Exploring the Impact of Audio-Visual Information in Omnidirectional Videos on User Behaviors for Virtual Reality

Chen Wang B.S.

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Aljosa Smolic

Co-supervisor: Cagri Ozcinar

09 2020

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

Chen Wang

September 16, 2020

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Chen Wang

September 16, 2020

Acknowledgments

I would like to express my warmest thanks to my supervisor Prof. Aljosa Smolic and co-supervisor Dr. Cagri Ozcinar, for all the valuable guidance and continuous encouragement throughout my dissertation. Then I would like to thank my second reader Prof. Michael Brady who taken out some of his time and gave me suggestions for improving my dissertation during my presentation.

I would also like to express my gratitude to all the staff and members from Trinity College Dublin, the School of Computer Science and Statistics, for teaching the knowledge of the field of computer science and helping me over this year.

Finally, I would like to thank my parents and friends for their support throughout my life.

CHEN WANG

University of Dublin, Trinity College
09 2020

Exploring the Impact of Audio-Visual Information in Omnidirectional Videos on User Behaviors for Virtual Reality

Chen Wang, Master of Science in Computer Science

University of Dublin, Trinity College, 2020

Supervisor: Aljosa Smolic

Co-supervisor: Cagri Ozcinar

Recent advances in Virtual reality (VR) technology have caused omnidirectional videos (ODVs) to emerge as an increasingly important vehicle for the delivery of immersive content provision. Unlike traditional videos, with the support of head-mounted displays (HMDs), the ODVs would bring a deeper sense of immersion to users. Ambisonics, which refers to a complete three-dimensional spherical audio scene, is used to present ODVs' audio content. The ODVs allow audience attention to be directed concurrently by both audio and visual stimuli. As a result, understanding how audio-visual information affects user behaviors is significant to improve the quality of experience of ODVs for VR applications. Thus far, research in this area is limited, a situation which this dissertation seeks to remedy. With this aim, we collected an audio-visual dataset containing trajectories and conducted a quantitative statistical analysis of the user navigation patterns. This analysis included visualization of viewport center trajectories and head-motion analysis while users were watching omnidirectional videos under ambisonics, mono, and mute. It was observed that there were variations in user behaviours when watching ODVs, which correlated with the three different audio modalities. We believe that this research contributes to the existing literature on the audio-visual perception of ODVs.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Virtual reality (VR) technology has made significant progress in academia and industry. The popularity of VR has shifted people's attention from traditional language narration to visual presentation. Users can see the real-time rendering of different 360-degree images by wearing a head-mounted display (HMD), and they can replicate real-world movement, such as walking and gaming, and decide in which direction to view the surrounding scenes. Scene simulation in the virtual world allows users to experience the invisible external world and immerse themselves in the virtual environment.

Essential to VR technology are 360-degree videos, also known as spherical videos or omnidirectional videos (ODVs). While wearing an HMD, the user can choose which portion of the spherical video content to watch. The portion of spherical surface content displayed is projected to a segment of plane, called the viewport. The audio and visual representation of ODVs is completely different from that of traditional 2D videos. At present, 360-degree videos are shot by a collection of cameras or an omnidirectional camera. Image-stitching technology is used to seamlessly stitch together multiple images, thus forming a clear and complete omnidirectional video utilizing compression coding and network-transmission technology.

Compared with traditional videos, ODVs can capture more scene information, and they usually need to be compressed before transmission. They are mostly stored in the equirectangular projection (ERP) format, which is the most widely used projection format for representing 360-degree video on a 2D plane. In this format, the longitude of a sphere is taken as the X coordinate and the latitude as the y coordinate of the projection. Specifically, ERP expands the sphere and stretches it into a rectangle. The longitude (horizontal axis) is in the range [0, 2 π], the latitude (vertical axis) is in the range [0, π], and where π is the ratio of a circle's circumference to its diameter. According to the relationship of the latitude and longitude, the length-width ratio of the plane after ERP is 2:1; it can subsequently be expressed in the range $360^\circ \times 180^\circ$.

The audio representation of ODVs is fully spherical (i.e., ambisonics), which improves users' sense of immersion by enabling them to experience the loudness and direction of sound sources. The technology of ambisonics can record the sound field information of the whole spherical space centered on the listener or microphone, and transform the information into the ambisonics B-format, which can be saved and played back accurately. The audio track of the ambisonics format does not correspond to the actual speaker channel but is an independently encoded signal that contains sound-field direction information. A microphone encodes the most common ambisonics format, called the first-order ambisonics format, with four diaphragms picking up the sound-field information simultaneously. It has four channels, W, X, Y, Z, which capture omnidirectional sound information: front and rear, left and right, and height sound information. Thus, ambisonics satisfies the sound requirements of ODVs and provides a sense of spatialization and immersion.

Omnidirectional videos are widely used in entertainment, medicine, education, and film and television. In recent years, technology giants such as Facebook and YouTube have offered platforms for 360-degree video content. They use deep learning to develop and deploy technologies, thus optimizing the way users make 360-degree video content. Omnidirectional video technology has attracted increasing amounts of attention, and so it is necessary to explore users' behavior and navigation patterns. A better understanding of users' behavior would benefit the design and optimization of VR functions such as rendering and streaming, thus improving user experience.

1.2 Motivation

As the demand for ODV technology in VR applications has increased, many studies have investigated users' behavior and predicted their visual attention. The human visual attention mechanism embodies a selective attention ability; that is, when faced with a scene, humans automatically process regions of interest and selectively ignore regions of non-interest. Therefore, it is necessary to detect the saliency of the information in ODVs to reduce redundant visual information.

Xu et al. [2] have reviewed a number of recent studies that have used saliency maps to analyze how users navigate and explore ODV content for VR. These studies are limited, however, because they ignore audio cues and pay too much attention to visual cues, despite the fact that the audio part plays as important a role as the visual part in the virtual reality experience. Evidence exists regarding the correlation between audio and visual cues and their joint contribution to visual attention [3]. Thus, studying how audio and visual cues affect ODV viewing behavior patterns is particularly important for optimizing VR systems, especially for video post production, coding, and streaming.

The present research tries to fill this research gap and contribute to the research of audio-visual attention for ODV.

1.3 Research Objective

The objective of this research is to study how audio and visual cues influence users' navigation patterns while watching 360-degree video. For this purpose, a dataset was developed containing viewpoint center trajectories (VCTs) information of 45 users and 12 videos with three audio modalities: ambisonics, mono and mute. These audio modalities provide users with different audio-visual experiences. The main difference is that, when watching video in mono, users can detect only the loudness of the sound; they cannot judge the direction of the sound. In contrast, when watching video in ambisonics, users can judge the direction of the sound source according to hints provided by the audio. (For the mute modality, users see the visual content without any sound.)

User behavior while watching 360-degree video in these three audio modalities has not been fully investigated. This dissertation presents a comprehensive analysis of the user navigation traces and head motion statistics (e.g., angular distance, angular velocity, and attendance cumulative density functions) to explore the user behavior while watching 360-degree video in ambisonics, mono, and mute.

1.4 Thesis Structure

The rest of this thesis is organized as follows. Chapter 2 introduces related work. Chapter 3 explains the methodologies and implementation of the experiments. The results are discussed in Chapter 4. Chapter 5 and 6 present the overall conclusion and recommendations for future work, respectively.

2 Literature Review

2.1 Omnidirectional Video

As VR (Virtual Reality) technology has attracted tremendous attention from academia and industry, the value of the VR industry has been experiencing explosive growth in recent years. ODVs are used to present VR contents with the help of head-mounted displays (HMDs) in VR technology [4]. Due to the immersion sense created by ODVs, VR technology is applied in many scenarios, such as game development, education and film production.

With an increasing demand for VR technology, a lot of famous companies, including Oculus, HTC and Sony, have launched VR headsets and equipment. On the other hand, ODVs have seen a significant increase in popularity. Facebook introduced the function of viewing 360-degree videos and supported playback across all platforms, including real-time 360-degree video streaming; YouTube also launched 360-degree video playback and uploads, and the combination of them enables a much wider audience to engage in this format than ever before[5].

Unlike the rectangular capture for traditional videography, ODVs is a technology which utilizes the omnidirectional cameras to seize different perspectives of space and then stitches together into a spherical video, thus bringing the users a deeper immersion sense. The viewer could wear HMDs to observe the surrounding scenes and control the navigation from all viewing directions. This enables the users to keep in the center of the viewing sphere, thus placing users within the context of an event or scene instead of presenting them as an external observer. The following section introduces the processing of 360-degree video, which is presented in **Figure 2.1**.

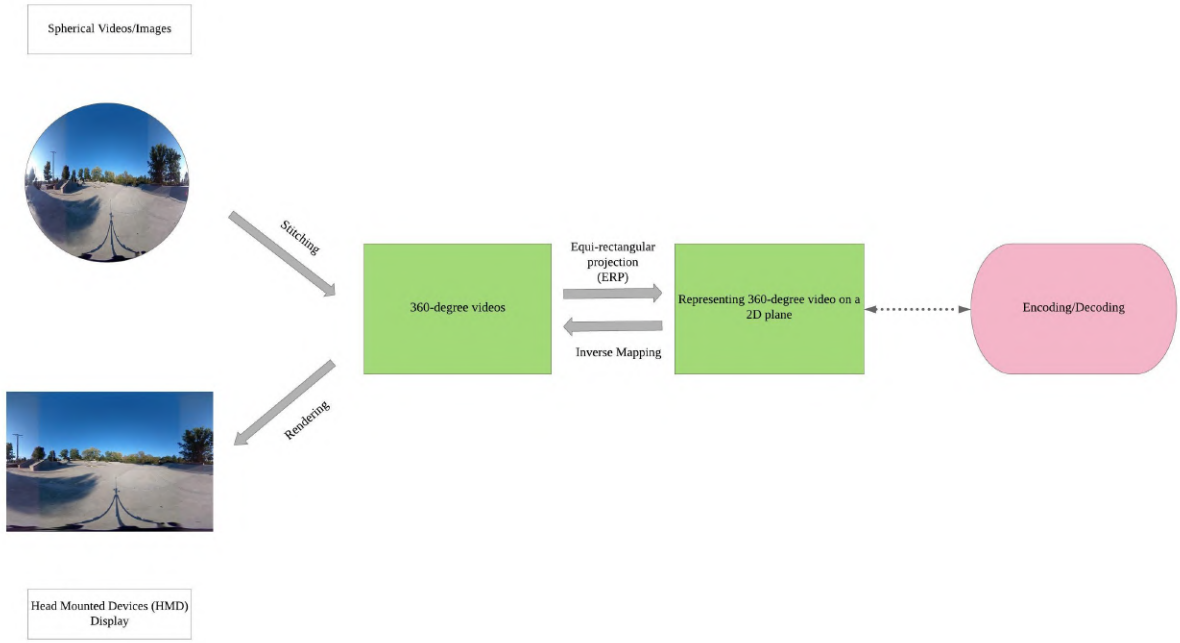


Figure 2.1: The procedure of a 360-degree video transmission chain[1]

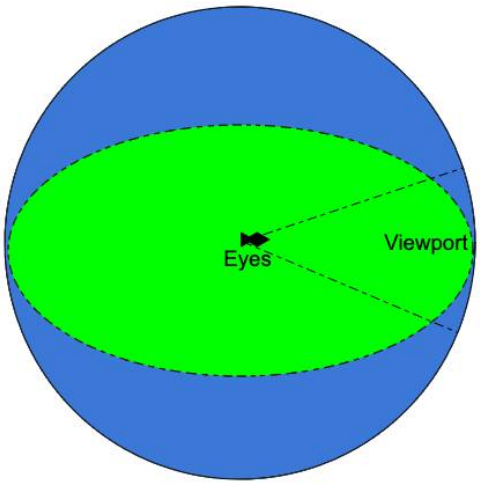


Figure 2.2: The viewport of a 360-degree video

The multiple lens panoramic camera has a 360-degree angle of view in both horizontal and vertical directions, which can cover all spherical environment centered on the shooting point. Therefore, it could be used to obtain multiple angle images or videos. After stitching images or videos, the 360-degree video obtained will be projected onto the 2D plane using a planar representation (e.g. Equi-rectangular projection (ERP)) [1]. Once the ERP is turned on, the transformed ERP picture will be encoded. After decoding, a planar representation 360-degree video will be converted back to a sphere by inverse mapping. When omnidirectional videos are rendered

through HMDs, only a part of the spherical viewing could be displayed and presented at a time, which is called viewport(see **Figure 2.2**). Finally, HMDs will be used to track and record the users' head movements to generate the corresponding viewport.

The visual presentation form of ODVs is consistent with the visual physiological and psychological characteristics of human viewing external objects. Due to the limited range of human visual angle, the human eye could only focus on a narrow visual range at a certain time. More specifically, during head-up viewing, the human eye could only see the region near the equator of the spherical video, rather than its polar region. Therefore, no matter what kind of media equipment is used to play ODVs, the information that the human eye could accept at a certain time is about the one-twelfth sector of the total visual information. In other words, users will only focus on a certain area of the screen when watching the ODVs. For film-producers and game development companies, it's important to present more attractive omnidirectional audio-visual content for users within the limited time and conditions. It's very necessary to research users' navigation patterns when watching 360-degree videos, thus finding what kind of content will attract users' attention, arranging the content reasonably and improving the quality of experience (QoE) of ODVs.

2.2 Ambisonics

Ambisonics is a method invented in the 1970s to record, mix, and play back fully spherical audio. With the rapid development of the VR industry, which demands 360-degree audio solutions, ambisonics technology was gradually commercialized. The basic method of ambisonics is to regard an audio scene as a sphere of sound coming from various directions around the center point (the location of the microphone during recording).

From a recording perspective, ambisonics can be understood as a 3D expansion of a mid-side stereo recording system. It uses multiple microphone head arrays combined in such a way as to record multi-track sound signals with height and depth information. It finally forms an omnidirectional sound field. The most widely used and mature technology is first-order ambisonics, which uses four channels. The first-order ambisonics microphone consists of four microphone heads with a heart-shaped pick-up pattern, pointing to front left, back left, front right, and back right. The original picked-up signal is called the A-format. B-format [6], which is the most popular ambisonics format today and widely used in VR and 360-degree video, can be obtained through the superposition or inverse superposition of the four channels (see **Figure 2.3**).

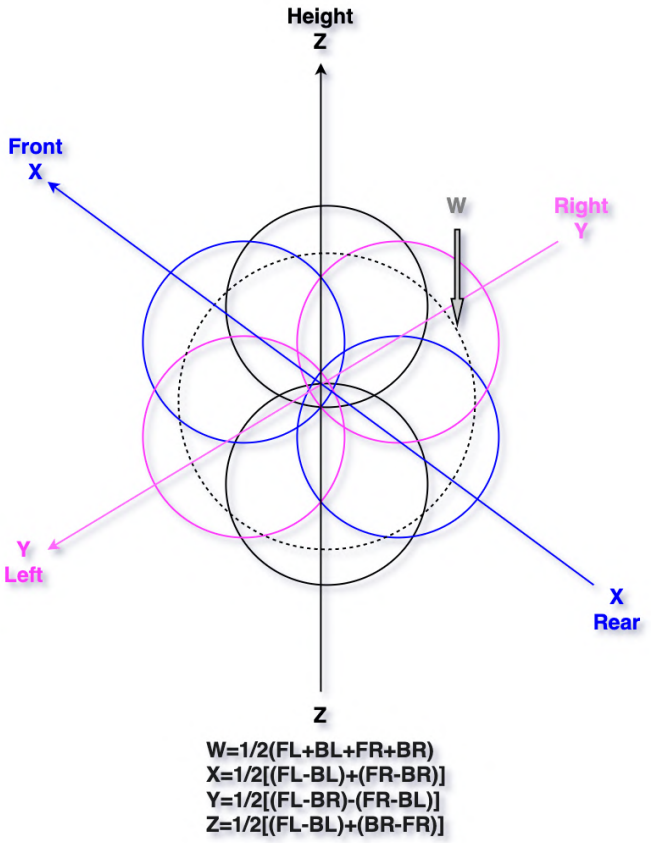


Figure 2.3: The convertion from the first-order ambisonics A-format to B-format

In first-order B format, there are four channels, called W, X, Y and Z. Each channel can be regarded as a different microphone polar pattern pointing in a specific direction, and the four channels are conjoined at the center point of the 360-degree sphere:

- W, as an omnidirectional polar pattern, contains all sounds coming from different directions at equal gain and phase in the 360-degree sphere;
- X, Y and Z, which are all figure-8 bi-directional polar patterns, point forward, left and up, respectively.

Each figure-8 microphone has two sides: a positive side and a negative (inverse phase) side. For example, the figure-8 polar pattern of the X channel points forward, and the negative side points backwards. The audio signal generated on the X channel includes all the sounds in the front of the sphere with positive phase and all the sounds from the back of the sphere with negative phase. Moreover, the gain picked up in figure-8 microphones for each direction is different. Specifically, full gain will be picked up for the audio signal in front or behind, but when the user leaves the bi-directional axis, the gain will decrease gradually until it is exactly 90 degrees from figure-8, and the gain will be zero. While the Y channel picks up the left side of the 360-degree

sphere with the positive phase and the right side with a negative phase, the Z channel picks up the bottom of the 360-degree sphere with a negative phase and the top with the positive phase. In this way, the combination of the four channels represents the three dimensions (i.e., a 360-degree sphere of sound) through the differential gain and phase relationships [7].

Ambisonics technology has been used to deliver 360 audio for omnidirectional videos, gaming and virtual reality experiences in the VR industry. Usually, the audio is experienced by the end user via headphones and an HMD. It has apparent advantages in the performance of 3D space and the positioning of sound elements and better resolution in height positioning.

Monaural sound (mono) uses only one channel of audio, which means that there is no difference in the information received by the left and right ears, and the auditory system does not allow for psychoacoustic localization (i.e., only the loudness of the audio, not the direction, can be detected) [8]. With ambisonics, however, users are able to position sounds in three dimensions—like a sphere around the listening position. In other words, an ODV with ambisonics not only provides auditory cues but also enables the direction of sound sources to be detected, whereas in mono, only the magnitude of auditory cues can be detected.

2.3 Equirectangular Projection

As mentioned before, projecting a spherical 360-degree video onto 2D planes is an important process to generate ODVs. Equirectangular Projection, which is the most widely used format for representing 360-degree images and videos, maps meridians(latitudes) and circles(longitudes) to the vertical and horizontal axes, respectively, by spacing them equally on the 2D plane. The north pole and south pole, located in the upper and lower edge of the whole plane respectively, are stretched across the entire frame width [9, 10]. The following **Figure 2.4** shows the Equirectangular perspective and flattening.

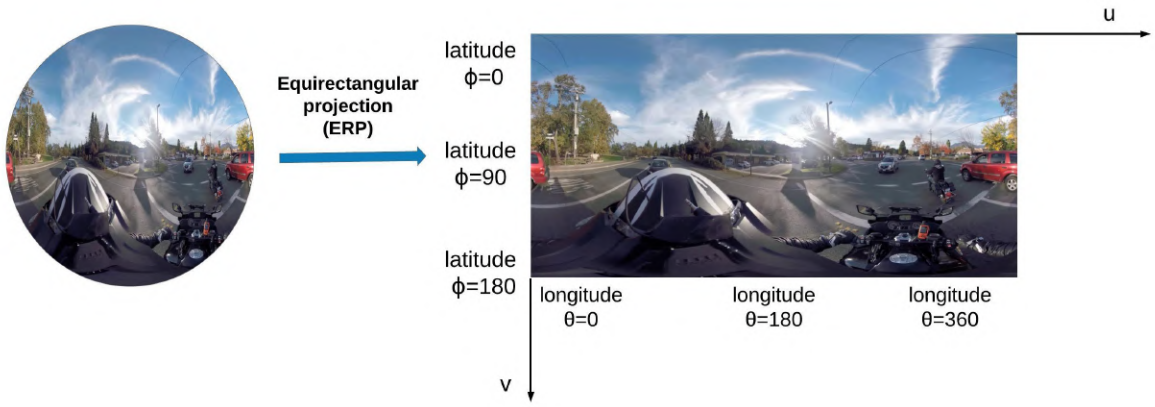


Figure 2.4: The Equirectangular perspective and flattening

The longitude (θ) and latitude (Φ) are used to describe the sphere coordinates (θ, Φ). The range of the longitude is [0, 2π], while the latitude is in the range [0, π], where is the ratio of a circle's circumference to its diameter. The sphere coordinates (θ, Φ) could be transformed to planar coordinates (u, v) in a 2D plane coordinate system, referred as the **uv** plane. In the **uv** plane, u and v are in the range [0, 1]. Thus, the longitude and latitude (θ, Φ) in the sphere can be calculated from (u, v) using Equation **1, 2**:

$$\theta = (u - 0.5) \times 2\pi \tag{1}$$

$$\Phi = (0.5 - v) \times \pi \tag{2}$$

In addition, the three-dimensional geometry of projection format representation is described with 3D Cartesian coordinate system. Starting from the center of the sphere, X axis points toward the front of the sphere, Y axis points toward the top of the sphere, and Z axis points toward the right of the sphere. The formula, which could convert longitude and latitude (θ, Φ) to 3D Cartesian coordinates (x, y, z), is defined as follows:

$$x = \cos(\Phi)\cos(\theta) \tag{3}$$

$$y = \sin(\Phi) \tag{4}$$

$$z = -\cos(\Phi)\cos(\theta) \tag{5}$$

Inversely, the longitude and latitude (θ, Φ) can be obtained from (x, y, z) coordinates using **6, 7**.

$$\theta = \arctan(-\frac{z}{x}) \tag{6}$$

$$\Phi = \arcsin \frac{y}{\sqrt{x^2 + y^2 + z^2}} \tag{7}$$

For the transformation from 2D to 3D, (θ, Φ)could be firstly calculated by using **1, 2**. Then, the formula **3, 4, 5** are used to evaluate 3D coordinates (x, y, z) [11].

2.4 Related Work

ODVs play an important part in providing users with a high quality of experience (QoE) in VR applications. HMDs allow viewers to explore the scene and control their navigation by simply rotating their heads. However, the limitations of HMD and ODV bring new challenges and requirements to the design of visual and audio content in VR. Exploring and understanding viewers' behaviors and navigation patterns when watching ODVs is significant for content producers of VR applications. A lot of researches about users' behavior analysis and navigation patterns have been investigated in recent years. Generally speaking, most of studies involved are predicting users' visual attention using different kinds of datasets and modelling approaches.

Saliency maps, which are two-dimensional probability distributions, represent the probability that the user is likely to fixate on a given region. It could be used to compute and present analysis outcome of users' behavior based on head datasets and eye movements datasets [12, 13, 14]. Sitzmann et al. [12] captured and analyzed gaze and head orientation datasets from 169 users. In their work, they conducted statistical analysis about similar behaviors between users for each of the scenes tested and the existence of a particular fixation bias using saliency maps and metrics related. Also, the work in Ozcinar and Smolic [13] analyzed the performance of standard video saliency detection methods using six ODVs rendered by an HMD. The results reveal that the quantity of fixations depends on motion complexity of ODV. The head and eye movements dataset from 51 participants dataset when watching 360-degree videos has been proposed in David et al. [14] to explore how the VR content influence the users' visual attention using saliency maps. Furthermore, to evaluate the performance of different saliency models when using saliency maps as inputs, evaluation metrics are introduced. Bylinskii et al. [15] presented a comprehensive analysis of eight different evaluation metrics and their properties.

Apart from saliency maps, there are also some studies regarding additional quantitative analysis based on viewport traces, such as the average angular velocity, frequency of fixation, and mean exploration angles [16, 17, 18]. Wu et al. [16] presented a dataset containing 18 omnidirectional videos with five categories, such as sports, performance, etc. They recruited 48 participants in two experiments: in the first experiment, participants were asked to watch the VR Spherical Videos; in the second experiment, participants were asked specific questions about the video content after each session. They also showed the analysis of user behavior patterns through the

experiments. Duanmu et al. [17] presented a dataset of view center traces gathered from over 50 viewers watching 360-degree videos on the computer. Also, they conducted the statistical analysis over the user navigation patterns and compared the differences of user behaviors between the head-mounted display (HMD) and computer viewing sessions. The authors in [18] presented a comprehensive analysis of user navigation patterns for watching VR videos across different contents and viewing devices(i.e. HMD, tablet and laptop) through the dataset of users' navigation trajectories. Also, an optimization of the user-centric server was proposed based on previous analysis consequences.In [19], a dataset of head movements from 59 users watching 360-degree videos on HMDs is introduced, and some examples of statistics for a content-dependent analysis of users' navigation patterns are shown.

To further predict visual scan paths, Rossi et al. [20] proposed a graph-based method for classifying users who pay attention to the same regions of the scene for a long time over ODVs. Furthermore, Nasrabadi et al. [21] proposed an improved clustering approach which increases the viewport prediction accuracy.

However, the feature of studies above is only focusing on visual cues. In recent times, the academic community is more and more interested in the aspects of the audio-visual perception for ODV. For example, Rana et al. [22] proposed a novel audio-visual video dataset of 265 videos and used different modeling approaches to generate ambisonics for ODV. Also, In [23], an unsupervised algorithm was presented to address the problem of localizing sound sources in visual scenes. 'DAVE' was proposed to study the applicability of audio cues in conjunction with visual ones in predicting saliency maps using deep neural networks [24]. Furthermore, in Chao et al. [25] work, a statistical analysis using saliency maps to study the audio-visual perception of ODVs was presented. Both of them illustrate audio and visual information contributes to affecting user navigation patterns and behaviors when watching 360-degree videos in VR applications. As a result, it's significant to further focus on researching audio-visual perception analysis for ODVs in VR applications.

3 Methodology

3.1 Data Collection

This chapter describes the dataset and technical details of the subjective experiment. The dataset comprises Viewport Center Trajectories (VCTs) from 45 participants across three audio modalities (ambisonics, mono, and mute) in three categories (Conversation, Music, and Environment).

3.1.1 Material

Fifteen ODVs were selected from YouTube, all of which were captured using a single-lens camera (i.e., they produced mono-scopic 360-degree videos). When viewed from an HMD, a mono-scopic ODV is immersive but appears a little flat. The audio modalities of the videos used in this experiment include ambisonics, mono, and mute. ODVs with first-order B-format ambisonics have four channels, called X, Y, Z and W, and ODVs in mono have only one channel, which can be distributed equally in the left and right headphones, thus allowing the user to detect the loudness of the audio but not the direction. The mute ODVs had all audio channels removed.

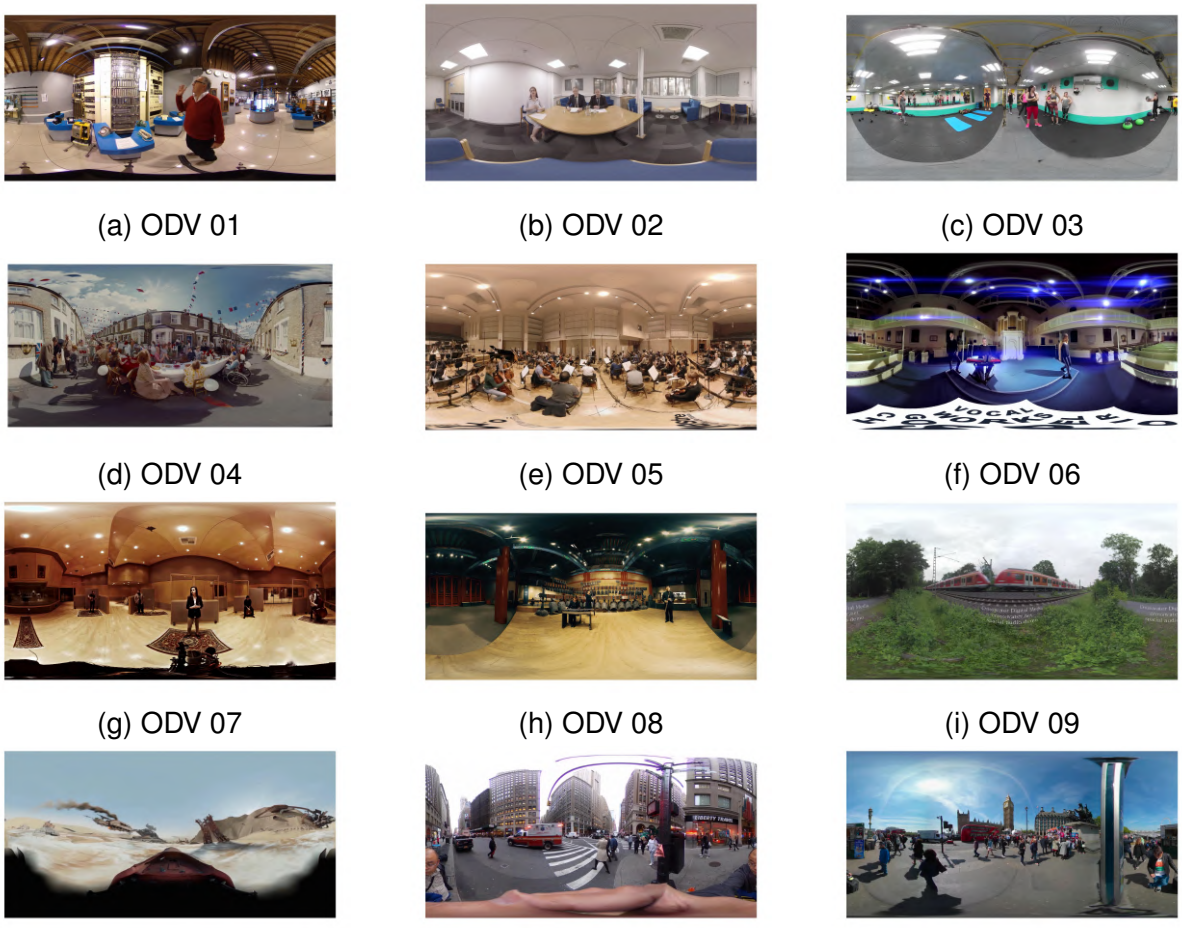
For this study, it was necessary that the audio-visual be as diverse as possible. Thus, the selected ODVs cover three representative categories of content: conversation, music, and environment. Videos in the category conversation consist of one person or several people talking, and those in the category music involve people playing instruments or singing. Videos in the category environment contain various background sounds such as vehicle engines and horns, the noise of crowds, and the sound of trains passing.

The source ODVs were all downloaded with the maximum available resolution and bitrate, which was 3840$\times$ 1920 in the ERP format. A representative segment of 25 s duration for each ODV was extracted and stored. The choice of segment was based on audio-visual cues in a pilot test with two experts. The 25 s duration is long enough for viewers to engage with the content and yet short enough to allow for a sufficient

number of experiments. To study the effect of audio-visual cues on the trajectories, the 15 ODVs were divided into 3 training ODVs and 12 testing ODVs. **Figure 3.1** summarizes the main features of the ODVs used in the dataset. In the figure, "Training Material" denotes the training set in each category. **Figure 3.2** presents a random selection of frames from the ODVs. The first ODV from each category was used as training material for the participants to familiarize themselves with the setup of the HMD.

	Dataset ID	Name	Fps	YouTube ID	Selected Segment
Conversation	**Training Material**	VoiceComic	24	5h95uTtPeck	00:30:10-00:55:10
	01	TelephoneTech	30	idLVnagj1_s	00:32:00-00:57:00
	02	Interview	50	ey9J7w98w1I	02:21:20-02:40:10
	03	GymClass	30	kZB3KMhqqyI	00:50:00-01:15:00
	04	CoronationDay	25	MzcdEI-tSUc	09:10:00-09:35:00
Music	**Training Material**	Chiaras	30	Bvu9m__ZX60	00:12:15-00:37:15
	05	Philharmonic	25	8ESEI0bqrJ4	00:40:00-01:05:00
	06	GospelChoir	25	1An41lDIJ6Q	00:09:10-00:34:10
	07	Riptide	60	6QUCaLvQ_3I	00:00:00-00:25:00
	08	BigBellTemple	30	8feS1rNYEbg	02:54:26-03:19:26
Environment	**Training Material**	Skatepark	30	gSueCRQO_5g	00:00:00-00:25:00
	09	Train	30	ByBF08H-wDA	00:20:10-00:45:10
	10	Animation	30	fryDy9YcbI4	00:01:00-00:26:00
	11	BusyStreets	30	RbgxpagCY_c	02:16:18-02:39:20
	12	BigBang	25	dd39herpgXA	00:00:00-00:25:00

Figure 3.1: Description of the ODVs used in the subjective experiment



(a) ODV 01 (b) ODV 02 (c) ODV 03

(d) ODV 04 (e) ODV 05 (f) ODV 06

(g) ODV 07 (h) ODV 08 (i) ODV 09

(j) ODV 10 (k) ODV 11 (l) ODV 12

Figure 3.2: ODV dataset used in the subjective experiment

3.1.2 Design of Apparatus

A test-bed was developed in which the participants could use an omnidirectional video player to watch a set of 360-degree videos in the three categories and under the three audio modalities. As they watched the videos, the participants' VCTs were recorded along with the current time-stamp, the ODV name, and its corresponding audio modality. The test-bed was implemented using JavaScript based on three APIs: three.js [26], WebVR [27], and JSAmbisonics [28]. These libraries enable users to watch a set of 360-degree videos via an HMD and provide a fully immersive experience through a web browser. The following technology was used in the experiment: a consumer version of the Oculus Rift as the HMD, Bose QuietComfort noise-cancelling headphones, and Firefox Nightly as the web browser.

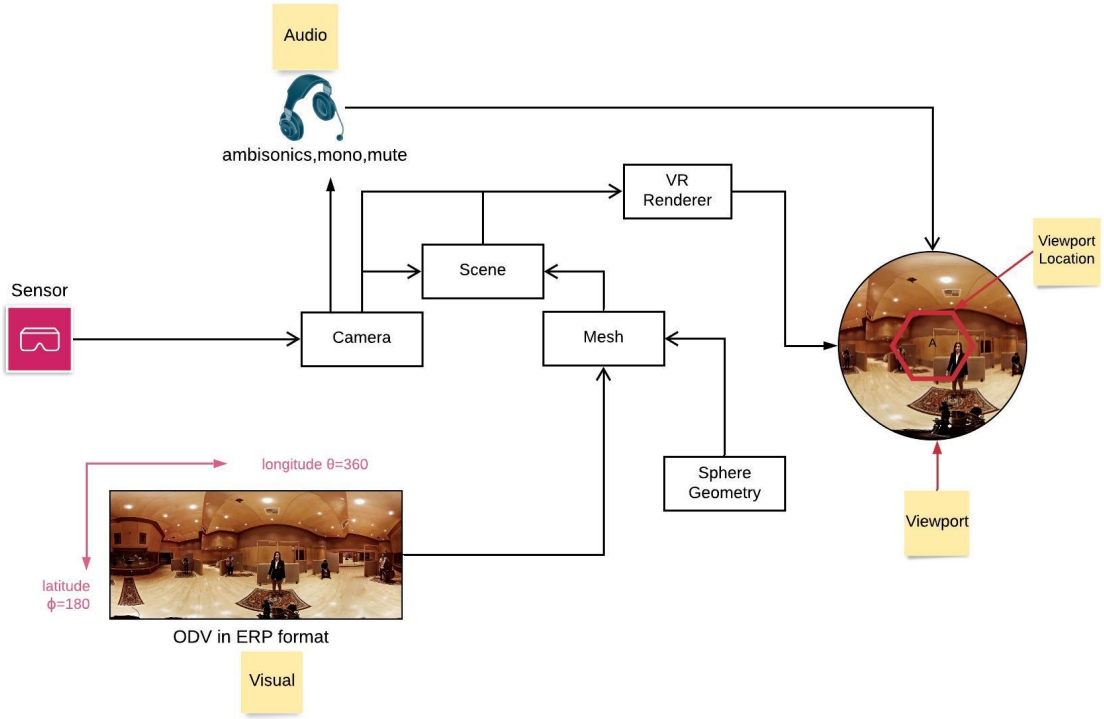


Figure 3.3: The schematic diagram of the designed test-bed

First, a given set of ODVs was loaded as the playlist file. While a given video was playing, the recorded data were transferred to the test-bed's server at the refresh rate of the device's graphics card. A schematic diagram of the test-bed for this experiment is shown in **Figure 3.3**. An Apache webserver with the MySQL database was used to implement the HTTP server. The audio-related (e.g., mute, mono, and ambisonics), sensor-related (e.g., viewing direction), and user-related (e.g., user ID, age, and gender) data were recorded and stored in the database.

3.1.3 Participants

Table 3.1 presents the demographic profile of the participants. The sample consisted of 45 participants, aged from 22 to 40 years, with an average age of 27 years. Fifteen (33%) are female, and eighteen (40%) are 20 to 24 years old. Eight of the participants (about 18%) were already familiar with VR technology and ODVs; the others had heard of VR before but had never used such technology. Twenty-four participants (around 55%) wore glasses during the experiment, and all of them were screened and reported having normal or corrected-to-normal visual acuity. **Figures 3.4** and **3.5** show the distribution of the participants' demographic information such as age and gender. Fifteen participants could watch each ODV per audio modality, and each participant watched each ODV content only once.

Age	Number
20-24	18
25-30	16
>30	11

(a) Age

Gender	Number
Male	30
Female	15

(b) Gender

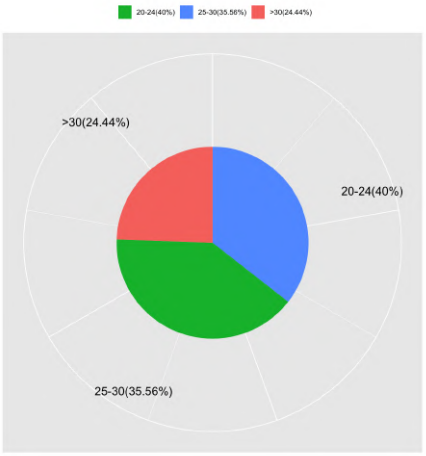
VR Experience	Number
Not Familiar	37
Familiar	8

(c) VR Experience

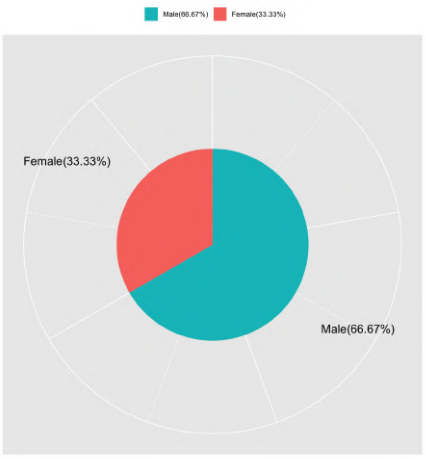
Glasses	Number
Yes	24
No	19

(d) Glasses

Table 3.1: Participants' demographic profile

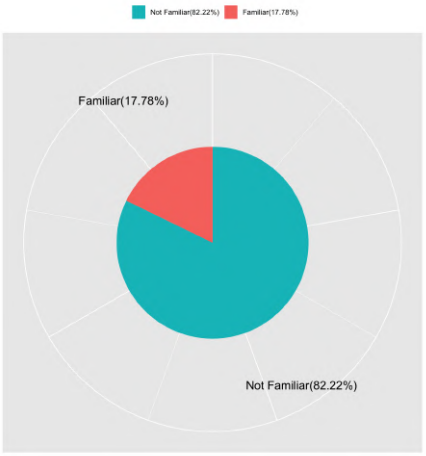


(a) Pie chart of age distribution

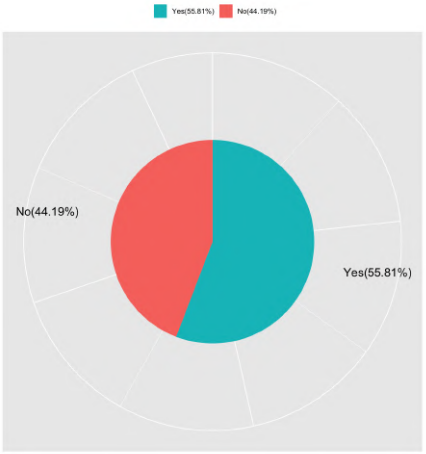


(b) Pie chart of gender distribution

Figure 3.4: Statistic on the profile of participants



(a) Pie chart of VR experience distribution



(b) Pie chart of wearing glasses distribution

Figure 3.5: Statistic on the profile of participants

3.1.4 Procedure

To avoid the memory bias effect, which may have affected the navigation trajectories, it was necessary to ensure that each participant watched each ODV content only once. Moreover, to balance the number of VCTs per audio modality for each ODV, three playlists were prepared. Each playlist contained a training ODV and four test ODVs per audio modality. Thus, there were 3 training ODVs and 12 ODVs for testing. The ODVs using the three audio modalities and the three content categories were allocated to the three playlists, and equal numbers of participants were assigned to the three playlists. Before the start of the subjective experiment, the test ODVs in each playlist were displayed in a random order while the VCTs of the participants were recorded.

Each test was performed as a task-free viewing session, and each participant was asked to look at each ODV naturally. During the experiment, participants were seated

in a rolling chair, which enabled them to turn freely in any direction. Moreover, they were asked to watch the ODVs alone in the laboratory to avoid the presence of the instructor influencing the results.

Before the experiment began, the procedure was explained to the participants, and they were informed that their VCTs were going to be recorded. The starting position of each viewing was fixed at the center point (θ = 180° and Φ = 90°) at the beginning of every ODV. Each participant was informed about the presence of a training video, which was displayed at the beginning of each playlist, so that they could familiarize themselves with the 360-degree viewing experience and adjust the HMD calibration. This was followed by a viewing session of 10 minutes in which the videos were displayed sequentially, separated by a gray screen. The screen was inserted to avoid motion sickness and eye fatigue and was shown for 5s between videos.

3.2 Statistical Analysis

3.2.1 Data Post-Processing

While users were watching the VR content with the HMD, only a portion of the ODV (i.e., the viewport) could be displayed at any given time. Since users' eyes tend to look straight ahead, with their heads following their eye movements to maintain the resting position of their eyes, it was decided to use the dynamicity of viewport center points to estimate the trajectories of users' gazes. VCTs in longitude ($0° \leq \theta \leq 360°$) and latitude ($0° \leq \Phi \leq 180°$) coordinates were recorded during the experiment. To analyze the collected VCTs, the interpolation was used to resample the collected data based on the frame rate of the corresponding video. In this way, each user is enabled to have a single value in longitude and latitude coordinates for a fair comparison in each frame of the ODV, thus ensuring the validity of the analysis, as shown below.

3.2.2 Data Analysis

To conduct the user behavior analysis, metrics were adopted such as the spatial distribution of the viewport center trajectories and the angular velocity:

- First, the areas that the participants were interested in are detected by analyzing the distributions of their viewport center traces in the longitudinal and latitudinal planes for various videos at different times;
- Second, the angular distance is used to evaluate the distance the viewer could travel over the sphere during the duration of an ODV segment. Here, the maximum angular distances of all viewers over different segment lengths for different

videos are calculated and presented in the cumulative density function (CDF) graphs;

- Finally, the angular velocities were calculated for all videos in all the categories and audio modalities. The purpose of this analysis is to reveal the change in users' navigation patterns and evaluate how fast they move their heads when watching a given ODV. The CDF of angular speed and mean angular velocity is used to show how the audio-visual information affects the user navigation patterns.

4 Results

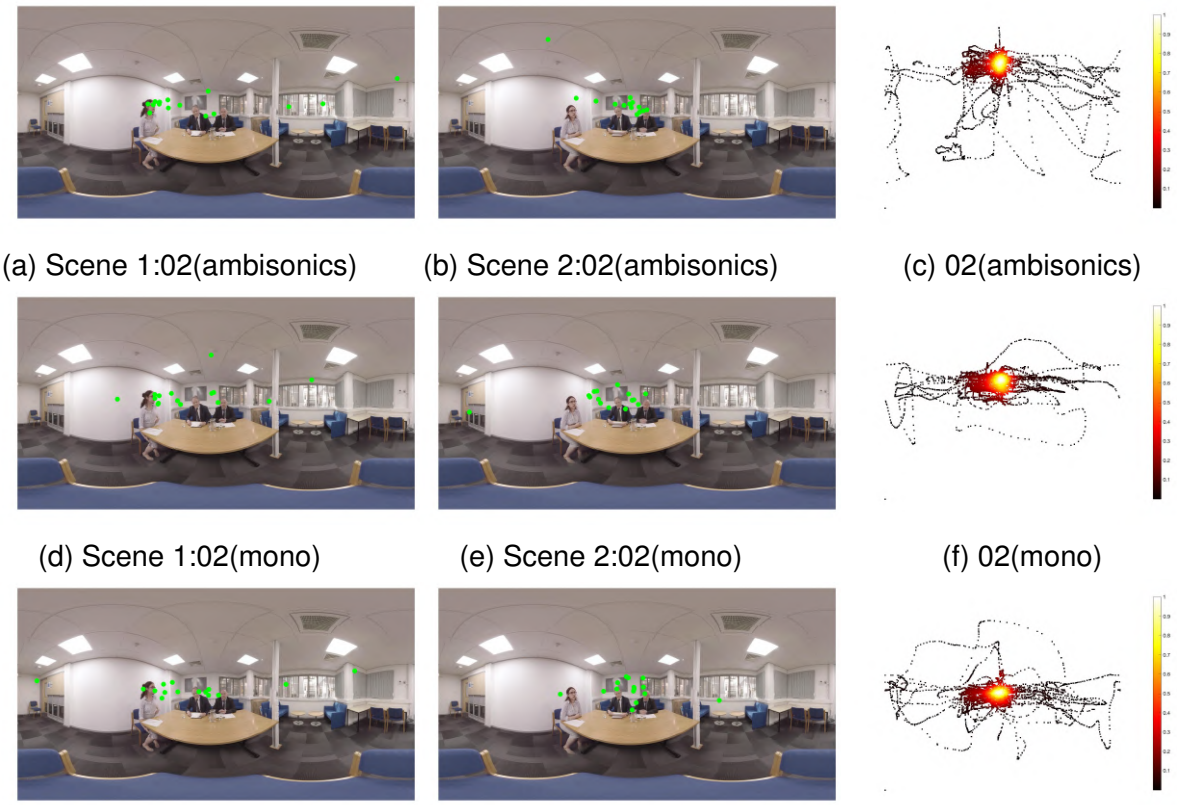
To understand how users explore different ODV content, this chapter presents the analysis of user behavior using the gathered navigation viewport center trajectories across the three audio modalities (mute, mono, and ambisonics) and in the three content categories (conversation, music, and environment). **Section 4.1** presents the analysis of the viewport center traces visualization. **Section 4.2** presents the head motion statistics.

4.1 Viewport Center Traces Visualization

This section describes how the viewport center locations were detected for each participant in the different video categories and audio modalities. Their fixation distributions over different videos at different scenes (i.e., different times) are analyzed. In addition, the sequence-level user fixation distributions are presented in heatmaps along with the longitude and latitude directions. In the fixation distributions, the green dots indicate participants' fixations; in the heatmaps, the color bars on the right show the intensity of the fixation distribution, and the yellow areas represent the high-density areas. 2D and 3D graphs are also plotted to analyze the traces over different videos.

4.1.1 Analysis of Fixation Distributions

Throughout the Interview (02) video from the category conversation (**Figure 4.1**), there are only slight changes in the fixation distributions under the three audio modalities. The fixation distributions for Scenes 1 and 2 under ambisonics, mono, and mute are similar. The focuses are almost on the area of three human faces which are in the central region of the ODV, as users' attention is driven by the salient visual and audio information (i.e., human faces and their sounds). The high-density areas in the heatmaps under the three audio modalities are all located on the center of the whole map, which is consistent with the previous observations of fixation distributions.



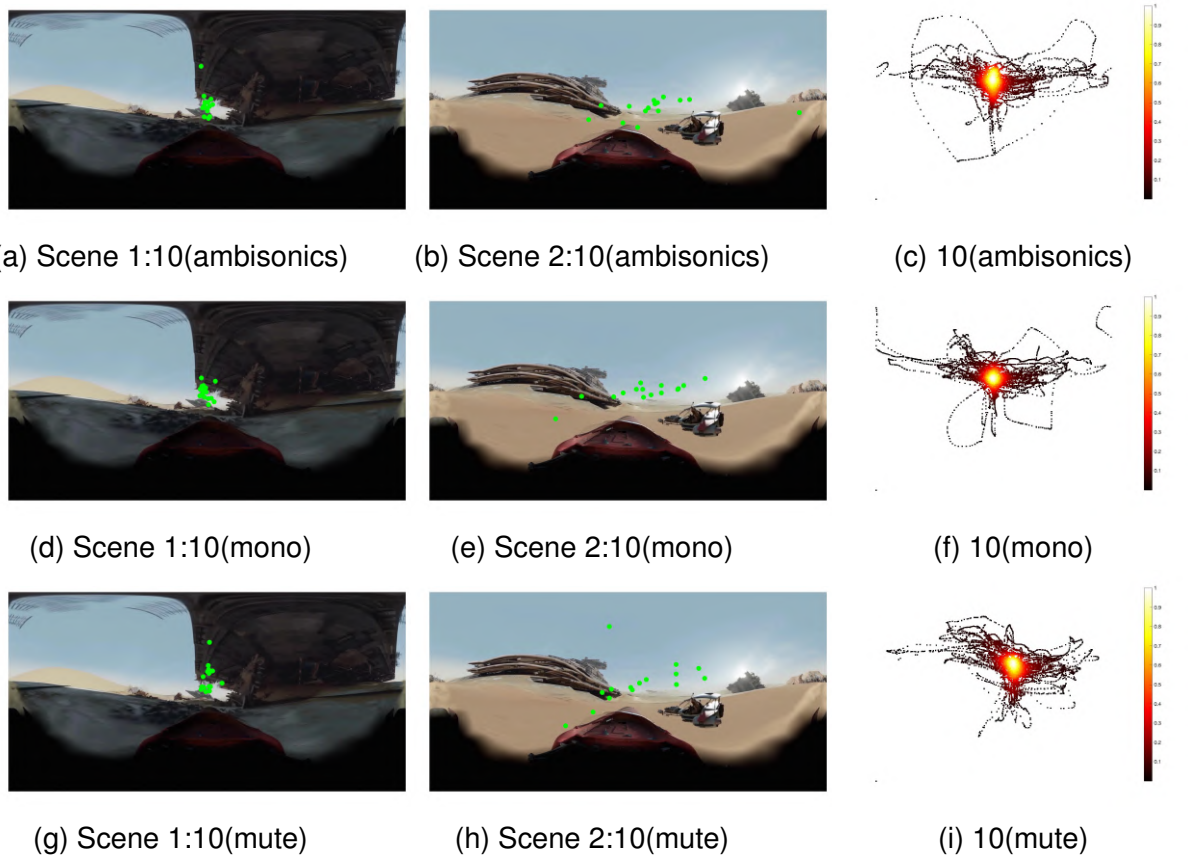
(a) Scene 1:02(ambisonics) (b) Scene 2:02(ambisonics) (c) 02(ambisonics)

(d) Scene 1:02(mono) (e) Scene 2:02(mono) (f) 02(mono)

(g) Scene 1:02(mute) (h) Scene 2:02(mute) (i) 02(mute)

Figure 4.1: Fixation distributions and heatmaps over Interview(02);
First and Second Columns: Viewer fixations at different scenes;
Third Column: Viewer trajectories heatmaps

The Animation (10) video belonging to the category environment depicts a series of rapidly changing background scenes captured by fast-moving cameras. **Figure 4.2** shows that in Scene 2 a racing car appears at high speed from the right, which catches the attention of some viewers, leading to greater dispersion of fixations than in Scene 1 under the three audio modalities. There is no difference between three heatmaps for the Animation video under the three audio modalities. The fixation distributions are more easily influenced by the salient visual cues (i.e., the fast-moving cameras) than the audio cues.



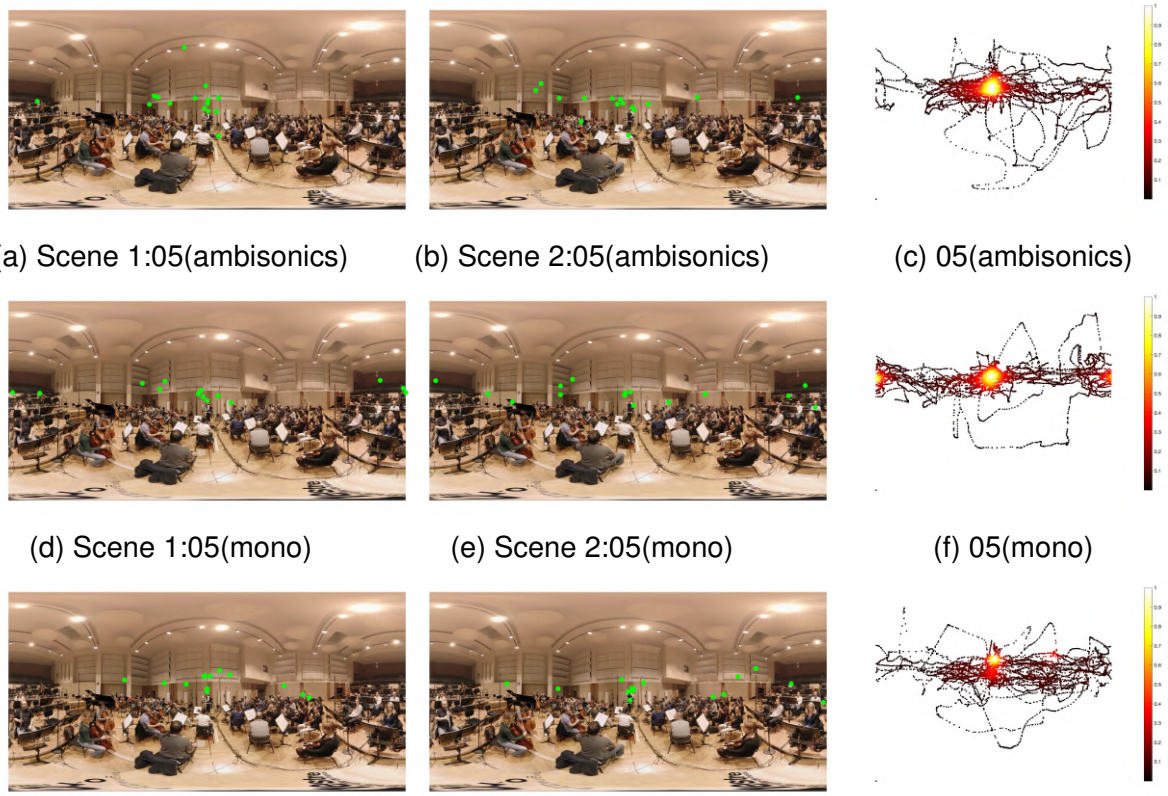
(a) Scene 1:10(ambisonics) (b) Scene 2:10(ambisonics) (c) 10(ambisonics)

(d) Scene 1:10(mono) (e) Scene 2:10(mono) (f) 10(mono)

(g) Scene 1:10(mute) (h) Scene 2:10(mute) (i) 10(mute)

Figure 4.2: Fixation distributions and heatmaps over Animation(10);
First and Second Columns: Viewer fixations at different scenes;
Third Column: Viewer trajectories heatmaps

For the Philharmonic (05) video in the category music, viewers mainly focused on the music director in Scene 1 under all three audio modalities (**Figure 4.3**). In Scene 2, the pianist outside the center of the field of view starts to play the piano. However, viewers' attention did not shift to the direction of the piano's sound in either ambisonics or mono. Moreover, in the heatmaps, the high-density areas under the three audio modalities are all in the center of the orchestra. It can be concluded that the moving objects (e.g., the conductor) have a more significant impact on users' attention than audio cues.



(a) Scene 1:05(ambisonics) (b) Scene 2:05(ambisonics) (c) 05(ambisonics)

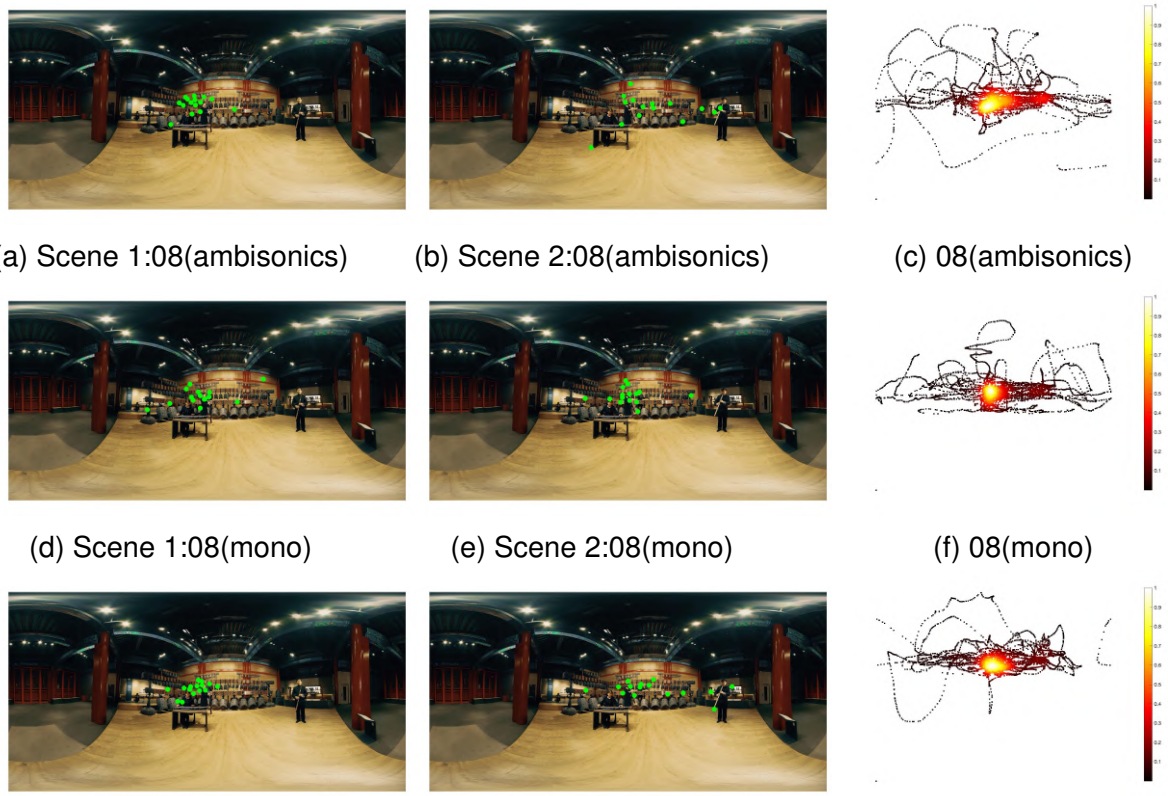
(d) Scene 1:05(mono) (e) Scene 2:05(mono) (f) 05(mono)

(g) Scene 1:05(mute) (h) Scene 2:05(mute) (i) 05(mute)

Figure 4.3: Fixation distributions and heatmaps over Philharmonic(05);
First and Second Columns: Viewer fixations at different scenes;
Third Column: Viewer trajectories heatmaps

For the BigBellTemple (08) video, all of the fixation distributions for the same scenes and heatmaps under the three audio modalities are similar to one another (**Figure 4.4**). The similarity in the fixation distributions is due to the subdued sound produced by the people playing musical instruments. Furthermore, there are no salient objects in this video, and so the areas of high-density in the heatmaps are concentrated on the central region of the view.

Based on the analysis of the Interview (02), Animation (10), and Philharmonic (05) videos, it was found that the salient visual cues, such as human faces, moving objects, and fast-moving camera motions, attract the user's attention more when presented along with audio information. In particular, the interaction of human voices (Interview) and salient visual objects catch more of the user's attention compared with the other two videos. Moreover, the comparison of the two category music videos, Philharmonic (05) and BigBellTemple (08), shows that users pay more attention to the louder sound produced by the instruments.



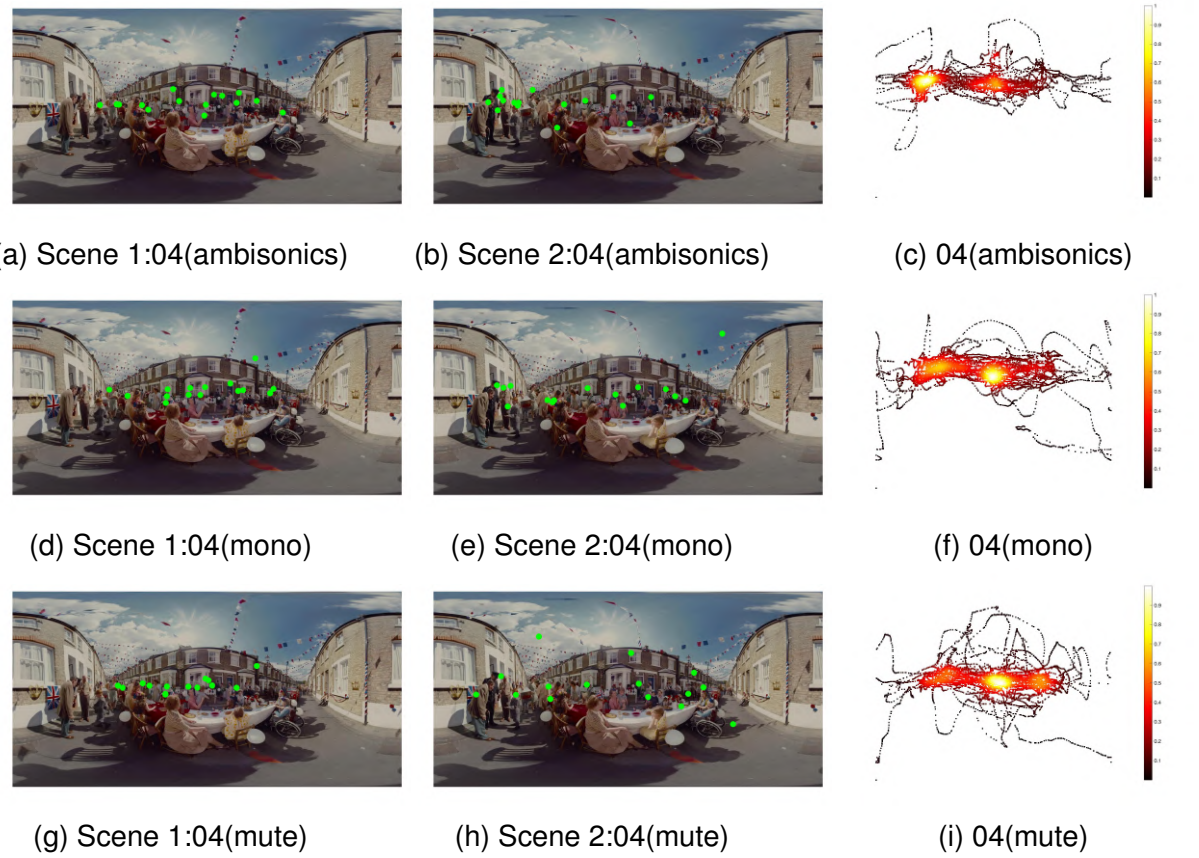
(a) Scene 1:08(ambisonics) (b) Scene 2:08(ambisonics) (c) 08(ambisonics)

(d) Scene 1:08(mono) (e) Scene 2:08(mono) (f) 08(mono)

(g) Scene 1:08(mute) (h) Scene 2:08(mute) (i) 08(mute)

Figure 4.4: Fixation distributions and heatmaps over BigBellTemple(08);
First and Second Columns: Viewer fixations at different scenes;
Third Column: Viewer trajectories heatmaps

In the CoronationDay (04) video of the category conversation, several people are talking at the back of the view area. At the beginning of the video (Scene 1), users tend to explore the environment more actively, which causes scattered fixations under the three audio modalities (**Figure 4.5**). In Scene 2, the four people standing in the back side of the viewing region start to talk, thereby catching more visual attention (i.e., more fixations) in ambisonics case, whereas in mono, viewers could only perceive the loudness of the audio. Thus, the number of fixations in mono is less than in ambisonics. The distribution of fixations in mute is relatively scattered throughout the video playback. These distributions are also shown in the heatmaps. The high-density areas are mainly concentrated on the talking persons in ambisonics. In contrast, in mono and mute, the high-density areas are more concentrated on the center field of the viewing area.



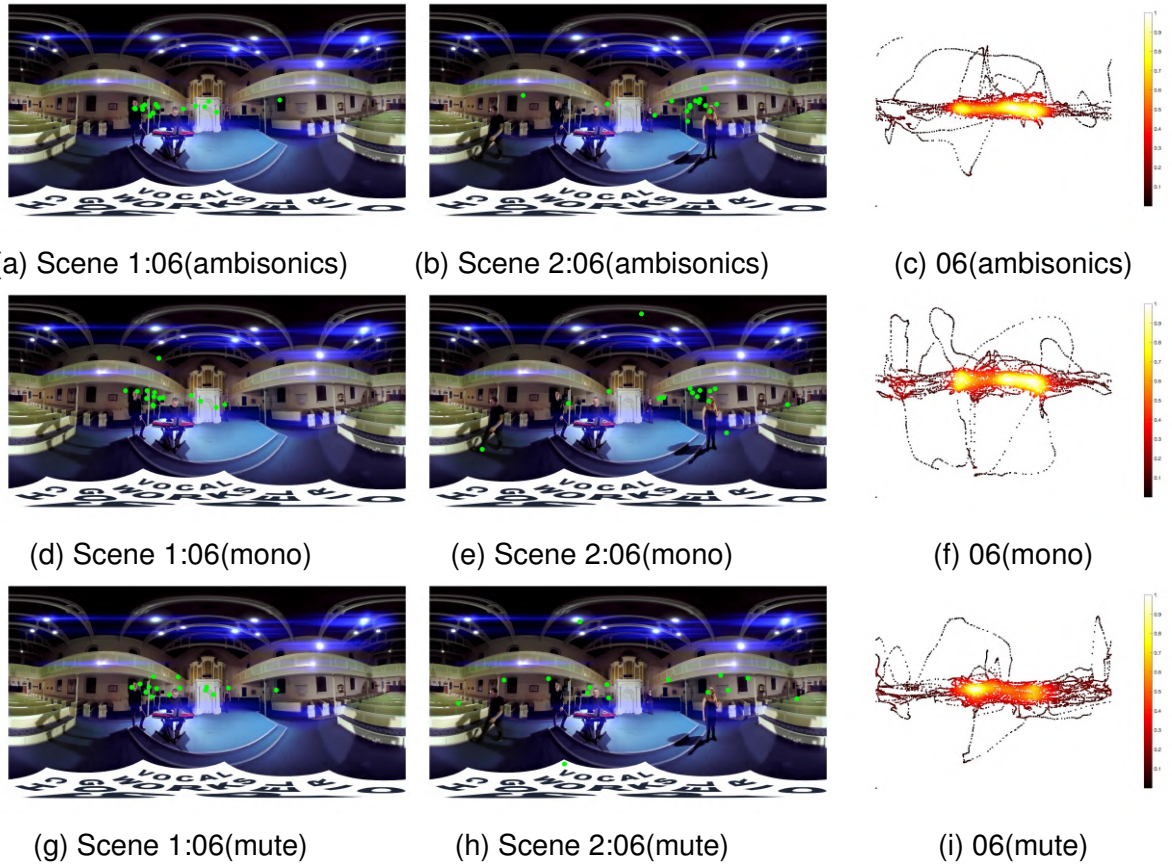
(a) Scene 1:04(ambisonics) (b) Scene 2:04(ambisonics) (c) 04(ambisonics)

(d) Scene 1:04(mono) (e) Scene 2:04(mono) (f) 04(mono)

(g) Scene 1:04(mute) (h) Scene 2:04(mute) (i) 04(mute)

Figure 4.5: Fixation distributions and heatmaps over CoronationDay (04);
First and Second Columns: Viewer fixations at different scenes;
Third Column: Viewer trajectories heatmaps

In Scene 1 of the GospelChoir (06) video in the category music, viewers focus their attention on the two human faces in all three audio modalities (**Figure 4.6**). Under ambisonics and mono, as a woman appears at the front and starts to sing, viewers' gazes gather on the position of the woman. In contrast, the distribution of fixations is more scattered in the mute modality. From the heatmaps, it can be observe that, under the three audio modalities, the high-density areas are much larger than those in the CoronationDay (04) video. This corresponds to that the fact that this video has multiple regions of interest. Besides, it confirms that the interaction of salient visual information (i.e., human faces) and audio information can attract more attention than visual cues alone.



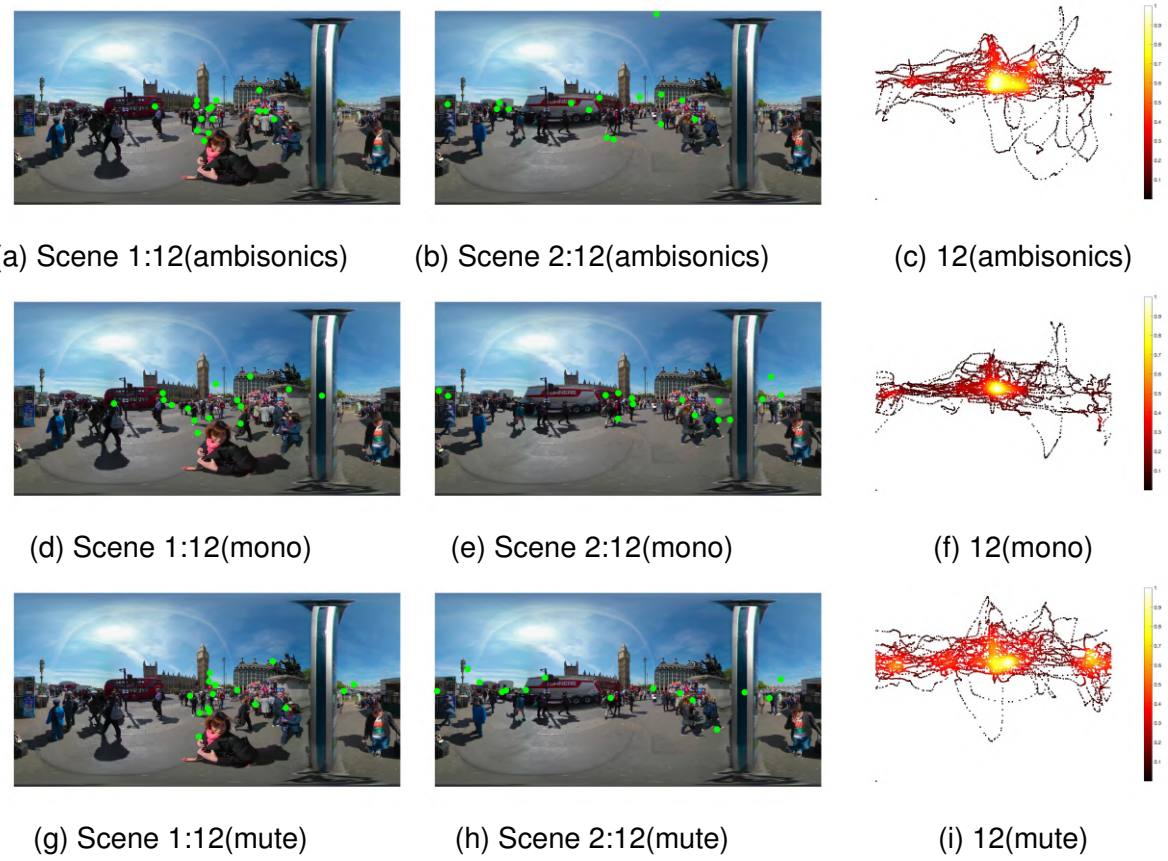
(a) Scene 1:06(ambisonics) (b) Scene 2:06(ambisonics) (c) 06(ambisonics)

(d) Scene 1:06(mono) (e) Scene 2:06(mono) (f) 06(mono)

(g) Scene 1:06(mute) (h) Scene 2:06(mute) (i) 06(mute)

Figure 4.6: Fixation distributions and heatmaps over GospelChoir (06);
First and Second Columns: Viewer fixations at different scenes;
Third Column: Viewer trajectories heatmaps

Compared with the CoronationDay (04) video (conversation category) and the GospelChoir (06) video (music category), the fixation distributions of the BigBang (12) video (environment category) under the three audio modalities are more scattered (**Figure 4.7**). In ambisonics and mono, users' gazes gathered on the sound direction of the vehicle engines. Moreover, from the heatmaps, it was found that the high-density areas in the mute modality are more scattered and much larger than in the ambisonics or mono modalities. This is due to the fact that moving objects such as pedestrians and vehicles on the streets attract viewers' more attention. These differences illustrate that the salient audio cues contribute to guiding the user's visual attention regardless of the loudness and direction of the sound.



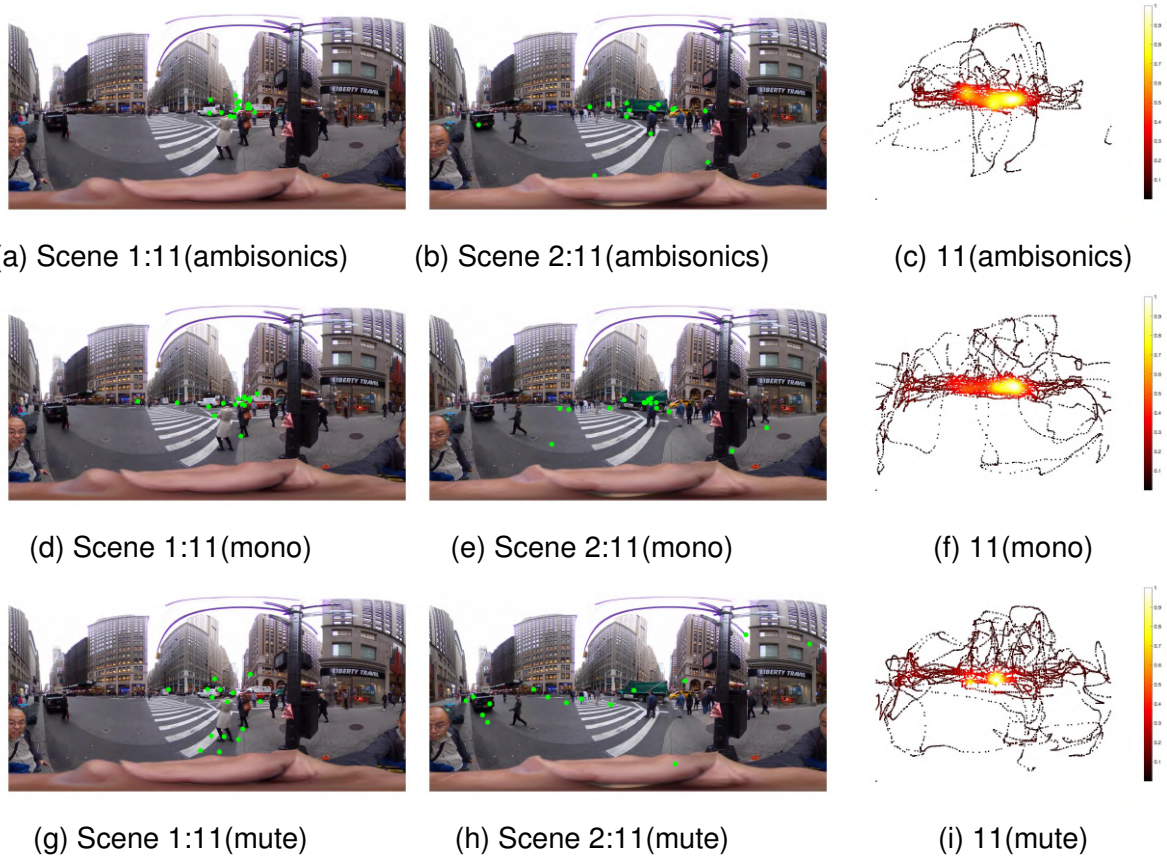
(a) Scene 1:12(ambisonics) (b) Scene 2:12(ambisonics) (c) 12(ambisonics)

(d) Scene 1:12(mono) (e) Scene 2:12(mono) (f) 12(mono)

(g) Scene 1:12(mute) (h) Scene 2:12(mute) (i) 12(mute)

Figure 4.7: Fixation distributions and heatmaps over BigBang (12);
First and Second Columns: Viewer fixations at different scenes;
Third Column: Viewer trajectories heatmaps

For the BusyStreets (11) video, the fixation distributions for the three audio modalities are similar to those of the BigBang (12) video. In the BusyStreets video, there is an ambulance siren, which makes the high-density area in the ambisonics modality (i.e. the direction of the ambulance siren) larger than those in the mono and mute modalities (see **Figure 4.8**).



(a) Scene 1:11(ambisonics) (b) Scene 2:11(ambisonics) (c) 11(ambisonics)

(d) Scene 1:11(mono) (e) Scene 2:11(mono) (f) 11(mono)

(g) Scene 1:11(mute) (h) Scene 2:11(mute) (i) 11(mute)

Figure 4.8: Fixation distributions and heatmaps over BusyStreets (11);
First and Second Columns: Viewer fixations at different scenes;
Third Column: Viewer trajectories heatmaps

Based on the analysis above, the following conclusions can be drawn:

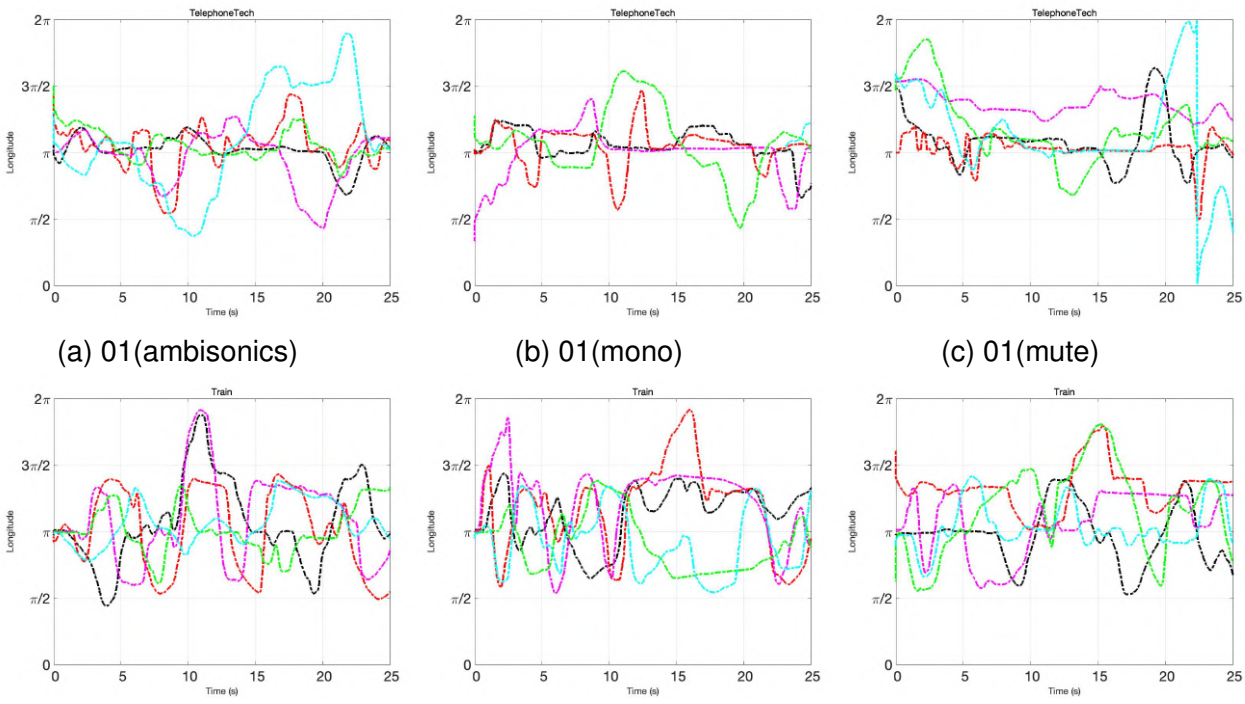
- First, the fixation distributions under the three audio modalities are different. In the mute modality, users' attention is more distracted than in the other modalities. Users' fixation distributions have a lower dispersion in the mono and ambisonics modalities. The loudness and direction of audio information help people track and find objects, even though the objects are not located in the current viewing area;
- Second, when both salient audio information, such as human voices and ambulance sirens, and visual information are presented, the audio information has a more significant influence on attracting users' visual attention compared with if only the visual information is presented;
- Third, there are also some differences in user behavior between the three categories. Users are more likely to follow audio information (e.g., human voices and the sound of a musical instrument) in the conversation and music categories than they are in the environment category, with the exception of clear audio in-

structions such as an ambulance siren. In videos in the environment category, users tend to explore the environment regardless of the audio information;

- Fourth, for different videos with three audio modalities and different video playback time, the users' fixation distributions are often different. Users' attention is mostly focused on the central region rather than the polar region, as can be seen from the heatmaps.

4.1.2 Analysis of View Traces

In the category conversation, the TelephoneTech (01) video shows a person talking on the phone in the center of the view area, and in the category environment the Train (09) video shows trains with engine noises driving at high speed from one side to the other. The 2D plots in **Figure 4.9** show that, for these two videos, there is a significant difference between the longitudes of the user trajectories. Specifically, in ambisonics and mono, the users' trajectories for the Train (09) video, which includes salient visual and audio information (i.e., fast-moving objects and loud engine noises), fluctuate in a wide range; in contrast, the trajectories for the TelephoneTech (01) video, which does not include such salient visual and audio cues, vary within a smaller range. It should be noted that the trajectories for both of these videos are more dynamic in the mute modality than in mono or ambisonics. In the absence of any audio cues, such as loudness and direction, users tend to explore the environment more.

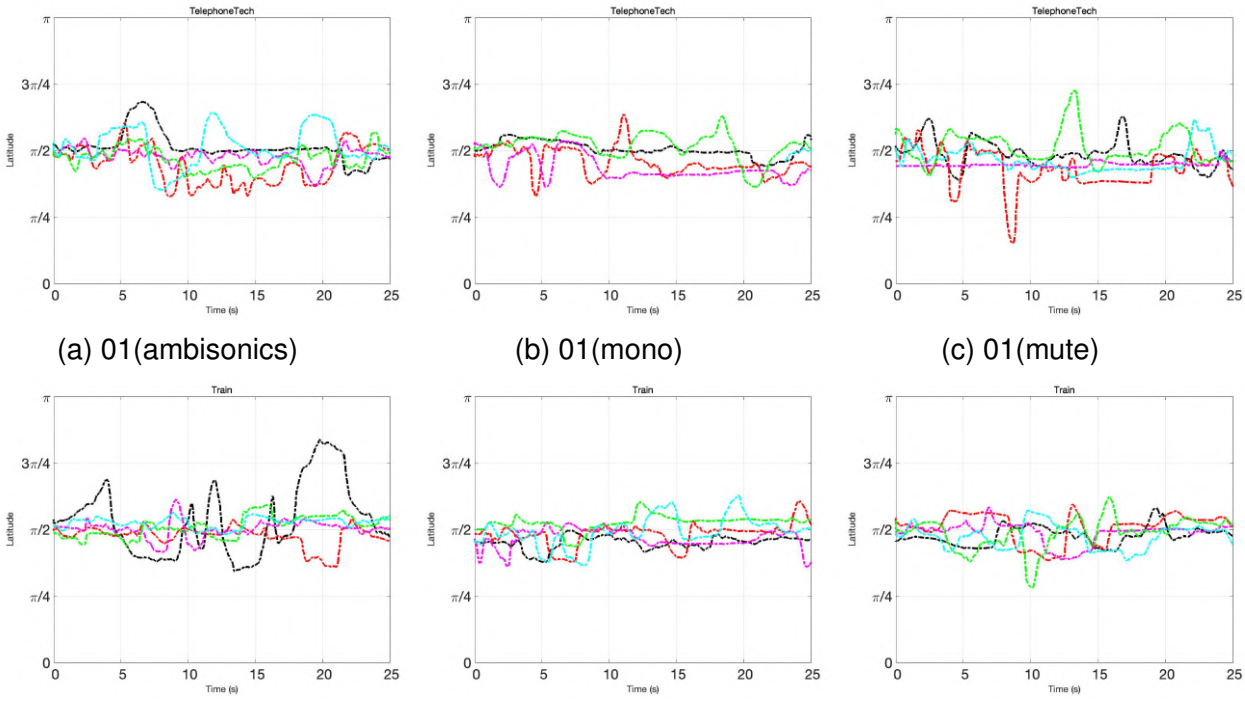


(a) 01(ambisonics) (b) 01(mono) (c) 01(mute)

(d) 09(ambisonics) (e) 09(mono) (f) 09(mute)

Figure 4.9: Viewport Center Trajectories 2D Plot in the longitude direction over TelephoneTech(01) and Train(09);
Each color represents a participant.

By comparing **Figure 4.9 and 4.10**, it can be seen that users' movements in the longitudinal direction are more predominant than those in the latitudinal direction when watching ODVs under the three audio modalities in all categories. This finding confirms the analyses presented in previous studies [18].

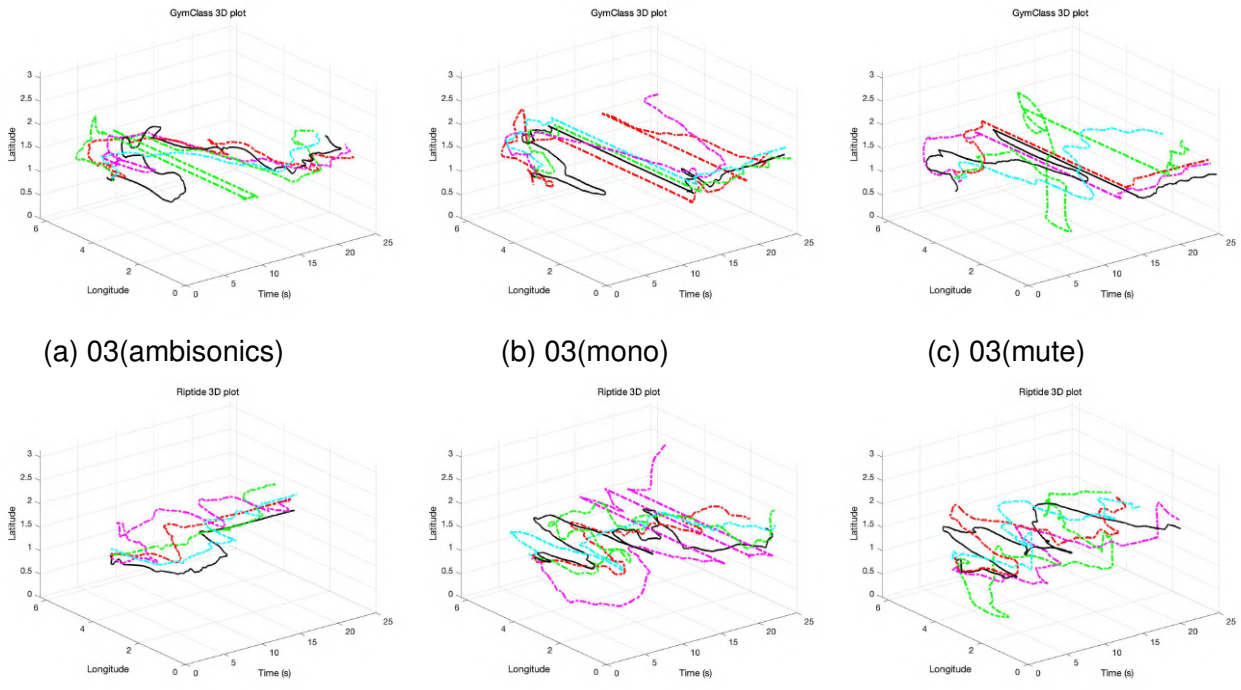


(a) 01(ambisonics) (b) 01(mono) (c) 01(mute)

(d) 09(ambisonics) (e) 09(mono) (f) 09(mute)

Figure 4.10: Viewport Center Trajectories 2D Plot in the latitude direction over TelephoneTech(01) and Train(09);
Each color represents a participant.

Finally, **Figure 4.11** presents the VCTs in 3D space of the GymClass (03) video from the category conversation and the Riptide (07) video from the category music. The first video includes scenes of several people talking while participating in gym classes, and the second video shows people singing in the center of the viewing area. Overall, it can be seen that for different videos with the three audio modalities in different categories, the users' viewport trajectories are often different.



(a) 03(ambisonics) (b) 03(mono) (c) 03(mute)

(d) 07(ambisonics) (e) 07(mono) (f) 07(mute)

Figure 4.11: Viewport Center Trajectories 3D Plot over GymClass(03) and Riptide(07);
Each color represents a participant.

4.2 Head Motion Statistics

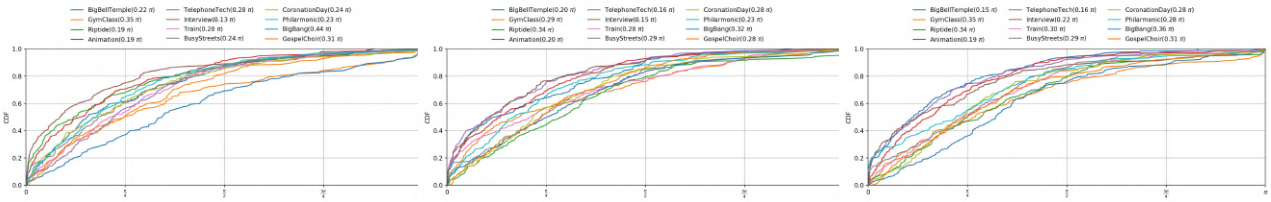
To determine the users' movement statistics, the metrics of angular distance and angular velocity were adopted. First, the maximum angular distance that the users could view over the sphere was calculated for each video under the three audio modalities and for different segment lengths. In addition, to reveal the change in users' navigation patterns and evaluate how fast the participants move their heads when watching a given ODV, the users' behavior was analyzed using the cumulative density function (CDF) of angular speed and the mean angular velocity for each user for different audio modalities and video categories.

4.2.1 Maximum Angular Distance

Figure 4.12 shows the CDF of the maximum angular distance within a time window of 2s for each video under the three audio modalities. There are differences between the distribution of all the videos with ambisonics, mono, and mute. For example, in ambisonics, over 70% of users move their heads no more than $\pi/2$ radians from the starting position within a time window of 2s. In mono and mute, this proportion is around 78%. The reason is that the audio information gives users a deeper feeling of

immersion.

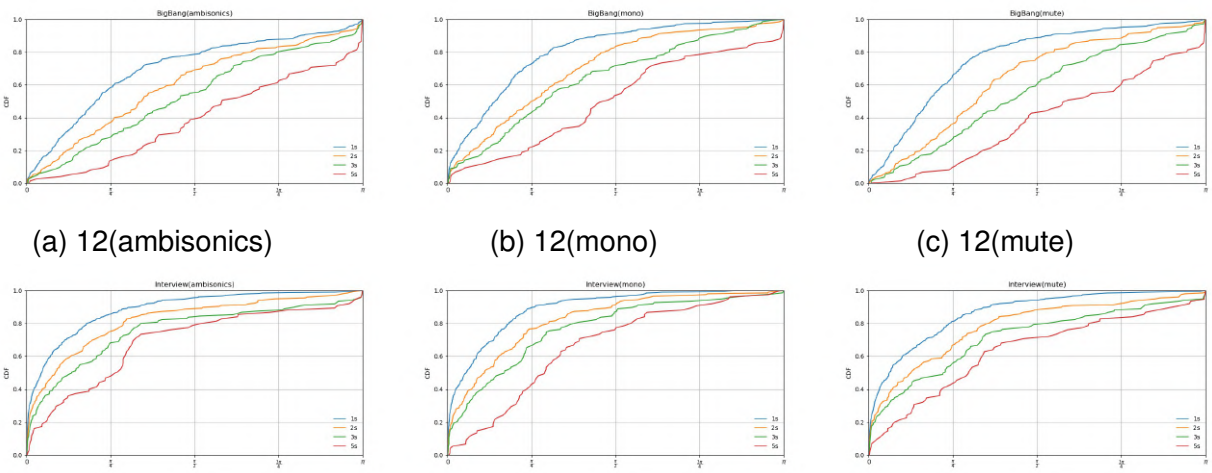
The average angular distances for each video under the three audio modalities were also calculated and are shown in the legends of **Figure 4.12**. The average angular distance of the BigBang (12) video in ambisonics is the highest (0.44π), while the value of the average angular distance of the Interview (02) video in ambisonics is 0.13π, which is the minimum. As suggested in [19], different segment lengths of 1 s, 2 s, 3 s, and 5 s were used to calculate the head motion statistics.



(a) Ambisonics (b) Mono (c) Mute

Figure 4.12: CDF of the maximum angular distance for each ODV under three audio modalities. Average angular distance for each video is also shown in the legend.

Figure 4.13 shows the CDFs of the maximum angular distances during time windows of 1 s, 2 s, 3 s, and 5 s for the BigBang(12) and Interview(02) videos under the three audio modalities. Under ambisonics, 90% of users who watch the Interview video do not move beyond $\pi/2$ within a 2 s time window, whereas over 30% of users who watch the BigBang video move beyond $\pi/2$ within the same duration. Under mono modality, the proportion of the users who move less than $\pi/2$ while watching Interview(91%) video is similar to the proportion who move less than $\pi/2$ while watching BigBang (85%) video within a 2-second time window. Moreover, within a segment duration of 5s, in the BigBang video, over 50% of users move more than $\pi/2$, whereas in the Interview video, around 70% of users move less than $\pi/2$ within the same duration. Furthermore, in mute modality, the proportion of the users who move less than $\pi/2$ while watching the Interview video is 90%, whereas for the BigBang video this proportion is around 78%.



(a) 12(ambisonics) (b) 12(mono) (c) 12(mute)

(d) 02(ambisonics) (e) 02(mono) (f) 02(mute)

Figure 4.13: CDF of the maximum angular distance over different segment lengths; Top: BigBang (12); Bottom: Interview (02)

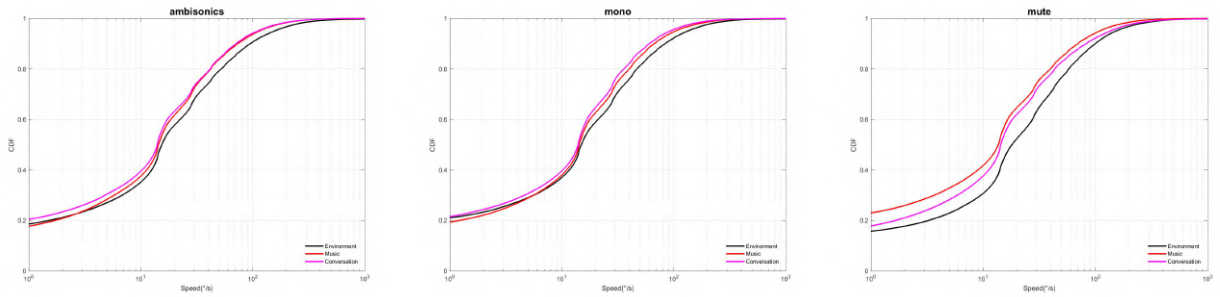
The analysis above confirms that the Interview video manages to drive users' visual attention toward the salient subject in the viewing area, which is consistent with the analysis in **Section 4.1.1**. More specifically, the Interview video contains salient visual and audio information (i.e., human faces and voices), which leads to a high level of consistency among users and lower angular distances than for the BigBang video. In contrast, the BigBang video, which belongs to the environment category, includes crowds, moving cars, and audio information from the surroundings, which causes users to tend to look around (i.e., the angular distances are higher).

4.2.2 Angular Velocity

Figure 4.14 presents the CDFs of the angular velocities of participants' head movements while watching all the ODVs in the three content categories under ambisonics, mono, and mute. In ambisonics, the angular velocity for the ODVs in the category environment is higher than those in categories conversation and music. For example, for ambisonics ODVs in the category environment, around 20% of the angular velocity values are beyond 50 degrees/second, whereas the proportions for ambisonics ODVs in the categories conversation and music are both about 15% (see **Figure 4.14a**).

Similarly, the proportion of users who have a high navigation speed in the category environment is higher than those in the categories conversation and music under mono and mute. This confirms that due to the lack of a main focus of attention in the ODVs in the category environment, the participants tend to change their navigation patterns more frequently and look around more than they do for ODVs in the categories conversation and music, without any influence of audio modalities.

Moreover, for the mute modality, there are clear differences between the distributions of the CDFs of the angular velocities of the three categories; whereas the distributions for the mono and ambisonics modalities are similar to each other. This shows that the audio information has a great effect on users' navigation patterns when watching different ODVs from different content categories.



(a) Ambisonics (b) Mono (c) Mute

Figure 4.14: CDF of the angular velocity of all the ODVs under ambisonics, mono, and mute in three content categories

Figure 4.15 shows the mean angular velocity for each video, and **Figure 4.16** shows the standard deviation. This analysis illustrates the dynamicity of users' behavior. The higher values indicate that users change their viewing direction more quickly. For example, the BigBang video in ambisonics has the highest mean angular speed, about 76 degrees/second, and the Interview video in mono has the lowest, 18 degrees/second (see **Figure 4.15**). From both figures, it can be seen that under the three audio modalities, the GymClass (3) video in the category conversation, the Philharmonic (05) video in the category music, and the BigBang (12) video in the category environment have larger mean angular speed and standard deviation values than the other ODVs, leading to more scattered navigation paths and more frequent viewing speed adjustment. In contrast, for the Interview (02) video in the category conversation, the BigBellTemple (08) video in the category music, and the Animation (10) video in the category environment, participants tend to follow and immerse themselves in the VR content, and thus the frequency of changes in the viewpoint center trajectories is lower.

Another observation is that, in the mute modality, each video has a higher mean angular speed and standard deviation, which confirms that in this modality the participants tend to explore the environment more and adjust the viewing speed more frequently. The analysis above corresponds to the previous analysis in **Section 4.1.1**.

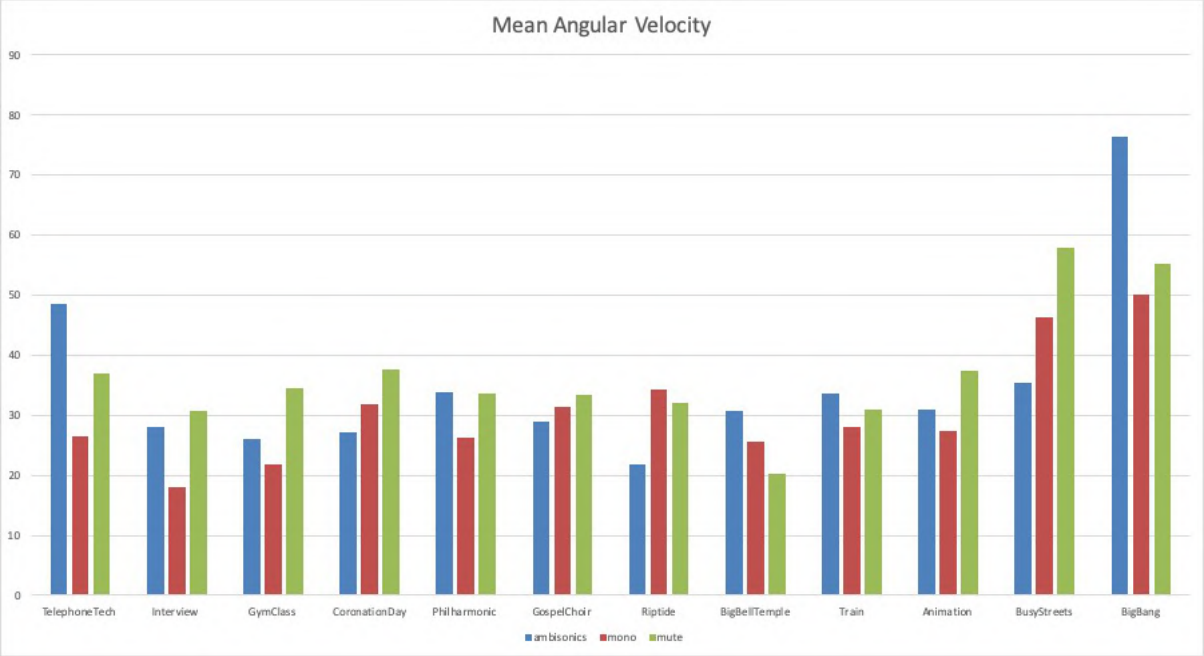


Figure 4.15: Mean velocity for all videos with ambisonics, mono, mute.

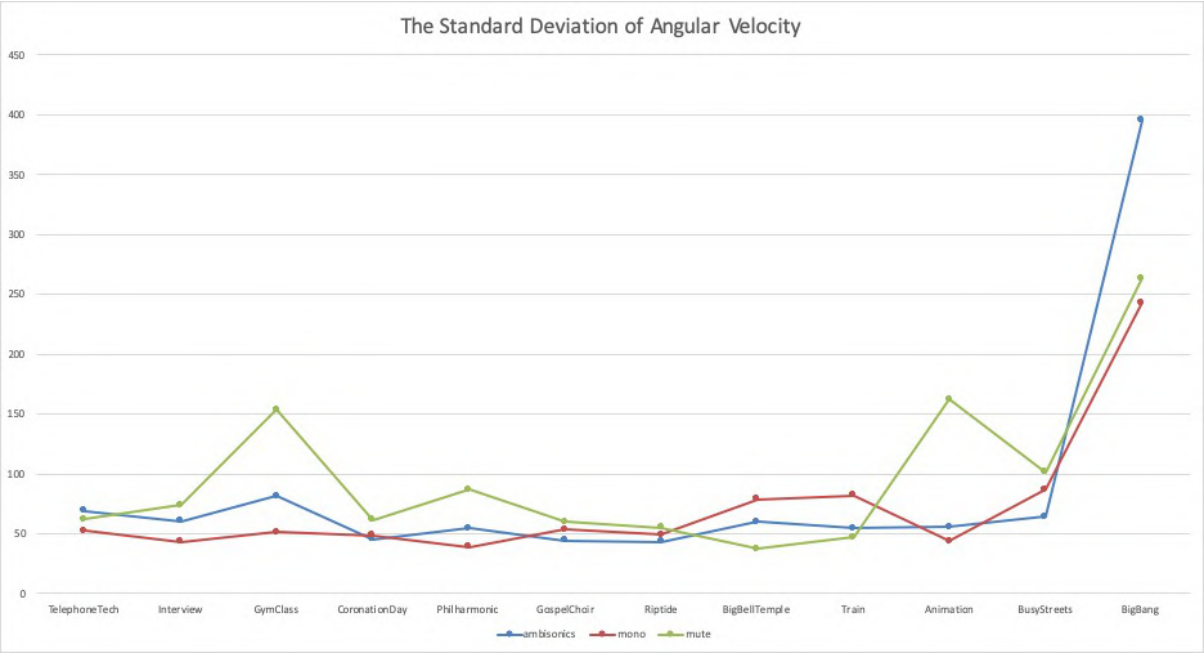
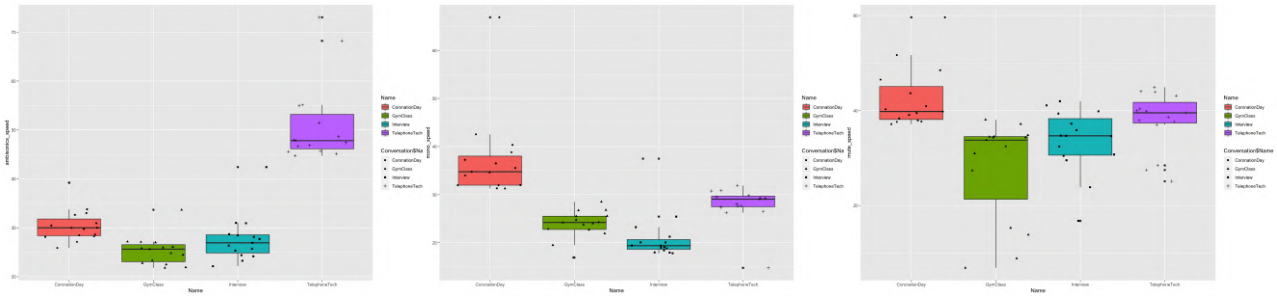


Figure 4.16: Standard deviation of speed for all videos with ambisonics, mono, mute.

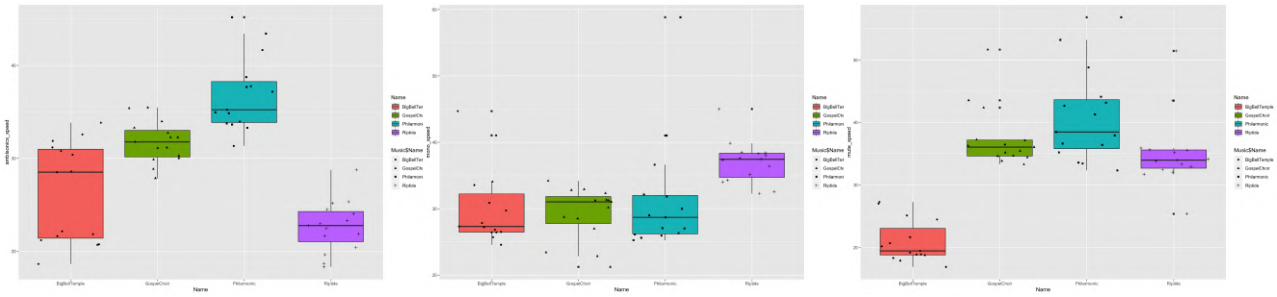
The overall distribution of users' mean angular velocity for each video under the three audio modalities is shown in **Figure 4.17,4.18,4.19**. It is worth noting that the distributions of mean speed per user in the category conversation under ambisonics and mono are more dispersed than under the mute modality. This shows that users present more varied mean speeds when watching ODVs in the category conversation in mute, which means that the participants exhibit a wider variety of navigation patterns

than they do in mono and ambisonics. In contrast, there are no clear differences in the distributions of users' mean speed while watching ODVs in the categories music and environment in the three audio modalities.



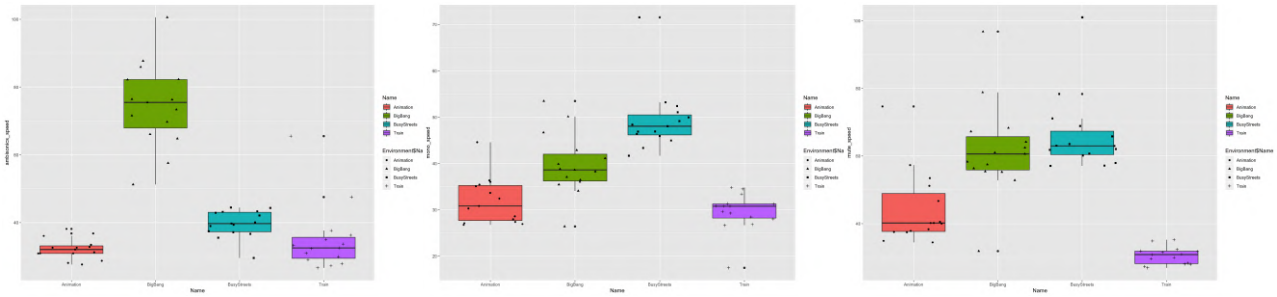
(a) Ambisonics (b) Mono (c) Mute

Figure 4.17: Boxplots per audio modality (ambisonics, mono, and mute) of mean angular speed for each video in the category conversation



(a) Ambisonics (b) Mono (c) Mute

Figure 4.18: Boxplots per audio modality (ambisonics, mono, and mute) of mean angular speed for each video in the category music



(a) Ambisonics (b) Mono (c) Mute

Figure 4.19: Boxplots per audio modality (ambisonics, mono, and mute) of mean angular speed for each video in the category environment

Here, two conclusions could be drawn:

- First, audio information, including audio direction and loudness, can influence the consistency of users' navigation patterns and behavior. More specifically, more users tend to change the viewing direction at a similar speed in mono and ambisonics than in mute – especially for ODVs in the category conversation, most of the users adjust their viewing navigation patterns at a similar velocity in mono and ambisonics. This finding illustrates that users are more sensitive to human voices than to sounds produced by instruments(i.e., category music) or vehicles(i.e., category environment);
- Second, users' navigation patterns or behaviors are often different for different kinds of video with different audio and visual information.

5 Conclusion

The aim of this study is to examine how audio-visual information affects users' behavior while watching 360-degree videos in three audio modalities (mute, mono, and ambisonics) in three different content categories. An audio-visual dataset was collected containing the VCTs from 45 participants and 12 ODVs. A comprehensive statistical analysis was performed using several metrics to analyze the users' behavior. From the results, three main conclusions can be drawn:

- First, salient audio cues help attract users' attention. The loudness and direction of audio information help users track and find objects, even if the objects are not located in the current viewing area. Moreover, salient visual cues attract users' attention to the regions in which they are located;
- Second, the viewport center distributions of users under the mute, mono, and ambisonics modalities are different from one another. In mute, the distribution of users' viewport center traces is more dispersed; in mono and ambisonics the navigation paths of users are more concentrated. This shows that, in the absence of sound, users tend to be more dynamic than they are in the presence of audio cues (loudness or direction);
- Third, differences in user behavior were also observed between the three categories (conversation, music, and environment). In ODVs in the category environment, users tend to look around and continuously change the viewing direction, regardless of the audio information.

6 Future Work

This dissertation contributes to the existing literature on the audio-visual perception of ODVs. The research tries to answer the research question in detail and leaves out room for further investigation.

Due to the limited time, there is room for further improvement in future work:

- The first thing could be improved is to design and conduct a new subjective experiment, for example, more participants would be recruited in this experiment, as is well known, more samples will help improve the accuracy of the analysis results;
- The second part that can be improved is to develop an algorithm to explore and predict the users' navigation patterns.

Bibliography

[1] F. Lopes, “Perceptual quality and bit rate models for omnidirectional video.”

[2] M. Xu, C. Li, S. Zhang, and P. Le Callet, “State-of-the-art in 360 video/image processing: Perception, assessment and compression,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 5–26, 2020.

[3] E. Van der Burg, C. N. Olivers, A. W. Bronkhorst, and J. Theeuwes, “Audiovisual events capture attention: Evidence from temporal order judgments,” *Journal of vision*, vol. 8, no. 5, pp. 2–2, 2008.

[4] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, “Assessing visual quality of omnidirectional videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3516–3530, 2018.

[5] U. SHUKLA, “An introduction to 360 video,” https://studio.knightlab.com/results/storytelling-layers-on-360-video/an-introduction-to-360-video/.

[6] C. Nachbar, F. Zotter, E. Deleflie, and A. Sontacchi, “Ambix-a suggested ambisonics format,” in *Ambisonics Symposium, Lexington*, 2011, p. 11.

[7] Waves, “Ambisonics explained: A guide for sound engineers,” https://www.waves.com/ambisonics-explained-guide-for-sound-engineers.

[8] S. Blog, “Understanding multi-channel audio and the stereo image,” https://splice.com/blog/multi-channel-audio-stereo-image.

[9] A. B. Araújo, “Drawing equirectangular vr panoramas with ruler, compass, and protractor,” *Journal of Science and Technology of the Arts*, vol. 10, no. 1, pp. 15–27, 2018.

[10] D. N. Tran and H.-J. Zepernick, “Spherical light-weight data hiding in 360-degree videos with equirectangular projection,” in *2019 International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2019, pp. 56–62.

[11] Y. Ye, E. Alshina, and J. Boyce, “Algorithm descriptions of projection format conversion and video quality metrics in 360lib,” *Joint Video Exploration Team of ITU-T SG*, vol. 16, 2017.

[12] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein, “Saliency in vr: How do people explore virtual environments?” *IEEE transactions on visualization and computer graphics*, vol. 24, no. 4, pp. 1633–1642, 2018.

[13] C. Ozcinar and A. Smolic, “Visual attention in omnidirectional video for virtual reality applications,” in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018, pp. 1–6.

[14] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet, “A dataset of head and eye movements for 360 videos,” in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, pp. 432–437.

[15] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, “What do different evaluation metrics tell us about saliency models?” *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 3, pp. 740–757, 2018.

[16] C. Wu, Z. Tan, Z. Wang, and S. Yang, “A dataset for exploring user behaviors in vr spherical video streaming,” in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 193–198.

[17] F. Duanmu, Y. Mao, S. Liu, S. Srinivasan, and Y. Wang, “A subjective study of viewer navigation behaviors when watching 360-degree videos on computers,” in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

[18] S. Rossi, C. Ozcinar, A. Smolic, and L. Toni, “Do users behave similarly in vr? investigation of the user influence on the system design,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–26, 2020.

[19] X. Corbillon, F. De Simone, and G. Simon, “360-degree video head movement dataset,” in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, pp. 199–204.

[20] S. Rossi, F. De Simone, P. Frossard, and L. Toni, “Spherical clustering of users navigating 360 content,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4020–4024.

[21] A. T. Nasrabadi, A. Samiei, and R. Prakash, “Viewport prediction for 360 videos: a clustering approach,” in *Proceedings of the 30th ACM Workshop on Network and Operating Systems Support for Digital Audio and Video*, 2020, pp. 34–39.

[22] A. Rana, C. Ozcinar, and A. Smolic, “Towards generating ambisonics using audio-visual cue for virtual reality,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2012–2016.

[23] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon, “Learning to localize sound sources in visual scenes: Analysis and applications,” *arXiv preprint arXiv:1911.09649*, 2019.

[24] H. R. Tavakoli, A. Borji, E. Rahtu, and J. Kannala, “Dave: A deep audio-visual embedding for dynamic saliency prediction,” *arXiv preprint arXiv:1905.10693*, 2019.

[25] F.-Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, “Audio-visual perception of omnidirectional video for virtual reality applications,” in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[26] Mr.doob, “Javascript 3d library,” https://threejs.org.

[27] WebXR, “Webxr device api specification,” https://github.com/immersive-web/webxr.

[28] D. P. Archontis Politis, “Jsambisonics,” https://github.com/polarch/JSAmbisonics.

A1 Appendix