

[228] Marcus Nyström and Kenneth Holmqvist. Effect of compressed offline foveated video on viewing behavior and subjective quality. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 6(1):1–14, 2010. 14

[229] Herbjørn Nysveen, Per E Pedersen, and Helge Thorbjørnsen. Intentions to use mobile services: Antecedents and cross-service comparisons. *Journal of the academy of marketing science*, 33(3):330–346, 2005. 11, 135

[230] Heather L O'Brien and Elaine G Toms. What is user engagement? a conceptual framework for defining user engagement with technology. *Journal of the American society for Information Science and Technology*, 59(6):938–955, 2008. 20, 29, 38, 43, 50, 51, 52, 53, 86, 124, 135

[231] Reza Oji. An automatic algorithm for object recognition and detection based on asift keypoints. *arXiv:1211.5829*, 2012. 18

[232] Bobbie O'Steen. *The invisible cut.* Michael Wiese Productions, 2009. 25

[233] ITU-T Recommendation ITU-T P.910. Subjective video quality assessment methods for multimedia applications. 2008. 11, 39

[234] Meenakshi Panwar and Pawan Singh Mehra. Hand gesture recognition for human computer interaction. In *2011 International Conference on Image Information Processing*, pages 1–7. IEEE, 2011. 18

[235] ALEXANDRU Păsărică, Radu Gabriel Bozomitu, DANIELA Tărniceriu, Gladiola Andruseac, Hariton Costin, and Cristian Rotariu. Analysis of eye image segmentation used in eye tracking applications. *Rev. Roum. Sci. Tech*, 62:215–222, 2017. 18

[236] Peter J Passmore, Maxine Glancy, Adam Philpot, Amelia Roscoe, Andrew Wood, and Bob Fields. Effects of viewing condition on user experience of panoramic video. 2016. 29

[237] Anjul Patney, Marco Salvi, Joohwan Kim, Anton Kaplanyan, Chris Wyman, Nir Benty, David Luebke, and Aaron Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016. 14, 15, 128, 130

[238] Amy Pavel, Björn Hartmann, and Maneesh Agrawala. Shot orientation controls for interactive cinematography with 360 video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, pages 289–297, 2017. 29

[239] Henry B Peters. Vision screening with a snellen chart. *Optometry and Vision Science*, 38(9):487–505, 1961. 39

[240] Thies Pfeiffer. Measuring and visualizing attention in space with 3d attention volumes. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 29–36, 2012. 16

[241] Michael Reynaldo Phangtriastu, Jeklin Harefa, and Dian Felita Tanoto. Comparison between neural network and support vector machine in optical character recognition. *Procedia computer science*, 116:351–357, 2017. 18

[242] Adam Philpot, Maxine Glancy, Peter J Passmore, Andrew Wood, and Bob Fields. User experience of panoramic video in cave-like and head mounted display viewing conditions. In *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, pages 65–75, 2017. 29

[243] Martin J Pickering, Steven Frisson, Brian McElree, and Matthew J Traxler. Eye movements and semantic composition. In *The on-line study of sentence comprehension*, pages 33–50. Psychology Press, 2004. 14, 38

[244] Johanna Pirker and Andreas Dengel. The potential of 360 virtual reality videos and real VR for education—a literature review. *IEEE computer graphics and applications*, 41(4):76–89, 2021. 132

[245] Felix Platter. *De corporis humani structura et usu libri III.* ap. Lud. König, 1583. 15

[246] Marc Pomplun, Helge Ritter, and Boris Velichkovsky. Disambiguating complex visual information: Towards communication of personal views of a scene. *Perception*, 25(8):931–948, 1996. 16

[247] Feng Qian, Lusheng Ji, Bo Han, and Vijay Gopalakrishnan. Optimizing 360 video delivery over cellular networks. In *Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges*, pages 1–6, 2016. 13

[248] Rui Qian, Robby T Tan, Wenhan Yang, Jiajun Su, and Jiaying Liu. Attentive generative adversarial network for raindrop removal from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2482–2491, 2018. 21

[249] Chaiyong Ragkhitwetsagul, Jens Krinke, and Bruno Marnette. A picture is worth a thousand words: Code clone detection based on image similarity. In *2018 IEEE 12th International workshop on software clones (IWSC)*, pages 44–50. IEEE, 2018. 21, 75

[250] Fitri N Rahayu, Ulrich Reiter, Touradj Ebrahimi, Andrew Perkis, and Peter Svensson. SS-SSIM and MS-SSIM for digital cinema applications. In *Human Vision and Electronic Imaging XIV*, volume 7240, pages 212–223. SPIE, 2009. 23, 75

[251] Rameshsharma Ramloll, Cheryl Trepagnier, Marc Sebrechts, and Jaishree Beedasy. Gaze data visualization tools: opportunities and challenges. In *Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004.*, pages 173–180. IEEE, 2004. 16

[252] Abhishek Ranjan, Jeremy P Birnholtz, and Ravin Balakrishnan. An exploratory analysis of partner action and camera control in a video-mediated collaborative task. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 403–412, 2006. 11

[253] Keith Rayner. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506, 2009. 15

[254] Keith Rayner, Albrecht Werner Inhoff, Robert E Morrison, Maria L Slowiaczek, and James H Bertera. Masking of foveal and parafoveal vision during eye fixations in reading. *Journal of Experimental Psychology: Human perception and performance*, 7(1):167, 1981. 15

[255] ITUTP Recommendation. 10/g. 100 amendment 2: New definitions for inclusion in recommendation itu-t p. 10/g. 100. *Vocabulary for performance and quality of service*, 2008. 10

[256] Erik D Reichle, Andrew E Reineberg, and Jonathan W Schooler. Eye movements during mindless reading. *Psychological science*, 21(9):1300–1310, 2010. 14, 38

[257] Jeremy R Reynolds, Jeffrey M Zacks, and Todd S Braver. A computational model of event segmentation from perceptual prediction. *Cognitive science*, 31(4):613–643, 2007. 25

[258] John TE Richardson. The use of latin-square designs in educational and psychological research. *Educational Research Review*, 24:84–97, 2018. 48

[259] Giuseppe Riva. Virtual reality for health care: the status of research. *Cyberpsychology & Behavior*, 5(3):219–225, 2002. 132

[260] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001. 12, 34

[261] Martin Rolfs. Attention in active vision: A perspective on perceptual continuity across saccades. *Perception*, 44(8-9):900–919, 2015. 15

[262] Sylvia Rothe, Daniel Buschek, and Heinrich Hußmann. Guidance in cinematic virtual reality-taxonomy, research status and challenges. *Multimodal Technologies and Interaction*, 3(1):19, 2019. 1, 24, 26, 27, 28, 30, 31, 44, 52, 53, 86, 127, 128, 132

[263] Sylvia Rothe and Heinrich Hußmann. Guiding the viewer in cinematic virtual reality by diegetic cues. In *Augmented Reality, Virtual Reality, and Computer Graphics: 5th International Conference, AVR 2018, Otranto, Italy, June 24–27, 2018, Proceedings, Part I 5*, pages 101–117. Springer, 2018. 27, 86, 123

[264] Virpi Ed Roto. User experience white paper. *http://www. allaboutux. org/uxwhitepaper*, 2011. 10

[265] David M Rouse and Sheila S Hemami. Analyzing the role of visual structure in the recognition of natural image content with multi-scale SSIM. In *Human vision and electronic imaging XIII*, volume 6806, pages 410–423. SPIE, 2008. 23, 75

[266] Leonardo Rundo, Andrea Tangherloni, Paolo Cazzaniga, Marco S Nobile, Giorgio Russo, Maria Carla Gilardi, Salvatore Vitabile, Giancarlo Mauri, Daniela Besozzi, and Carmelo Militello. A novel framework for mr image segmentation and quantification by using medga. *Computer methods and programs in biomedicine*, 176:159–172, 2019. 21

[267] Michael A Rupp, James Kozachuk, Jessica R Michaelis, Katy L Odette, Janan A Smither, and Daniel S McConnell. The effects of immersiveness and future VR expectations on subjective-experiences during an educational 360 video. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 60, pages 2108–2112. SAGE Publications Sage CA: Los Angeles, CA, 2016. 29, 127

[268] Michael S Ryoo and Jake K Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *2009 IEEE 12th international conference on computer vision*, pages 1593–1600. IEEE, 2009. 19

[269] Santeri Saarinen, Ville Mäkelä, Pekka Kallioniemi, Jaakko Hakulinen, and Markku Turunen. Guidelines for designing interactive omnidirectional video applications. In *Human-Computer Interaction–INTERACT 2017: 16th IFIP TC 13 International Conference, Mumbai, India, September 25-29, 2017, Proceedings, Part IV 16*, pages 263–272. Springer, 2017. 28, 86

[270] Norma S Said. An engaging multimedia design model. In *Proceedings of the 2004 conference on Interaction design and children: building a community*, pages 169–172, 2004. 29, 38

[271] Débora Pereira Salgado, Felipe Roque Martins, Thiago Braga Rodrigues, Conor Keighrey, Ronan Flynn, Eduardo Lázaro Martins Naves, and Niall Murray. A qoe assessment method based on eda, heart rate and eeg of a virtual reality assistive technology system. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pages 517–520, 2018. 12, 34, 38

[272] Khena MA Sallow and Jeffrey M Zacks. Event segmentation. Current directions in psychological science. 14(2):80–84, 2007. 25, 128

[273] Mehul P Sampat, Zhou Wang, Shalini Gupta, Alan Conrad Bovik, and Mia K Markey. Complex wavelet structural similarity: A new image similarity index. *IEEE transactions on image processing*, 18(11):2385–2401, 2009. 23, 75

[274] Ana Luisa Sánchez Laws. Can immersive journalism enhance empathy? *Digital journalism*, 8(2):213–228, 2020. 29

[275] Maria V Sanchez-Vives and Mel Slater. From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, 6(4):332–339, 2005. 1, 6

[276] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019. 23

[277] Raimund Schatz, Andreas Sackl, Christian Timmerer, and Bruno Gardlo. Towards subjective quality of experience assessment for omnidirectional video streaming. In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2017. 12, 38

[278] Christian M Schulz, Erich Schneider, L Fritz, Johannes Vockeroth, Alexander Hapfelmeier, Thomas Brandt, EF Kochs, and G Schneider. Visual attention of anaesthetists during simulated critical incidents. *British journal of anaesthesia*, 106(6):807–813, 2011. 14

[279] Philip Sedgwick and Nan Greenwood. Understanding the Hawthorne effect. *Bmj*, 351, 2015. 56, 57

[280] Daniel Senkowski, Till R Schneider, John J Foxe, and Andreas K Engel. Crossmodal binding through neural coherence: implications for multisensory processing. *Trends in neurosciences*, 31(8):401–409, 2008. 12

[281] Ana Serrano, Vincent Sitzmann, Jaime Ruiz-Borau, Gordon Wetzstein, Diego Gutierrez, and Belen Masia. Movie editing and cognitive event segmentation in virtual reality video. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 16, 24, 25, 26, 43, 48, 56, 128

[282] Michael Seufert, Sebastian Egger, Martin Slanina, Thomas Zinner, Tobias Hoßfeld, and Phuoc Tran-Gia. A survey on quality of experience of http adaptive streaming. *IEEE Communications Surveys & Tutorials*, 17(1):469–492, 2014. 9, 30

[283] Yuanjie Shao, Lerenhan Li, Wenqi Ren, Changxin Gao, and Nong Sang. Domain adaptation for image dehazing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2808–2817, 2020. 21

[284] Alia Sheikh, Andy Brown, Zillah Watson, and Michael Evans. Directing attention in 360-degree video. 2016. 27, 28, 86, 123

[285] Mohammadreza Sheykhmousa, Masoud Mahdianpari, Hamid Ghanbari, Fariba Mohammadimanesh, Pedram Ghamisi, and Saeid Homayouni. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:6308–6325, 2020. 18

[286] D Shin and F Biocca. Exploring immersive experience in journalism. New Media&Society, 19 (11), 1-24, 2017. 29

[287] Herbert A Simon. Designing organizations for an information-rich world. *International Library of Critical Writings in Economics*, 70:187–202, 1996. 29

[288] F Simone, Jesús Gutiérrez, and Le Callet. Complexity measurement and characterization of 360-degree content. 2019. 19, 30

[289] Lisa K Simone, Maria T Schultheis, Jose Rebimbas, and Scott R Millis. Head-mounted displays for clinical virtual reality applications: pitfalls in understanding user behavior while using technology. *CyberPsychology & Behavior*, 9(5):591–602, 2006. 16, 34

[290] Ashutosh Singla, Stephan Fremerey, Werner Robitza, Pierre Lebreton, and Alexander Raake. Comparison of subjective quality evaluation for HEVC encoded omnidirectional videos at different bit-rates for UHD and FHD resolution. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 511–519, 2017. 1, 11, 16, 19, 30, 34, 37, 39, 43, 135

[291] Ashutosh Singla, Stephan Fremerey, Werner Robitza, and Alexander Raake. Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays. In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2017. 12

[292] Ashutosh Singla, Steve Göring, Alexander Raake, Britta Meixner, Rob Koenen, and Thomas Buchholz. Subjective quality evaluation of tile-based streaming for omnidirectional videos. In *Proceedings of the 10th ACM Multimedia Systems Conference*, pages 232–242, 2019. 12, 38

[293] Ashutosh Singla, Werner Robitza, and Alexander Raake. Comparison of subjective quality evaluation methods for omnidirectional videos with DSIS and modified ACR. *Electronic Imaging*, 2018(14):1–6, 2018. 11, 30, 39

[294] Mel Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557, 2009. 28, 130

[295] Mel Slater, Angus Antley, Adam Davison, David Swapp, Christoph Guger, Chris Barker, Nancy Pistrang, and Maria V Sanchez-Vives. A virtual reprise of the stanley milgram obedience experiments. *PloS one*, 1(1):e39, 2006. 28, 29, 86

[296] Mel Slater, David-Paul Pertaub, Chris Barker, and David M Clark. An experimental study on fear of public speaking using a virtual environment. *CyberPsychology & Behavior*, 9(5):627–633, 2006. 29, 86

[297] Mel Slater and Sylvia Wilbur. A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 6(6):603–616, 1997. 13, 28, 86, 130

[298] Jeroen BJ Smeets and Eli Brenner. Grasping Weber's law. *Current Biology*, 18(23):R1089–R1090, 2008. 22

[299] Tim J Smith. The attentional theory of cinematic continuity. *Projections*, 6(1):1–27, 2012. 27, 86

[300] Tim J Smith and John M Henderson. Edit blindness: The relationship between attention and global change blindness in dynamic scenes. *Journal of eye movement research*, 2(2), 2008. 25

[301] Tim J Smith and Parag K Mital. Attentional synchrony and the influence of viewing task on gaze behavior in static and dynamic scenes. *Journal of vision*, 13(8):16–16, 2013. 17

[302] Will Smith. Stop calling Google Cardboard's 360-degree videos 'VR'. *Wired. Retrieved June*, 19:2019, 2015. 29

[303] Noah Snavely. Scene reconstruction and visualization from internet photo collections: A survey. *IPSJ Transactions on Computer Vision and Applications*, 3:44–66, 2011. 18

[304] Jake Snell, Karl Ridgeway, Renjie Liao, Brett D Roads, Michael C Mozer, and Richard S Zemel. Learning to generate images with perceptual similarity metrics. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 4277–4281. IEEE, 2017. 23, 75

[305] Jacob Søgaard, Lukáš Krasula, Muhammad Shahid, Dogancan Temel, Kjell Brunnström, and Manzoor Razaak. Applicability of existing objective metrics of perceptual quality for adaptive video streaming. In *Electronic Imaging, Image Quality and System Performance XIII*, 2016. 24, 61, 75

[306] Marco Speicher, Christoph Rosenberg, Donald Degraen, Florian Daiber, and Antonio Krúger. Exploring visual guidance in 360-degree videos. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, pages 1–12, 2019. 1, 27, 44, 86

[307] Anthony Steed, Sebastian Frlston, Maria Murcia Lopez, Jason Drummond, Ye Pan, and David Swapp. An 'in the wild' experiment on presence and embodiment using consumer virtual reality equipment. *IEEE transactions on visualization and computer graphics*, 22(4):1406–1414, 2016. 13

[308] Eckehard Steinbach, Sandra Hirche, Marc Ernst, Fernanda Brandi, Rahul Chaudhari, Julius Kammerl, and Iason Vittorias. Haptic communications. *Proceedings of the IEEE*, 100(4):937–956, 2012. 10, 50

[309] Lena Steindorf and Jan Rummel. Do your eyes give you away? A validation study of eye-movement measures used as indicators for mindless reading. *Behavior research methods*, 52:162–176, 2020. 14, 38

[310] Sophie Stellmach, Lennart Nacke, and Raimund Dachselt. Advanced gaze visualizations for three-dimensional virtual environments. In *Proceedings of the 2010 symposium on eye-tracking research & Applications*, pages 109–112, 2010. 16

[311] Joseph N Stember, Haydar Celik, E Krupinski, Peter D Chang, Simukayi Mutasa, Bradford J Wood, A Lignelli, Gul Moonis, LH Schwartz, Sachin Jambawalikar, et al. Eye tracking for deep learning segmentation using convolutional neural networks. *Journal of digital imaging*, 32:597–604, 2019. 18

[312] CM Sukanya, Roopa Gokul, and Vince Paul. A survey on object recognition methods. *International Journal of Science, Engineering and Computer Technology*, 6(1):48, 2016. 18

[313] Khena M Swallow, Jovan T Kemp, and Ayse Candan Simsek. The role of perspective in event segmentation. *Cognition*, 177:249–262, 2018. 26, 128

[314] Mohsen Tavakol and Reg Dennick. Making sense of Cronbach's alpha. *International journal of medical education*, 2:53, 2011. 81

[315] TDG. Tdg: Global consumer VR revenue to top $18 billion in 2025. *GlobeNewswire News Room*, Nov 2020. 1, 6

[316] Unity Technologies. Unity, 2023. 55

[317] Hadish Habte Tesfamikael, Adam Fray, Israel Mengsteab, Adonay Semere, and Zebib Amanuel. Simulation of eye tracking control based electric wheelchair construction by image segmentation algorithm. *Journal of Innovative Image Processing (JIIP)*, 3(01):21–35, 2021. 18

[318] Mai Thi Thanh Thai, Li Choy Chong, and Narendra M Agrawal. Straussian grounded theory method: An illustration. *The Qualitative Report*, 17(5), 2012. 67, 115

[319] Truong Cong Thang, Quang-Dung Ho, Jung Won Kang, and Anh T Pham. Adaptive streaming of audiovisual content using mpeg dash. *IEEE Transactions on Consumer Electronics*, 58(1):78–85, 2012. 7, 40

[320] Truong Cong Thang, Hung T Le, Anh T Pham, and Yong Man Ro. An evaluation of bitrate adaptation methods for http live streaming. *IEEE Journal on Selected Areas in Communications*, 32(4):693–705, 2014. 7, 40

[321] Jérôme Thevenot, Miguel Bordallo López, and Abdenour Hadid. A survey on computer vision for assistive medical diagnosis from faces. *IEEE journal of biomedical and health informatics*, 22(5):1497–1511, 2017. 19

[322] Clark Thompson. Depth perception in stereo computer vision. Technical report, STANFORD UNIV CA DEPT OF COMPUTER SCIENCE, 1975. 18

[323] Christian Timmerer, Markus Waltl, Benjamin Rainer, and Hermann Hellwagner. Assessing the quality of sensory experience for multimedia presentations. *signal processing: image communication*, 27(8):909–916, 2012. 10, 50

[324] Takahiro Toizumi, Simone Zini, Kazutoshi Sagi, Eiji Kaneko, Masato Tsukada, and Raimondo Schettini. Artifact-free thin cloud removal using gans. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3596–3600. IEEE, 2019. 21

[325] Lingwei Tong, Sungchul Jung, Richard Chen Li, Robert W Lindeman, and Holger Regenbrecht. Action units: Exploring the use of directorial cues for effective storytelling with swivel-chair virtual reality. In *Proceedings of the 32nd Australian Conference on Human-Computer Interaction*, pages 45–54, 2020. 24

[326] Melanie Tory, M Stella Atkins, Arthur E Kirkpatrick, Marios Nicolaou, and G-Z Yang. Eyegaze analysis of displays with combined 2d and 3d views. In *VIS 05. IEEE Visualization, 2005.*, pages 519–526. IEEE, 2005. 16

[327] Alexander Toshev, Jianbo Shi, and Kostas Daniilidis. Image matching via saliency region correspondences. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007. 18

[328] Huyen TT Tran, Nam Pham Ngoc, Cuong T Pham, Yong Ju Jung, and Truong Cong Thang. A subjective study on qoe of 360 video for VR communication. In *2017 IEEE 19th international workshop on multimedia signal processing (MMSP)*, pages 1–6. IEEE, 2017. 9, 12, 14, 30, 34, 35, 39, 43

[329] Matthew Turk. Computer vision in the interface. *Communications of the ACM*, 47(1):60–67, 2004. 18

[330] Ayşegül Uçar, Yakup Demir, and Cüneyt Güzeliş. Object recognition and detection with deep learning for autonomous driving applications. *Simulation*, 93(9):759–769, 2017. 18

[331] Scott E Umbaugh. *Computer vision and image processing: a practical approach using cviptools with cdrom*. Prentice Hall PTR, 1997. 18

[332] Lucia R Valmaggia, Daniel Freeman, Catherine Green, Philippa Garety, David Swapp, Angus Antley, Corinne Prescott, David Fowler, Elizabeth Kuipers, Paul Bebbington, et al. Virtual reality and paranoid ideations in people with an 'at-risk mental state'for psychosis. *The British Journal of Psychiatry*, 191(S51):s63–s68, 2007. 29

[333] Johanna C Van Niekerk and JD Roode. Glaserian and Straussian grounded theory: similar or completely different? In *Proceedings of the 2009 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists*, pages 96–103, 2009. 67

[334] Ashish V Vanmali and Vikram M Gadre. Visible and NIR image fusion using weight-map-guided laplacian–gaussian pyramid for improving scene visibility. *Sādhanā*, 42:1063–1082, 2017. 23

[335] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. User acceptance of information technology: Toward a unified view. *MIS quarterly*, pages 425–478, 2003. 11, 135

[336] Rafael Veras and Christopher Collins. Discriminability tests for visualization effectiveness and scalability. *IEEE transactions on visualization and computer graphics*, 26(1):749–758, 2019. 23, 24, 75

[337] Peter Vorderer, Werner Wirth, Feliz Ribeiro Gouveia, Frank Biocca, Timo Saari, Lutz Jäncke, Saskia Böcking, Holger Schramm, Andre Gysbers, Tilo Hartmann, et al. Mec spatial presence questionnaire. *Retrieved Sept*, 18(2004):2015, 2004. 13

[338] Mirjam Vosmeer, Christian Roth, and Hartmut Koenitz. Who are you? Voice-over perspective in surround video. In *Interactive Storytelling: 10th International Conference on Interactive Digital Storytelling, ICIDS 2017 Funchal, Madeira, Portugal, November 14–17, 2017, Proceedings 10*, pages 221–232. Springer, 2017. 27, 86

[339] Nicholas Wade, Benjamin W Tatler, et al. *The moving tablet of the eye: The origins of modern eye movement research*. Oxford University Press, USA, 2005. 14, 15

[340] Diane Walker and Florence Myrick. Grounded theory: An exploration of process and procedure. *Qualitative health research*, 16(4):547–559, 2006. 67, 115

[341] Jan Oliver Wallgrün, Mahda M Bagher, Pejman Sajjadi, and Alexander Klippel. A comparison of visual attention guiding approaches for 360 image-based VR tours. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 83–91. IEEE, 2020. 27, 86

[342] Cong Wang, Wanshu Fan, Yutong Wu, and Zhixun Su. Weakly supervised single image dehazing. *Journal of Visual Communication and Image Representation*, 72:102897, 2020. 21

[343] Guan Wang, Wenying Gu, and Ayoung Suh. The effects of 360-degree VR videos on audience engagement: evidence from the new york times. In *HCI in Business, Government, and Organizations: 5th International Conference, HCIBGO 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15-20, 2018, Proceedings 5*, pages 217–235. Springer, 2018. 6, 29, 30, 31, 132

[344] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube UGC dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019. 19

[345] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 21, 23, 24, 75, 123

[346] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 23, 24, 75

[347] Thomas M Ward, Pietro Mascagni, Yutong Ban, Guy Rosman, Nicolas Padoy, Ozanan Meireles, and Daniel A Hashimoto. Computer vision in surgery. *Surgery*, 169(5):1253–1256, 2021. 18

[348] Josh Weberruss, Lindsay Kleeman, David Boland, and Tom Drummond. Fpga acceleration of multilevel orb feature extraction for computer vision. In *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*, pages 1–8. IEEE, 2017. 18

[349] Nadir Weibel, Adam Fouse, Colleen Emmenegger, Sara Kimmich, and Edwin Hutchins. Let's look at the cockpit: exploring mobile eye-tracking for observational research on the flight deck. In *Proceedings of the symposium on eye tracking research and applications*, pages 107–114, 2012. 14

[350] Ben G Weinstein. A computer vision for animal ecology. *Journal of Animal Ecology*, 87(3):533–545, 2018. 18

[351] Cedric Westphal. Challenges in networking to support augmented reality and virtual reality. *IEEE ICNC*, 2017. 132

[352] Laurie M Wilcox, Robert S Allison, Samuel Elfassy, and Cynthia Grelik. Personal space in virtual reality. *ACM Transactions on Applied Perception (TAP)*, 3(4):412–428, 2006. 7, 28, 86

[353] Eric Williams, Carrie Love, and Matt Love. *Virtual reality cinema: narrative tips and techniques*. Routledge, 2021. 132

[354] Stefan Winkler. *Digital video quality: vision models and metrics*. John Wiley & Sons, 2005. 12, 34

[355] Hui-Yin Wu and Marc Christie. Stylistic patterns for generating cinematographic sequences. In *4th Workshop on Intelligent Cinematography and Editing Co-Located w/Eurographics 2015*, 2015. 25

[356] Wanmin Wu, Ahsan Arefin, Raoul Rivas, Klara Nahrstedt, Renata Sheppard, and Zhenyu Yang. Quality of experience in distributed interactive multimedia environments: toward a theoretical framework. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 481–490, 2009. 1, 2, 10, 11, 30, 34, 49, 50, 51, 61, 64, 123

[357] Yue Wu, Yicong Zhou, George Saveriades, Sos Agaian, Joseph P Noonan, and Premkumar Natarajan. Local shannon entropy measure with statistical tests for image randomness. *Information Sciences*, 222:323–342, 2013. 78, 123

[358] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2693–2708, 2018. 13, 130

[359] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360 immersive videos. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5333–5342, 2018. 13, 18, 24, 43, 130

[360] 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. 30, 130

[361] Kentaro Yamada, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Attention prediction in egocentric video using motion and visual saliency. In *Advances in Image and Video Technology: 5th Pacific Rim Symposium, PSIVT 2011, Gwangju, South Korea, November 20-23, 2011, Proceedings, Part I 5*, pages 277–288. Springer, 2012. 44

[362] Abid Yaqoob, Ting Bi, and Gabriel-Miro Muntean. A survey on adaptive 360 video streaming: Solutions, challenges and opportunities. *IEEE Communications Surveys & Tutorials*, 22(4):2801–2838, 2020. 7, 9, 12, 35, 39, 132

[363] Alfred L Yarbus and Alfred L Yarbus. Eye movements during perception of complex objects. *Eye movements and vision*, pages 171–211, 1967. 15

[364] Subiya Yaseen, Shireen Fathima, KV Surendra, P Jebran, and Saba Sanober. Gesture controlled touch less response using image processing. In *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS)*, pages 2610–2614. IEEE, 2017. 18

[365] Jiang Yu and Yong Liu. Field-of-view prediction in 360-degree videos with attention-based neural encoder-decoder networks. In *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems*, pages 37–42, 2019. 13

[366] Matt Yu, Haricharan Lakshman, and Bernd Girod. A framework to evaluate omnidirectional video coding schemes. In *2015 IEEE international symposium on mixed and augmented reality*, pages 31–36. IEEE, 2015. 7, 9

[367] Youngjae Yu, Sangho Lee, Joonil Na, Jaeyun Kang, and Gunhee Kim. A deep ranking model for spatio-temporal highlight detection from a 360 ◦ video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 19

[368] Jeffrey M Zacks. How we organize our experience into events. *Psychological Science Agenda*, 24(4), 2010. 25

[369] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, pages 492–511. Springer, 2020. 18

[370] Alireza Zare, Alireza Aminlou, Miska M Hannuksela, and Moncef Gabbouj. HEVC-compliant tile-based streaming of panoramic video for virtual reality applications. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 601–605, 2016. 7

[371] Bo Zhang, Junzhe Zhao, Shu Yang, Yang Zhang, Jing Wang, and Zesong Fei. Subjective and objective quality assessment of panoramic videos in virtual reality environments. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 163–168. IEEE, 2017. 2, 12, 30, 38, 39

[372] Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Jiashi Feng. Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4372–4381, 2017. 18

[373] Jing Zhao, Ruiqin Xiong, Jizheng Xu, Feng Wu, and Tiejun Huang. Learning a deep convolutional network for subband image denoising. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1420–1425. IEEE, 2019. 21

[374] Simone Zini, Simone Bianco, and Raimondo Schettini. Deep residual autoencoder for blind universal JPEG restoration. *IEEE Access*, 8:63283–63294, 2020. 21

[375] Michael Zink, Ramesh Sitaraman, and Klara Nahrstedt. Scalable 360 video stream delivery: Challenges, solutions, and opportunities. *Proceedings of the IEEE*, 107(4):639–650, 2019. 1, 6, 7, 8, 12, 24, 27, 40, 86, 123

[376] Wenjie Zou, Fuzheng Yang, Wei Zhang, Yi Li, and Haoping Yu. A framework for assessing spatial presence of omnidirectional video on virtual reality device. *IEEE Access*, 6:44676–44684, 2018. 12, 13, 38, 43, 52, 124, 135

[377] Haichun Zuo. Implementation of hci software interface based on image identification and segmentation algorithms. In *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*, pages 1–6. IEEE, 2016. 18

Appendices

The research as conducted in this thesis utilised a variety of methodologies, material and data-sets. Relevant content and supporting materials, such as developed Python scripts, advanced eye-tracking imagery, documentation, utilised questionnaires and other complementary material is presented in the subsequent appendices. The appendices were referenced accordingly within this thesis to ensure comprehensiveness and conciseness. This addendum is structured as follows:

It is important to note that each of the above appendices contain one or more sub-appendices. The specificity of which are detailed separately per Appendix.

Appendix A

Eye-Tracking Data Appendix

This Appendix comprises the resulting dataset containing the computed values of N_ψ (degree of gaze distribution) across all participants. Furthermore, the aggregate gaze distribution heatmaps for each of the utilised 360-degree videos are presented, as well as the aggregate gaze heatmaps per group. This Appendix is structured as follows:

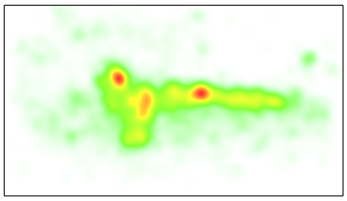
Important: the aggregate heatmaps were generated based on the cumulative eye-tracking data of the corresponding groups (i.e., all users, group R and group F). As such, the aggregate heatmaps provide an initial indication of overall exploration, but do not represent mean μ degrees of gaze distribution nor do they contain the nuances of individual computation. Thereby, the individual computations of N_ψ per participant for each of the 360-degree videos is detailed in Table A.1.

A1 Dataset of N_ψ

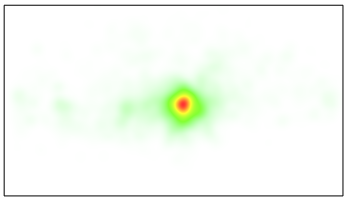
	Degree of Gaze Distribution N_ψ					
	A1	**A2**	**B1**	**B2**	**C1**	**C2**
R1	.272	.122	.190	.086	.219	.109
R2	.223	.132	.106	.064	.175	.216
R3	.209	.158	.198	.198	.199	.189
R4	.327	.175	.199	.143	.162	.145
R5	.209	.158	.198	.198	.199	.189
R6	.209	.158	.198	.198	.199	.189
R7	.251	.149	.190	.075	.215	.185
R8	.236	.083	.291	.096	.181	.134
R9	.144	.141	.158	.128	.255	.174
R10	.255	.142	.175	.182	.206	.145
R11	.185	.139	.216	.116	.212	.154
R12	.221	.086	.178	.116	.272	.181
R13	.227	.117	.161	.149	.177	.139
R14	.275	.140	.189	.122	.136	.189
R15	.195	.141	.181	.103	.176	.147
R16	.204	.134	.141	.093	.253	.151
R17	.211	.121	.163	.136	.144	.146
R18	.211	.115	.046	.050	.117	.127
R19	.293	.185	.245	.085	.226	.076
R20	.150	.111	.259	.060	.166	.168
R21	.211	.087	.216	.120	.209	.161
R22	.303	.137	.147	.185	.289	.132
R23	.180	.103	.095	.042	.146	.095
R24	.209	.104	.174	.101	.150	.115
R25	.226	.214	.255	.112	.156	.120
R26	.216	.070	.064	.054	.165	.110
F1	.071	.045	.047	.032	.074	.054
F2	.083	.057	.134	.082	.115	.079
F3	.141	.074	.090	.110	.051	.127
F4	.118	.062	.131	.060	.117	.095
F5	.153	.037	.12	.087	.126	.105
F6	.100	.093	.125	.072	.105	.088
F7	.167	.109	.130	.146	.165	.103
F8	.085	.069	.097	.114	.165	.066
F9	.111	.130	.124	.091	.141	.107
F10	.124	.179	.071	.079	.161	.104
F11	.116	.109	.078	.117	.179	.068
F12	.132	.141	.116	.083	.168	.091
F13	.112	.090	.077	.059	.146	.066
F14	.115	.121	.097	.095	.151	.072
F15	.115	.127	.144	.115	.168	.134
F16	.127	.171	.159	.147	.151	.103
F17	.187	.188	.105	.112	.198	.113
F18	.204	.143	.226	.124	.199	.147
F19	.175	.102	.175	.151	.158	.094
F20	.194	.134	.187	.105	.165	.134
F21	.147	.129	.224	.127	.153	.185
F22	.190	.093	.213	.091	.117	.091
F23	.159	.161	.238	.125	.190	.116
F24	.169	.150	.210	.163	.146	.127
F25	.216	.130	.163	.162	.192	.119
F26	.149	.125	.164	.087	.118	.081

Table A.1: N_ψ Dataset

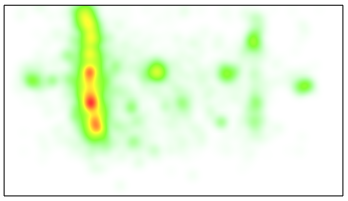
A2 Aggregate Gaze Distribution Heatmaps



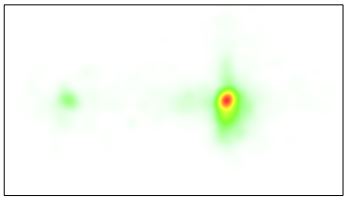
(a) A1: $N_\psi = .443$



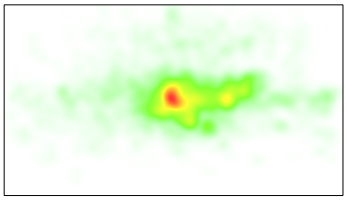
(b) A2: $N_\psi = .232$



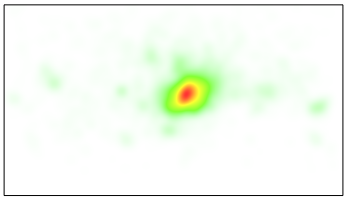
(c) B1: $N_\psi = .421$



(d) B2: $N_\psi = .183$

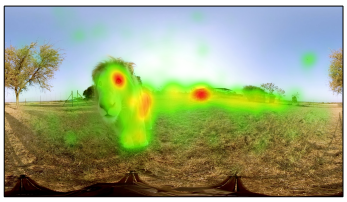


(e) C1: $N_\psi = .417$



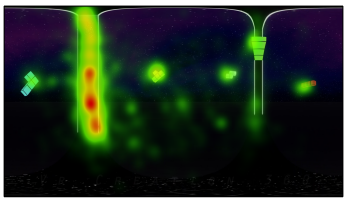
(f) C2: $N_\psi = .275$

Figure A.1: Aggregate gaze distribution heatmaps.

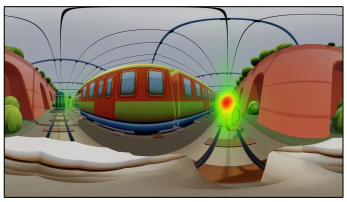


(a) A1: $N_\psi = .443$

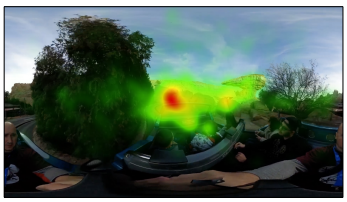
(b) A2: $N_\psi = .232$



(c) B1: $N_\psi = .421$



(d) B2: $N_\psi = .183$

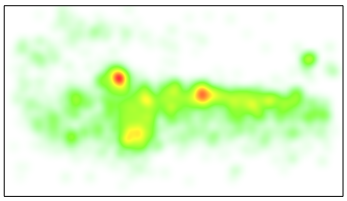


(e) C1: $N_\psi = .417$

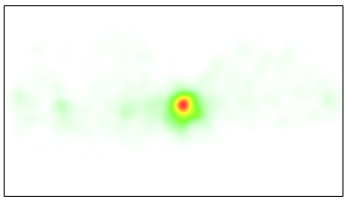
(f) C2: $N_\psi = .275$

Figure A.2: Aggregate gaze distribution heatmaps (projected).

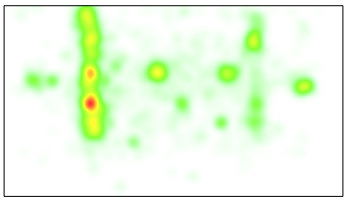
A3 Aggregate Gaze Distribution Heatmaps (Per Group)



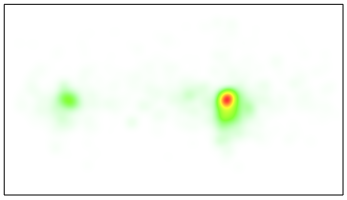
(a) A1: $N_\psi = .461$



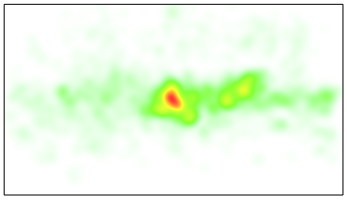
(b) A2: $N_\psi = .243$



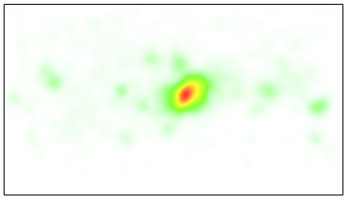
(c) B1: $N_\psi = .403$



(d) B2: $N_\psi = .188$

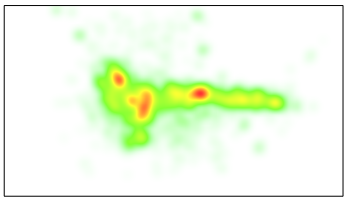


(e) C1: $N_\psi = .415$

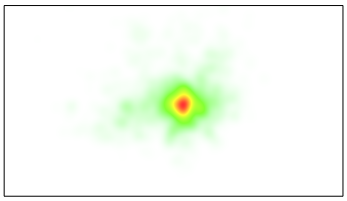


(f) C2: $N_\psi = .309$

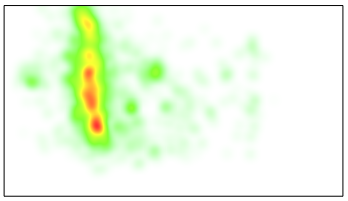
Figure A.3: Aggregate gaze distribution heatmaps (group R).



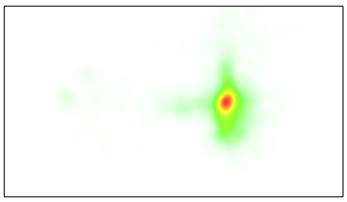
(a) A1: $N_\psi = .301$



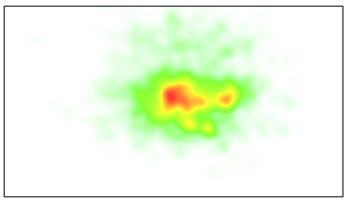
(b) A2: $N_\psi = .188$



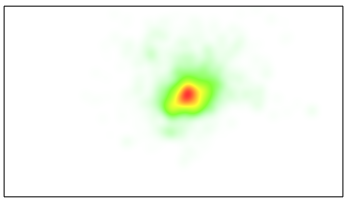
(c) B1: $N_\psi = .340$



(d) B2: $N_\psi = .166$



(e) C1: $N_\psi = .301$



(f) C2: $N_\psi = .183$

Figure A.4: Aggregate gaze distribution heatmaps (group F).

Appendix B

Python Repository

This Appendix comprises the repository of written and devised Python scripts to conduct the following computations: format conversion, image segmentation, image structure analyses, coding and mathematical calculations. This Appendix is structured as follows:

The Python scripts provided in this repository are designed for execution within suitable Python interpreters or development environments. It is important to note that the input and output file paths in each script are replaced by placeholders to ensure readability. When running the scripts, please take caution in substituting the correct file paths.

B4 EAC to ERP Conversion

```
import cv2
import numpy as np
import sys

def main(input_file, output_file):
    cap = cv2.VideoCapture(input_file)
    if not cap.isOpened():
        print(f"Error: file {input_file} not compatible")
        sys.exit(1)

    fps = int(cap.get(cv2.CAP_PROP_FPS))
    width = int(cap.get(cv2.CAP_PROP_FRAME_WIDTH))
    height = int(cap.get(cv2.CAP_PROP_FRAME_HEIGHT))
    fourcc = int(cap.get(cv2.CAP_PROP_FOURCC))

    out = cv2.VideoWriter(output_file, fourcc, fps, (width, height // 2))

    while cap.isOpened():
        ret, frame = cap.read()
        if not ret:
            break

        top = frame[:height // 2, :]
        bottom = frame[height // 2:, :]
        erp = cv2.resize(top, (width, height // 2),
        interpolation=cv2.INTER_CUBIC)
        * 0.5 + cv2.resize(bottom, (width, height // 2),
        interpolation=cv2.INTER_CUBIC) * 0.5

        out.write(erp.astype(np.uint8))

        if cv2.waitKey(1) & 0xFF == ord('q'):
            break

    cap.release()
    out.release()
    cv2.destroyAllWindows()

if __name__ == "__main__":
    if len(sys.argv) != 3:
        print("python eac_to_erp.py input_file output_file")
        sys.exit(1)

    input_file = sys.argv[1]
    output_file = sys.argv[2]
    main(input_file, output_file)
```

B5 Computation of Spatiotemporal Complexity

```
import cv2
import numpy as np
from tqdm.notebook import tqdm
import matplotlib.pyplot as plt

def sobel_filter(frame):
    grey_frame = cv2.cvtColor(frame, cv2.COLOR_BGR2GRAY)
    sobel_x = cv2.Sobel(grey_frame, cv2.CV_64F, 1, 0, ksize=3)
    sobel_y = cv2.Sobel(grey_frame, cv2.CV_64F, 0, 1, ksize=3)
    sobel = np.sqrt(np.square(sobel_x) + np.square(sobel_y))
    return sobel

def motion_difference_feature(frame1, frame2):
    grey_frame1 = cv2.cvtColor(frame1, cv2.COLOR_BGR2GRAY)
    grey_frame2 = cv2.cvtColor(frame2, cv2.COLOR_BGR2GRAY)
    motion_difference = np.abs(grey_frame2.astype(np.int16)
    - grey_frame1.astype(np.int16))
    return motion_difference

def calculate_complexities(video_path):
    cap = cv2.VideoCapture(video_path)
    total_frames = int(cap.get(cv2.CAP_PROP_FRAME_COUNT))

    ret, prev_frame = cap.read()

    si_values = []
    ti_values = []

    for _ in tqdm(range(1, total_frames), desc=f"Processing {video_path}"):
        ret, frame = cap.read()

        if not ret:
            break

        # Spatial complexity
        sobel_frame = sobel_filter(frame)
        current_std_spatial = np.std(sobel_frame)
        si_values.append(current_std_spatial)

        # Temporal complexity
        motion_diff = motion_difference_feature(prev_frame, frame)
        current_std_temporal = np.std(motion_diff)
        ti_values.append(current_std_temporal)

        prev_frame = frame

    cap.release()

    return np.array(si_values), np.array(ti_values)

def plot_per_frame_complexities(video_complexities):
    plt.figure(figsize=(12, 6))
    plt.title('Spatial and Temporal Complexity of Videos (Per Frame)')

    for video_name, (si_values, ti_values) in video_complexities.items():
        plt.scatter(si_values, ti_values, label=video_name, marker='o',
        s=50, alpha=0.5)

    plt.xlabel('Spatial Complexity (SI)')
    plt.ylabel('Temporal Complexity (TI)')
    plt.legend()
    plt.grid(True)
    plt.show()
```

```
    # Video plots
    for video_name, (si_values, ti_values) in video_complexities.items():
        plt.figure(figsize=(12, 6))
        plt.title(f'Spatial and Temporal Complexity of {video_name}
        (Per Frame)')

        plt.scatter(si_values, ti_values, label=video_name, marker='o',
        s=50, alpha=0.5)

        plt.xlabel('Spatial Complexity (SI)')
        plt.ylabel('Temporal Complexity (TI)')
        plt.legend()
        plt.grid(True)
        plt.show()


video_paths = [
    "/path/to/video",
]

video_complexities = {}

for video_path in video_paths:
    video_name = video_path.split('/')[-1]
    si_values, ti_values = calculate_complexities(video_path)
    video_complexities[video_name] = (si_values, ti_values)
    print(f"{video_name}:")
    print("Spatial Complexity (SI):", np.mean(si_values))
    print("Temporal Complexity (TI):", np.mean(ti_values))
    print()

plot_complexities(video_complexities)
plot_per_frame_complexities(video_complexities)
```

B6 Multiplicative Index E_1 of A and I_{norm}

```python
import cv2
import numpy as np

def area_intensity_index(heatmap):

    img = cv2.imread(heatmap)

    # Greyscale
    grey_img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)

    threshold = 254

    # Binary mask for non-white pixels
    non_white_mask = grey_img < threshold

    # Weighted sum of non-white pixel intensities
    weighted_sum = np.sum(grey_img[non_white_mask])

    non_white_pixels = non_white_mask.sum()
    total_pixels = grey_img.size

    # Average intensity of non-white pixels
    average_intensity = weighted_sum / non_white_pixels if non_white_pixels > 0
        else 0

    area_coverage = non_white_pixels / total_pixels

    E1 = (area_coverage * (average_intensity / 255))

    return E1


image_path = '/path/to/heatmapimage'
E1 = area_intensity_index(heatmap)
print(f"Area Ratio and Avg. Intensity Index E1: {E1:.2f}")
```

B7 MS-SSIM

```
import cv2
import numpy as np
from skimage.metrics import structural_similarity as compare_ssim

def ms_ssim(img1, img2, max_val=255, filter_size=11, filter_sigma=1.5, k1=0.01, k2
    =0.03):
    weights = np.array([0.0448, 0.2856, 0.3001, 0.2363, 0.1333])

    # Imgs to floats
    img1 = img1.astype(np.float64)
    img2 = img2.astype(np.float64)

    # Downsampling
    mssim_values = []
    for _ in range(len(weights)):
        ssim = compare_ssim(img1, img2, win_size=filter_size, sigma=filter_sigma,
                            data_range=max_val, use_sample_covariance=False,
                            K1=k1, K2=k2, full=True)[0]
        mssim_values.append(ssim)


        img1 = cv2.resize(img1, (0, 0), fx=0.5, fy=0.5)
        img2 = cv2.resize(img2, (0, 0), fx=0.5, fy=0.5)


    ms_ssim = np.prod(np.power(mssim_values, weights))

    return ms_ssim

img1 = cv2.imread("/path/to/heatmapimage", cv2.IMREAD_GRAYSCALE)
img2 = cv2.imread("/path/to/referenceimage", cv2.IMREAD_GRAYSCALE)

ms_ssim_value = ms_ssim(img1, img2)
print("MS-SSIM:", ms_ssim_value)
```

B8 Weighted Sum E_2 of d and $H(x)_{norm}$

```
import cv2
import numpy as np

def calculate_entropy(image):
    # Convert image to grayscale
    gray_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)

    # Histogram and normalisation
    hist = cv2.calcHist([gray_image], [0], None, [256], [0, 256])
    hist /= hist.sum()

    #Entropy
    entropy = -np.sum(hist * np.log2(hist + np.finfo(float).eps))

    return entropy

heatmap_image_path = '/path/to/heatmap'
heatmap_image = cv2.imread(heatmap_image_path)

entropy = calculate_entropy(heatmap_image)

# Normalise entropy
normalised_entropy = entropy / np.log2(256)

ms_ssim_value =

E2 = (1 - ms_ssim_value) + normalised_entropy

print('Entropy:', entropy)
print('Normalised Entropy:', normalised_entropy)
print('Dissimilarity and Entropy Index E2:', E2)
```

B9 Principal Component Analysis

```
import pandas as pd
import numpy as np
from sklearn.preprocessing import MinMaxScaler
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt
import seaborn as sns

data_path = '/path/to/database'
df = pd.read_excel(data_path)

X = df[['AREA_INTENSITY', 'D_ENTROPY']].values

# PCA
pca = PCA(n_components=2)
X_pca = pca.fit_transform(X)

weights = pca.components_
print(f"Weights: {weights}")

# PSI calculation
df['PSI'] = np.round(weights[0,0] * df['AREA'] + weights[0,1] * df['E2'], 3)

df.to_excel('/path/to/newdatabase', index=False)
```

B10 Quadrifactorial Exploration Index N_ψ

```
import cv2
import numpy as np
from skimage.metrics import structural_similarity as compare_ssim


def area_intensity_index(heatmap):
    img = cv2.imread(heatmap)
    grey_img = cv2.cvtColor(img, cv2.COLOR_BGR2GRAY)
    threshold = 254
    non_white_mask = grey_img < threshold
    weighted_sum = np.sum(grey_img[non_white_mask])
    non_white_pixels = non_white_mask.sum()
    total_pixels = grey_img.size
    average_intensity = weighted_sum / non_white_pixels if non_white_pixels > 0
        else 0
    area_coverage = non_white_pixels / total_pixels
    E1 = area_coverage * (average_intensity / 255)
    return E1


def calculate_entropy(image):
    gray_image = cv2.cvtColor(image, cv2.COLOR_BGR2GRAY)
    hist = cv2.calcHist([gray_image], [0], None, [256], [0, 256])
    hist /= hist.sum()
    entropy = -np.sum(hist * np.log2(hist + np.finfo(float).eps))
    return entropy


def ms_ssim(img1, img2, max_val=255, filter_size=11, filter_sigma=1.5, k1=0.01, k2
    =0.03):
    weights = np.array([0.0448, 0.2856, 0.3001, 0.2363, 0.1333])
    img1 = img1.astype(np.float64)
    img2 = img2.astype(np.float64)
    mssim_values = []
    for _ in range(len(weights)):
        ssim, _ = compare_ssim(img1, img2, win_size=filter_size, sigma=filter_sigma
            ,
                               data_range=max_val, use_sample_covariance=False,
                               K1=k1, K2=k2, full=True)
        mssim_values.append(ssim)
        img1 = cv2.resize(img1, (0, 0), fx=0.5, fy=0.5)
        img2 = cv2.resize(img2, (0, 0), fx=0.5, fy=0.5)
    ms_ssim = np.prod(np.power(mssim_values, weights))
    return ms_ssim


heatmap_image_path = '/path/to/heatmapimagesignal'
reference_image_path = '/path/to/whitebackground'

# Calculate E1
E1 = area_intensity_index(heatmap_image_path)
print(f"Area Ratio and Avg. Intensity Index E1: {E1:.2f}")

# Calculate E2
heatmap_image = cv2.imread(heatmap_image_path)
entropy = calculate_entropy(heatmap_image)
normalised_entropy = entropy / np.log2(256)
img1 = cv2.imread(heatmap_image_path, cv2.IMREAD_GRAYSCALE)
img2 = cv2.imread(reference_image_path, cv2.IMREAD_GRAYSCALE)
ms_ssim_value = ms_ssim(img1, img2)
d = (1 - ms_ssim_value)
E2 = 0.388 * d + 0.289 * normalised_entropy
print('Entropy:', entropy)
print('Normalised Entropy:', normalised_entropy)
```

```
print('Dissimilarity and Entropy Index E2:', E2)

# Calculate psi
psi = 0.726 * E1 + E2

# Calculate N_psi
N_psi = psi / 1.414
print(f"N_psi: {N_psi:.3f}")
```

B11 Diegetic Coding

```
import pandas as pd

columns = ['Artefact ID', 'Artefact Label', 'Artefact Start Time', 'Artefact End
    Time', 'Artefact Size', 'Artefact Duration Score', 'Artefact DAS Score']
df = pd.DataFrame(columns=columns)

# Size classes and point values
size_values = {'s': 1, 'm': 2, 'l': 3}

# Duration classes and point values
def duration_score(duration):
    if duration <= 2:
        return 1
    elif duration <= 10:
        return 2
    else:
        return 3

video_name = 'video'

artefact_id = 1

# Manual annotation loop
while True:

    start_time = input("t_start (in s): ")
    if start_time.lower() == 'q':
        break
    end_time = input("t_end (in s): ")
    size = input("size (s/m/l): ")
    artefact_name = input("label: ")

    duration = float(end_time) - float(start_time)
    duration_points = duration_score(duration)
    artefact_das_score = size_values[size.lower()] * duration_points

    df = df.append({'Artefact ID': artefact_id,
                    'Artefact Label': artefact_name,
                    'Artefact Start Time': float(start_time),
                    'Artefact End Time': float(end_time),
                    'Artefact Size': size,
                    'Artefact Duration Score': duration_points,
                    'Artefact DAS Score': artefact_das_score}, ignore_index=True)

    # Increment
    artefact_id += 1

total_das_score = df['Artefact DAS Score'].sum()
print(f'Total DAS {video_name}: {total_das_score}')

df.to_csv(f'{video_name}_annotations.csv', index=False)
```

Appendix C

Ethics and Privacy Quick Scan

This Appendix comprises the Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences was conducted to assess potential ethical and privacy concerns related to this research, as well as the approval correspondence. The ethical approval can be found in Appendix C12, on page 192.

The Ethics and Privacy Quick Scan of the Utrecht University Research Institute of Information and Computing Sciences:

- P1. Does your project involve human participants? This includes for example use of observation, (online) surveys, interviews, tests, focus groups, and workshops where human participants provide information or data to inform the research. If you are only using existing data sets or publicly available data (e.g. from Twitter, Reddit) without directly recruiting participants, please answer no. | Yes.
- P2. Does your project involve participants younger than 18 years of age? | No.
- P3. Does your project involve participants with learning or communication difficulties of a severity that may impact their ability to provide informed consent? | No.
- P4. Is your project likely to involve participants engaging in illegal activities? | No.
- P5. Does your project involve patients? | No.
- P6. Does your project involve participants belonging to a vulnerable group, other than those listed above? | No.
- P8. Does your project involve participants with whom you have, or are likely to have, a working or professional relationship: for instance, staff or students of the university, professional colleagues, or clients? | Yes.
- P9. Is it made clear to potential participants that not participating will in no way impact them (e.g. it will not directly impact their grade in a class)? | Yes.
- PC1. Do you have set procedures that you will use for obtaining informed consent from all participants, including (where appropriate) parental consent for children or consent from legally authorised representatives? (See suggestions for information sheets and consent forms on the website. | Yes.
- PC2. Will you tell participants that their participation is voluntary? | Yes.
- PC3. Will you obtain explicit consent for participation? | Yes.
- PC4. Will you obtain explicit consent for any sensor readings, eye tracking, photos, audio, and/or video recordings? | Yes.
- PC5. Will you tell participants that they may withdraw from the research at any time and for any reason? | Yes.
- PC6. Will you give potential participants time to consider participation? | Yes.
- PC7. Will you provide participants with an opportunity to ask questions about the research before consenting to take part (e.g. by providing your contact details)? | Yes.
- PC8. Does your project involve concealment or deliberate misleading of participants? | No.
- D1. Are you gathering or using personal data (defined as any information relating to an identified or identifiable living person)? | Yes.
- DR1. Will you process personal data that would jeopardise the physical health or safety of individuals in the event of a personal data breach? | No.
- DR2. Will you combine, compare, or match personal data obtained from multiple sources, in a way that exceeds the reasonable expectations of the people whose data it is? | No.
- DR3. Will you use any personal data of children or vulnerable individuals for marketing, profiling, automated decision-making, or to offer online services to them? | No.

- DR4. Will you profile individuals on a large scale? | No.
- DR5. Will you systematically monitor individuals in a publicly accessible area on a large scale (or use the data of such monitoring)? | No.
- DR6. Will you use special category personal data, criminal offence personal data, or other sensitive personal data on a large scale? | No.
- DR7. Will you determine an individual's access to a product, service, opportunity, or benefit based on an automated decision or special category personal data? | No.
- DR8. Will you systematically and extensively monitor or profile individuals, with significant effects on them? | No.
- DR9. Will you use innovative technology to process sensitive personal data? | No.
- DM1. Will you collect only personal data that is strictly necessary for the research? | Yes.
- DM4. Will you anonymise the data wherever possible? | Yes.
- DM5. Will you pseudonymise the data if you are not able to anonymise it, replacing personal details with an identifier, and keeping the key separate from the data set? | Yes.
- DC1. Will any organisation external to Utrecht University be involved in processing personal data (e.g. for transcription, data analysis, data storage)? | No.
- DI1. Will any personal data be transferred to another country (including to research collaborators in a joint project)? | No.
- DF1. Is personal data used to recruit participants? | No.
- DP1. Will participants be provided with privacy information? (Recommended is to use as part of the information sheet: For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information at www.uu.nl/en/organisation/privacy.) | Yes.
- DP2. Will participants be aware of what their data is being used for? | Yes.
- DP3. Can participants request that their personal data be deleted? | Yes.
- DP4. Can participants request that their personal data be rectified (in case it is incorrect)? | Yes.
- DP5. Can participants request access to their personal data? | Yes.
- DP6. Can participants request that personal data processing is restricted? | Yes.
- DP7. Will participants be subjected to automated decision-making based on their personal data with an impact on them beyond the research study to which they consented? | No.
- DP8. Will participants be aware of how long their data is being kept for, who it is being shared with, and any safeguards that apply in case of international sharing? | Yes.
- DP9. If data is provided by a third party, are people whose data is in the data set provided with (1) the privacy information and (2) what categories of data you will use? | N/a.
- DE1. Will you use any personal data that you have not gathered directly from participants (such as data from an existing data set, data gathered by a third party, data scraped from the internet)? | No.
- DS1. Will any data be stored (temporarily or permanently) anywhere other than on password-protected University authorised computers or servers? | No.

- DS4. Excluding (1) any international data transfers mentioned above and (2) any sharing of data with collaborators and contractors, will any personal data be stored, collected, or accessed from outside the EU? | No.
- H1. Does your project give rise to a realistic risk to the national security of any country? | No.
- H2. Does your project give rise to a realistic risk of aiding human rights abuses in any country? | No.
- H3. Does your project (and its data) give rise to a realistic risk of damaging the University's reputation? (E.g., bad press coverage, public protest.) | No.
- H4. Does your project (and in particular its data) give rise to an increased risk of attack (cyber- or otherwise) against the University? (E.g., from pressure groups.) | No.
- H5. Is the data likely to contain material that is indecent, offensive, defamatory, threatening, discriminatory, or extremist? | No.
- H6. Does your project give rise to a realistic risk of harm to the researchers? | No.
- H7. Is there a realistic risk of any participant experiencing physical or psychological harm or discomfort? If yes, provide detail on how you will minimise risks. | Yes, see below.
- H8. Is there a realistic risk of any participant experiencing a detriment to their interests as a result of participation? | No.
- H9. Is there a realistic risk of other types of negative externalities? | No.
- C1. Is there any potential conflict of interest (e.g. between research funder and researchers or participants and researchers) that may potentially affect the research outcome or the dissemination of research findings? | No.
- C2. Is there a direct hierarchical relationship between researchers and participants? | No.
- Z10. In case you encountered warnings in the survey, does your supervisor already have ethical approval for a research line that fully covers your project? | No.

The potential physical discomfort the participant may experience is cybersickness. During the study, participants will be asked to wear a head-mounted display (Virtual Reality headset). The usage of this type of device can potentially result in the sensation of cybersickness, which a cluster of symptoms that occur in the absence of physical motion, similar to motion sickness. Also, while interacting with the HMD, the participant might experience physical strain of looking around extensively when viewing a 360-degree video. To minimise the risk and impact of these effect, a series of precautions will be put in place:

- During the recruitment process, participants will be asked to fill out a survey in which they are asked about risk inducing parameters, such as: sensitivity to motion sickness or dizziness, recent surgery, back, neck or other physical conditions and other related factors that are a risk. Participants with any of these conditions/sensitivity limitations will be excluded from participation.
- A pilot study will be conducted with a small sample size to determine optimal duration of the main user study. During the pilot study, the participants will be informed about the exploratory nature of the pilot study and that the main goal is to determine the parameters of the main study so that participants do not experience a sense of physical strain, discomfort or cybersickness.
- After the pilot study, participants will be asked to fill out the SSQ (simulator sickness questionnaire) which will be used to determine the level of sickness and dizziness the participants have experienced. This data will be used to determine the duration and other parameters of the main study so that cybersickness or physical discomfort is minimised.
- Only study designs will be utilised that are proven to minimise the risk of cybersickness, such as the M-ACR methodology for QoE assessment.
- Participants will be seated on an ergonomic, (rotational) chair to relieve any stress on the participant's spine or neck and minimise risk as much as possible.
- Constant monitoring of the participant's well-being and mandatory breaks.

C12 Ethical Approval

Dear Rik Hazekamp,

Thank you for completing the ethics scan. Based on the information provided below, I can tell you that the research can go ahead.

Best wishes,

Maartje

Dr. M.M.A. (Maartje) de Graaf
Assistant Professor of Human-Computer Interaction
Coordinator of HCI Master Graduations

Utrecht University
Faculty of Science
Department of Information and Computing Sciences
Princetonplein 5
3584 CC UTRECHT

Visiting: Buys Ballot Building - Room 4.21
Contact by phone +31 6 4851 2017 or email m.m.a.degraaf@uu.nl
https://robonarratives.wordpress.com/

Appendix D

Sampling Correspondence

This Appendix comprises the written correspondence utilised during the sampling and recruitment of participants for the study. This Appendix is structured as follows:

- Appendix D13: Written recruitment for the pre-test parameter study, presented on page 194.
- Appendix D14: Written recruitment for the eye tracking study, presented on page 195.

D13 Recruitment Pre-Test Parameter Study

Dear [Participant],

As part of my master's degree thesis, I am conducting oculesics research to study eye movement while viewing 360-degree videos in a virtual reality environment. Before conducting the main study, I am running a pilot study to determine the parameters and identify potential risks. The pilot study will involve performing the entire main study, but with the main goal of assessing the duration, the potential for cybersickness, and any other risks associated with the study.

I am currently seeking participants for the pilot study, in which you will view a sequence of 360-degree videos in a virtual reality environment, while your eye movements and gaze data are acquired using eye tracking technology. Afterwards, you will be interviewed about your experience. Your participation in this pilot study will help me ensure that the experiment design minimises risks as much as possible.

If you are interested in participating in the pilot study, please note that participation is not free of risk, and you may experience physical discomfort such as motion sickness or dizziness. The study will take approximately 30-60 minutes to complete. To participate in the study, you must meet the following criteria:

- Be over 18 years old
- Have normal or corrected-to-normal vision (incl. color-blindness)
- Have no history of motion sickness or epilepsy
- Have no physical conditions that may limit or be aggravated by using a VR headset
- Haven't participated in similar research in the past.

Please note that individuals who have undergone eye surgery or have eye diseases, wear heavy makeup, or have high myopia may be excluded from participating in the study due to potential effects on eye tracking performance.

If you meet the criteria and would like to participate, please click on the following link [LINK] to let me know your availability. The pilot study will be conducted in a lab setting, at the Utrecht University.

Please note that participation in the pilot study is voluntary, and you may withdraw from the study at any time. Your data will be kept confidential, and only used for research purposes.

Thank you for considering participation in my study. If you have any questions, please do not hesitate to contact me.

Sincerely,

Rik Hazekamp

D14 Recruitment Eye Tracking Study

Dear [Participant],

As part of my master's degree thesis, I am conducting oculesics research to study eye movement while viewing 360-degree videos in a virtual reality environment. I am currently seeking participants for an experimental study in which you will view a sequence of 360-degree videos in a virtual reality environment, while your eye movements and gaze data are acquired using eye tracking technology.

Your participation in this study will help me gain insights into how people interact with 360-degree videos in a virtual reality environment and how different content factors can affect user behaviour. The study will take approximately 30 minutes to complete.

To participate in the study, you must meet the following criteria:

- Be over 18 years old
- Have normal or corrected-to-normal vision (incl. colour-blindness)
- Have no history of motion sickness or epilepsy
- Have no physical conditions that may limit or be aggravated by using a VR headset
- Haven't participated in similar research in the past.

Please note that individuals who have undergone eye surgery or have eye diseases, wear heavy makeup, or have high myopia may be excluded from participating in the study due to potential effects on eye tracking performance.

If you are interested in participating in this study, please click on the following link [LINK] to read more about this research and to get in contact. The study will be conducted in a lab setting, at the Utrecht University.

Please note that participation is voluntary, and you may withdraw from the study at any time. Your data will be kept confidential, and only used for research purposes.

Thank you for considering participation in my research. If you have any questions, please do not hesitate to contact me.

Sincerely,

Rik Hazekamp

Appendix E

Information and Consent

This Appendix comprises the research participation information sheets of the pre-test parameter study and eye tracking study which was conducted as part of this thesis. Furthermore, the consent form used to acquire written consent is presented as well. This Appendix is structured as follows:

E15 Information Sheet Pre-Test Parameter Study

Research Participant Information Sheet
"The Significance of Spatiotemporal Image Complexity on Gaze Dynamics in VR-based 360º Video Interactions: An Integrated Oculesics and Computer Vision Approach"

May, 2023

1. Introduction This information sheet is presented as part of participation in scientific research at Utrecht University. In this information sheet, information about the nature of the research, the background thereof, the researchers involved, the research activities, data acquisition and usage, ethical approval and contact information are presented.

2. What is the background and purpose of this study? The objective of the research is to determine the risk-reducing parameters of an upcoming research project on how spatiotemporal complexities affect user behaviour in 360-degree videos. To do so, you will conduct the experiment of this research project, and are afterwards evaluated on the physical risks of conducting the experiment. Background information on the research scope is as follows. While previous research mainly focused on technical aspects, this study takes a more holistic approach that takes into account content characteristics to investigate how users perceive and behave when interacting with 360-degree video content. User behaviour is assessed through both objective and subjective metrics: the first part of the research design involves an eye-tracking study that utilises gaze data to measure the user's gaze while viewing a variety of 360-degree video content. The second part is a user evaluation study that enables subjective analysis of the user's interaction and perception. The main aim of this study is to evaluate the duration and how aspects of cybersickness or physical discomfort are present in the current study design to ensure participant comfort and prevent physical discomfort.

3. Who will carry out the study? This study is carried out by R. Hazekamp (r.hazekamp@students.uu.nl) as part of my master thesis under supervision of W. Hürst (huerst@uu.nl). The research project is conducted as part of the Human-Computer Interaction research curriculum in the Research Institute of Information and Computing Sciences of Utrecht University.

4. How will the study be carried out? In this study, you will perform a series of tasks. Firstly, you will be asked to fill out a demographic survey. After the demographic survey has been filled out, you will view a sequence of 360-degree videos using a VR system. After the viewing session, you will be evaluated on the experience of viewing those videos in an evaluation. This consists of a short survey and small interview regarding your perception and experience of the 360-degree videos. Important; you will *not* be interviewed on the content of the videos, therefore it is *not required* to memorise the content. Lastly, you will be asked to fill out a questionnaire regarding cybersickness and physical discomfort, and you are evaluated on the duration of the experiment. The experiment will take about 40 to 45 minutes. During the experiment, refreshments are provided and short breaks are implemented. You will not receive any monetary compensation for participating, however mutual participation in research as a quid pro quo can be discussed with the researcher.

5. What will we do with your data? If you consent to this, audio, photo, video, eye tracking and sensor recordings will be acquired. The data acquired will only be used to set parameters of the upcoming research. Sensory recordings will not be further analysed in this thesis. The recordings will be stored on a secure university server. The recordings will be transcribed so that participants' opinions are captured into text. All recordings will be securely deleted after analysis (within 3 months of the study). The transcribed text and recordings will be anonymised so that you will not be identifiable. All data will be processed using pseudonymisation techniques, using a pseudonym to link the constituted personal data and all data recordings, ensuring that your data will be completely anonymised. The pseudonym will be deleted within 3 months of the study. The data will become part of my thesis and will also be stored in a data repository for use by other researchers and research users. My thesis, any publications based on this research, and the data repository will not include your name or any other individual information by which you could be identified.

6. What are your rights? Participation is voluntary. We are only allowed to collect your data for our study if you consent to this. If you decide not to participate, you do not have to take any further action. You do not need to sign anything. Nor are you required to explain why you do not want to participate. If you decide to participate, you can always change your mind and stop participating at any time, including *during* the study. You will even be able to withdraw your consent *after* you have participated. However, if you choose to do so, we will not be required to undo the processing of your data that has taken place up until that time. The personal data and sensor recordings we have obtained from you up until the time when you withdraw your consent will be erased (where personal data is any data that can be linked to you, so this excludes any already anonymised data).

7. Approval of this study The exploratory nature of this study necessitates the notion that risks are not yet minimized during this study. Therefore, risk of cybersickness or physical discomfort are present. Please contact the researcher in case of need. This study has been allowed to proceed by the Research Institute of Information and Computing Sciences on the basis of an Ethics and Privacy Quick Scan. If you have a complaint about the way this study is carried out, please send an email to: ics-ethics@uu.nl. If you have any complaints or questions about the processing of personal data, please send an email to the Faculty of Sciences Privacy Officer: privacy-beta@uu.nl. The Privacy Officer will also be able to assist you in exercising the rights you have under the GDPR. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information at www.uu.nl/en/organisation/privacy.

8. More information about this study? If you have any questions or concerns about this research please contact R. Hazekamp at r.hazekamp@students.uu.nl or the research supervisor W. Hürst at huerst@uu.nl.

9. Appendices The included Appendix, the consent form, can be found in Appendix E17.

E16 Information Sheet Eye Tracking Study

Research Participant Information Sheet
"The Significance of Spatiotemporal Image Complexity on Gaze Dynamics in VR-based 360º Video Interactions: An Integrated Oculesics and Computer Vision Approach"

May, 2023

1. Introduction This information sheet is presented as part of participation in scientific research at Utrecht University. In this information sheet, information about the nature of the research, the background thereof, the researchers involved, the research activities, data acquisition and usage, ethical approval and contact information are presented.

2. What is the background and purpose of this study? The objective of the research is to examine how spatiotemporal complexities affect user behaviour in 360-degree videos. While previous research mainly focused on technical aspects, this study takes a more holistic approach that takes into account visual content characteristics to investigate how users perceive and behave when interacting with varying 360-degree video content. User behaviour is assessed through both objective and subjective metrics: the first part of the research design involves an eye-tracking study that utilises gaze data to measure the user's gaze while viewing a variety of 360-degree video content. The second part is a user evaluation study that enables subjective analysis of the user's interaction and perception. To ensure participant comfort and prevent physical discomfort, precautions were taken during the viewing sessions.

3. Who will carry out the study? This study is carried out by R. Hazekamp (r.hazekamp@students.uu.nl) as part of my master thesis under supervision of W. Hürst (huerst@uu.nl). The research project is conducted as part of the Human-Computer Interaction research curriculum in the Research Institute of Information and Computing Sciences of Utrecht University.

4. How will the study be carried out? In this study, you will perform a series of tasks. Firstly, you will be asked to fill out a demographic survey. After the demographic survey has been filled out, you will view a sequence of 360-degree videos using a VR system. After the viewing session, you will be evaluated on the experience of viewing those videos in an evaluation. This consists of a short survey and small interview regarding your perception and experience of the 360-degree videos. Important; you will *not* be interviewed on the content of the videos, therefore it is *not required* to memorise the content. The experiment will take about 30 minutes. During the experiment, refreshments are provided and short breaks are implemented. You will not receive any monetary compensation for participating, however mutual participation in research as a quid pro quo can be discussed with the researcher.

5. What will we do with your data? If you consent to this, audio, photo, video, eye tracking and sensor recordings will be acquired. The recordings will be stored on a secure university server. The recordings will be transcribed so that participants' opinions are captured into text. All recordings will be securely deleted after analysis (within 3 months of the study). The transcribed text and recordings will be anonymised so that you will not be identifiable. All data will be processed using pseudonymisation

techniques, using a pseudonym to link the constituted personal data and all data recordings, ensuring that your data will be completely anonymised. The pseudonym will be deleted within 3 months of the study. The data will become part of my thesis and will also be stored in a data repository for use by other researchers and research users. My thesis, any publications based on this research, and the data repository will not include your name or any other individual information by which you could be identified.

6. What are your rights? Participation is voluntary. We are only allowed to collect your data for our study if you consent to this. If you decide not to participate, you do not have to take any further action. You do not need to sign anything. Nor are you required to explain why you do not want to participate. If you decide to participate, you can always change your mind and stop participating at any time, including *during* the study. You will even be able to withdraw your consent *after* you have participated. However, if you choose to do so, we will not be required to undo the processing of your data that has taken place up until that time. The personal data and sensor recordings we have obtained from you up until the time when you withdraw your consent will be erased (where personal data is any data that can be linked to you, so this excludes any already anonymised data).

7. Approval of this study This study has been allowed to proceed by the Research Institute of Information and Computing Sciences on the basis of an Ethics and Privacy Quick Scan. If you have a complaint about the way this study is carried out, please send an email to: ics-ethics@uu.nl. If you have any complaints or questions about the processing of personal data, please send an email to the Faculty of Sciences Privacy Officer: privacy-beta@uu.nl. The Privacy Officer will also be able to assist you in exercising the rights you have under the GDPR. For details of our legal basis for using personal data and the rights you have over your data please see the University's privacy information at www.uu.nl/en/organisation/privacy.

8. More information about this study? If you have any questions or concerns about this research please contact R. Hazekamp at r.hazekamp@students.uu.nl or the research supervisor W. Hürst at huerst@uu.nl.

9. Appendices The included Appendix, the consent form, can be found in Appendix E17.

E17 Consent Form

Consent Form for Participation in the Research Project:
"The Significance of Spatiotemporal Image Complexity on Gaze Dynamics in VR-based 360º Video Interactions: An Integrated Oculesics and Computer Vision Approach"

Please read the statements below and tick the final box to confirm you have read and understood the statements and upon doing so agree to participate in the project.

I confirm that I am 18 years of age or over.

I confirm that the research project "*The Significance of Spatiotemporal Image Complexity on Gaze Dynamics in VR-based 360º Video Interactions: An Integrated Oculesics and Computer Vision Approach*" has been explained to me. I have had the opportunity to ask questions about the project and have had these answered satisfactorily. I had enough time to consider whether to participate.

I consent to the material I contribute being used to generate insights for the research project "*The Significance of Spatiotemporal Image Complexity on Gaze Dynamics in VR-based 360º Video Interactions: An Integrated Oculesics and Computer Vision Approach*".

I consent to audio, video, photo, eye tracking and sensor recordings being used in this study as explained in the information sheet. I understand that I can request to stop recordings at any time.

I understand that if I give permission, the audio, video, photo, eye tracking and sensor recordings will be held confidentially so that only R. Hazekamp has access to the recording. The recordings will be anonymised and stored on a secure, password protected server and anonymised for up to 3 months after which period they will be securely destroyed, fully anonymised, transcribed/encoded in an anonymous form and the original securely destroyed. In accordance with the General Data Protection Regulation (GDPR) I can have access to my recordings and can request them to be deleted at any time during this period.

I understand that in addition to the recordings, other personal data will be collected from me and that this information will be held confidentially so that only R. Hazekamp has access to this data and is able to trace the information back to me personally. The information will be anonymised and stored on a secure, password protected server and anonymised for up to 3 months after which period if will be deleted. In accordance with the General Data Protection Regulation (GDPR) I can have access to my information and can request my data to be deleted at any time during this period.

I understand that my participation in this research is voluntary and that I may withdraw from the study at any time without providing a reason, and that if I withdraw any personal data already collected from me will be erased.

I consent to allow the fully anonymised data to be used in future publications and other scholarly means of disseminating the findings from the research project.

I understand that the data acquired will be securely stored by researcher(s), but that appropriately anonymised data may in future be made available to others for research purposes. I understand that the University may publish appropriately anonymised data in appropriate data repositories for verification purposes and to make it accessible to researchers and other research users.

I understand that I can request any personal data collected from me to be deleted.

☐ I confirm that I have read and understood the above statements and agree to participate in the study (Check the box).

Appendix F

Questionnaires

This Appendix comprises the set of questionnaires developed as part of the research experiment. This Appendix is structured as follows:

F18 Demographic Questionnaire

The demographic questionnaire was used at the start of the experiment to assess eligibility, acquire written consent and obtain demographic information from each of the participants. The questionnaire contains the following set of questions:

What is your age?

What is your gender?
Male / Female / Other

Have you ever used a virtual reality headset before?
Yes / No / Not sure

On a scale of 1 to 5, how often do you use a VR headset?
1. Never
2. Once a year or less
3. A few times a year
4. Once a month or more
5. Once a week or more

Do you have normal or corrected-to-normal vision?
Yes / No / Not sure

Do you have any visual impairments that affect your ability to see clearly or use VR technology?
Yes / No

Have you ever experienced motion sickness or epilepsy?
Yes / No / Not sure

Do you have any physical conditions that may limit or be aggravated by using a VR headset?
Yes / No / Not sure

Do you have any back, neck or similar physical conditions, or have you undergone recent surgery that may limit your ability to wear a VR headset or move your head?
Yes / No

Have you participated in similar research in the past?
Yes / No

Have you read and understood the information sheet and do you agree with the terms and conditions of participating in this research?
Yes / No / Not sure

Do you give consent with the researcher and Utrecht University using any anonymised data and sensor recordings for the duration of this experiment?
Signature

F19 Simulator Sickness Questionnaire

The Simulator Sickness Questionnaire (SSQ) [163] utilised during the pre-test parameter study, and functions as an important tool in establishing the parameters on minimising cybersickness during the eye tracking experiment. The participants were asked to rate the amount of effect on each of the following, potential, symptoms:

	None	Slight	Moderate	Severe
1. General discomfort				
2. Fatigue				
3. Headache				
4. Eye strain				
5. Difficulty focusing				
6. Salivation increasing				
7. Sweating				
8. Nausea				
9. Difficulty concentrating				
10. "Fullness of head"				
11. Blurred vision				
12. Dizziness with eyes open				
13. Dizziness with eyes closed				
14. * Vertigo				
15. ** Stomach awareness				
16. Burping				

Table F.1: Simulator Sickness Questionnaire

* Vertigo entails the sensation of loss of orientation from respect to a vertical position.
** Stomach awareness entails the sensation of discomfort in the stomach area, commonly prior to nausea.

The rating was done using the following 4-point scale: *none - slight - moderate - severe.*

F20 User Evaluation Questionnaire

User Evaluation Questionnaire (UEQ) was used during the experiment phase of the thesis, and was used to acquire subjective measurements on the viewing experience across all participants. The UEQ comprises closed-ended questions and statements entailing attributes of engagement, attention, spatial awareness and usability context in regards to viewing behaviour. The UEQ contains the following set of questions and statements:

Engagement and Attention (scale 1-7):

How enjoyable did you find the overall 360-degree video experience?
How would you rate the richness and quality of the graphics in the 360-degree videos?
How quickly did time seem to pass while watching the 360-degree videos?
To what extent did you feel situated in the story being depicted in the 360-degree videos?
How much control did you feel you had over your viewing experience while watching the 360-degree videos?
To what extent were you unaware of the presence of others while watching the 360-degree videos?
To what extent were you able to maintain your attention on the 360-degree video throughout the entire viewing experience?

"The fast and dynamic camera motion in some of the videos made it difficult for me to focus on a particular area of the scene."
"I found myself getting distracted by the background elements when watching the videos with a static camera."
"I found myself more focused on the details of the scene when the camera was moving slowly."

Spatial Awareness and Usability Context (scale 1-7):

How well were you able to locate and identify important objects or landmarks within the virtual environment?
How well were you able to understand the layout of the virtual environment?
How well were you able to navigate through the virtual environment?

"I had a better understanding of the layout of the environment when watching the 360-degree content with a static camera."
"I found it difficult to orient myself and understand the layout of the environment when watching the 360-degree video with a moving camera."

How comfortable was the use of the [chair type] during the viewing session?
To what extent did the [chair type] affect your overall enjoyment of the 360-degree video?

Fixed chair:
"I felt limited in the amount of exploring I could do due to the fixed chair."
"I found it harder to keep track of the camera movements because of the fixed chair."
Rotating chair:
"I felt more encouraged to look around because of the rotating chair."
"The rotating chair made it easier for me to follow the camera movements."

The rating was done using the following 7-point Likert scale.

F21 Semi-Structured Interview

The semi-structured interview (SSI) was used during the experiment phase of the thesis, and was used to acquire subjective measurements on the viewing experience across all participants. The SSI comprises a series of open-ended questions relating to the participant's perception of viewing behaviour, content and interpretation of both. The SSI contains the following set of open-ended questions:

On a scale of 1 to 7, with 1 being 'not at all interesting' and 7 being 'extremely interesting', how interesting did you find the 360-degree videos?

On a scale of 1 to 7, with 1 being 'not at all' and 7 being 'extremely', how much did you experience the sense of FOMC, due to loss of information or missed out content? If so, can you describe when and why?

Can you provide a brief description of each of the videos you watched in this study? Please provide elements or details you found interesting or remember vividly from each of the videos.

Some of the videos were more dynamic, with the camera moving relatively fast or having more movement. Other videos were more static, with the camera moving relatively slow or remaining stationary. Can you describe the effect this had on how you viewed the content?

How would you describe your viewing behaviour between the different genres of videos (scenery, roller coaster, video game)? How was it different and why do you think that was the case?