Demo:A QoE and Visual Attention Evaluation on the Influence of Audio in 360° Videos

Amit Hirway
Department of Computer Engineering
Athlone Institute of Technology
Athlone, Ireland
a.hirway@research.ait.ie

Yuansong Qiao
Software Research Institute
Athlone Institute of Technology
Athlone, Ireland
ysqiao@research.ait.ie

Niall Murray
Department of Computer Engineering
Athlone Institute of Technology
Athlone, Ireland
nmurray@research.ait.ie

***Abstract*—360° video, also known as immersive video, is the recording of video content which simultaneously captures scene information in every direction, using an omnidirectional camera. Due to their immersive nature, the popularity of 360° videos has grown significantly. Understanding user Visual Attention when watching 360° videos is very important. This knowledge can help develop effective techniques for processing, encoding, distributing, and rendering 360° content. Whilst major efforts have concentrated on the visual element of immersive experiences, recently there has been growing interest in different forms of audio and in particular high-quality spatial audio. Spatial audio allows listeners to experience sound in all directions. Ambisonics or 3D audio is one such technique which offers a complete 360° soundscape. Although several models of visual and audio-visual attention have been proposed, very few have investigated the role of spatial audio in guiding attention in 360° videos. This demo shows our dataset and our methodological approach to understanding the user's audio-visual attention and QoE when experiencing 360° videos enhanced with spatial sound (first and third order ambisonic). Our research focus is to understand how audio affects Visual Attention in 360° videos and to evaluate its impact on the user's Quality of Experience (QoE).**

Keywords*—*360° Video, Spatial Audio, Ambisonics, QoE, Audio-Visual Attention

I. INTRODUCTION

Immersive media experiences involve immersive technologies such as Augmented Reality (AR), Virtual Reality (VR), Mixed Reality (MR) and 360° video. With such technologies, the physical world is emulated through a digital simulated world [1]. Virtual Reality aims to captivate users by delivering 360° visuals, immersive audio and environments in which the user can interact. Sound is an important part of immersive experiences. It contributes to immersion and presence [2]. However, previous research on "attention" in immersive media experiences has rarely taken the impact of the audio modality into consideration. Although public datasets [3][4][5] with user viewing behaviors (head-tracking, eye-tracking) while watching 360° videos are available, these are video-only or video with non-spatial audio datasets. In terms of attention and behavioral analysis, none of these works have considered the influence of spatial audio.

One of the biggest challenges in a 360° video is that the user can look in any direction. Head movement and eye gaze are important user behaviors, which can reflect a user's visual attention, preference, and even unique motion pattern. Adding spatial audio to the VR environment may completely change the way users watch the videos: how they move their heads; directions in which they focus; and what content they can remember after each session.

To date, we have built our own testbed for collecting traces from real viewers watching 360° videos (using a Head-Mounted Device (HMD)) with different types of audio (including spatial). We are currently collecting user responses via self-reporting and multimodal sensor data tracking [16] [18]. The resultant dataset can be used to predict which parts of 360° videos attract viewers to watch the most when the video is accompanied without and with different types of audio. This will help us to understand whether observers looking at videos with spatial sound have different eye movements than observers looking at the same videos when sound is not included or when the sound is non-spatial. The dataset can also be leveraged to compute the most common FoV (field of view) among viewers which may be used when the immersive media includes spatial sound. Table I mentions stages of the media delivery chain which can benefit from insights into this comprehensive dataset.

TABLE I. DATASET APPLICATION AREAS

Stage	*Benefit*
Multimedia Storage	Improving source coding
Multimedia Delivery	Improving resilience to transmission errors
Multimedia Compute	Optimizing field of view for 360° video streaming applications
Multimedia Quality	Video and Audio quality assessment

II. RELATED WORK

Understanding users' attention is critical to interpreting their behavior in VR. There has been some previous work related to Visual and Audio-Visual Attention wherein authors have recorded stimuli or used previously recorded stimuli to perform eye-tracking experiments. Also, they have evaluated QoE and published datasets for further research.

In [3], the authors studied subjects' 360° viewing behavior with an application that played several 360° videos. They collected both the orientation and rotation velocity of the HMD which was worn by 32 subjects whilst watching the videos. Each video was classified according to a set of categories: exploration; static focus; moving focus; rides; and a combination of these. The purpose of the categories were to help determine if video content produced different viewing patterns. It is common to see 360° videos with diverse characteristics available from YouTube employed for user tests and this was the case in [4]. They used an open-source head tracking tool, to record the viewer orientations, including yaw, pitch and roll from the HMD sensors. They also recorded and timestamped the viewer positions, including the x, y, and z coordinates.

In [5], the authors investigated the circumstances under which sound influenced visual attention. They collected a set of videos from YouTube to perform eye-tracking experiments. The experiments were performed with the same videos but in different test conditions: with and without soundtracks. Specifically, their focus was to understand the influence of non-spatial audio on visual attention with non-360° videos. The eye-tracking dataset proposed in [6] contains the eye positions gathered during four eye-tracking experiments. Observers were recorded whilst exploring video in different audio conditions (with or without sound). They also defined three categories of videos: moving objects; landscapes; and faces). The authors observed that audio-visual dispersion was always lower than visual dispersion. The presence or absence of sound seemed to influence the spatial distributions of eye positions for some visual categories. However, the media used for this study were non-360° videos with non-spatial audio.

In [7], how stereoscopic video and ambisonic sound contributed to the users' perceived QoE was studied. Participants were asked to rate specific aspects of their experience on a Likert scale in a questionnaire with rating of subjective quality measures. The questionnaire was based on imagery, sound, presence and motion sickness to produce an acceptability score, used as a measure of QoE. The study did not mention the level of ambisonic sound used - also, it did not intend to understand or evaluate subjects' Visual or Audio-Visual Attention during the experience. Based on the literature, there is a need to understand audio-visual cross-modal interaction for 360° spherical content with spatial audio (first and high-order Ambisonics [8]) and evaluate associated Quality of Experience [15].

III. METHODOLOGY

The research method employed is experimental and inspired by [9][10] and [17]. The user assessments are conducted in a controlled lab environment. Subjects for the actual experiment are expected to be in the age group of 20-50 years, with 50 percent gender distribution. For the pilot, 8 subjects had participated, two for each independent variable; for the actual experiment, a minimum of 10 subjects for each independent variable will participate. As per Table II, a number of different phases are defined which will facilitate within and between group statistical analysis.

TABLE II. METHODOLOGY, ACTIVITIES AND TOOLS

Phase, Duration	***Activity***	***Apparatus***
Informative, 10-min	Explain details of the test to the subject	Information sheet and consent form
Screening, 10-min	Assess visual acuity and color perception	Snellen Chart, Ishihara color blindness test
Training, 5-min	Get subject to be familiar and comfortable with the VR environment	Training Video
Testing, 15-min	Subject views a sequence of 360° videos with 1 of the 4 audio conditions	360 videos with different audio conditions viz. NS, ST, FO, TO*
Questions, 10-min	Subject answers questionnaire	Subjective Questionnaire

*NS-No Sound, ST-Stereo, FO/HO – First/Third Order

IV. TESTBED

Fig 1. shows a test subject in the lab as captured during the pilot. Details of the components used for the testbed are described in Table III. The stimuli for the experiment have been obtained from [11]. These have been selected from the many files located at [12] considering recording duration, content, categories, resolution and order of ambisonic sound. The stimuli have been categorized broadly into Indoor and Outdoor scenes. Further, these are subcategorized as Opera, Instrument, Riding and Exploration. The stimuli have been processed using ffmpeg tool to: a) set the maximum duration to 60 seconds; b) remove sound from the original videos; c) convert ambisonic sound to stereo. There are no narratives or subtitles in the videos. With inputs from [13] [14], a questionnaire with twenty questions has been developed to evaluate participants perception of presence, immersion, and spatiality of sound after watching the stimuli. Participants are asked to rate each question using the absolute category rating (ACR) system as outlined in [9]. The rating system will use a five-point Likert scale to determine if a user agreed or disagreed with the statements.

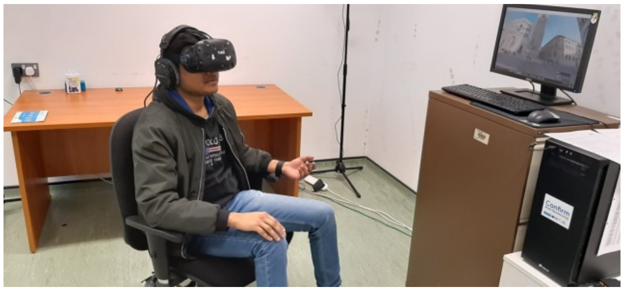

Fig. 1. Test subject wearing Vive HMD with integrated Tobii eye-tracker watching source stimuli. E4 wristband is strapped on wrist to record EDA

TABLE III. TESTBED COMPONENT DETAILS

Component	***Manufacturer***	***Used For***
HMD	HTC with Tobii Pro VR Integration	Watching 360° videos
Headphones	Beyerdynamic DT 990 Pro	Listening to non-spatial/ spatial audio
360° Player	GoPro VR Player	Head orientation capture
Wristband	Empatica E4	Recording EDA and heart rate
SDK	Tobii Pro Python SDK	Obtaining gaze origin, direction and pupil diameter

V. BRIEF ANALYSIS OF PILOT DATA

This section outlines some preliminary results from this early stage research. With reference to Fig. 2, from the first plot for head pose, it is evident that the subject movement was more prominent along the yaw-axis compared to the pitch and roll-axes. Also, in the second and third plots for left and right-eye gaze direction, movement along the yaw-axis is more prominent. In the fourth plot, there are various changes to pupil diameter. The content available at these times in the stimuli may be a reason for emotional arousal which causes the pupil diameter to change. From Fig. 3, the initial findings suggest that first and high order ambisonics outperform stereo on sound realism, localization, identification, attention retainment and enjoyment.

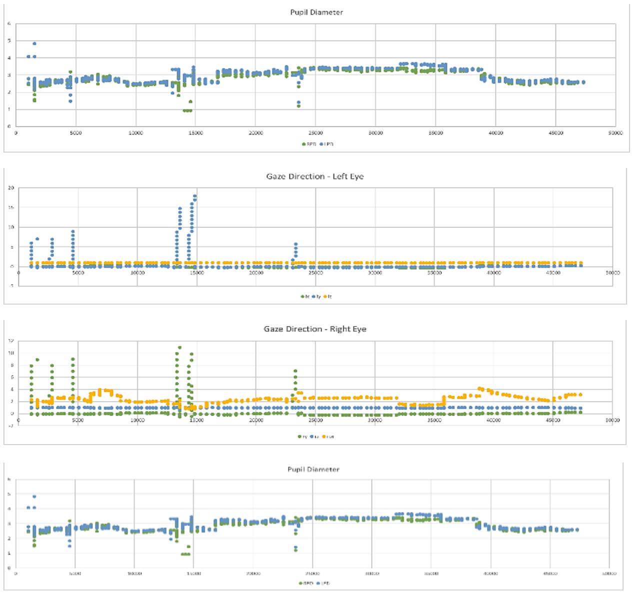

Fig. 2. Head Pose, Gaze Direction and Pupil Diameter of a subject for 60-sec duration of the stimuli

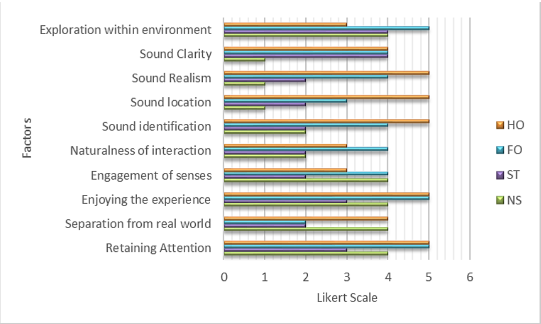

Fig. 3. Questionnaire Evaluation across the four audio conditions (8 subjects)

VI. DEMONSTRATION

Demonstration setup for the attendees will be using a conventional PC monitor, VR headset, external headphones and a 360-degree VR player (refer IV. TESTBED). During the demonstration, participants can experience the available 360° video sequences in any of the four audio conditions. After completion of the demo, sensor data from the HMD and eye-tracker and physiological data from the E4 wristband will be available for further analysis. Since the participants will be part of independent groups, we intend to use an appropriate statistical test like ANOVA to analyze the collected data using the IBM statistical analysis software package SPSS [19]. In terms of data interpretation, preliminary analysis using the SPSS package has given us some insights about the differences of each sound condition.

REFERENCES

[1] A. Suh and J. Prophet, "The state of immersive technology research: A literature analysis", Computers in Human Behavior, vol. 86, pp. 77-90, 2018.

[2] S. Poeschl-Guenther, K. Wall and N. Doering, "Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence", 2013, pp. 129-130.

[3] Almquist, M., & Almquist, V., "Analysis of 360° Video Viewing Behaviours", 2018.

[4] Lo, Wen-Chih & Fan, Ching-Ling & Lee, Jean & Huang, Chun-Ying & Chen, Kuan-Ta & Hsu, Cheng-Hsin, "360° Video Viewing Dataset in Head-Mounted Virtual Reality," 2017, pp. 211-216.

[5] Xiongkuo Min, Guangtao Zhai, Zhongpai Gao, Chunjia Hu and Xiaokang Yang, "Sound influences visual attention discriminately in videos," 2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX), Singapore, 2014, pp. 153-158.

[6] P. Marighetto et al., "Audio-visual attention: Eye-tracking dataset and analysis toolbox," 2017 IEEE International Conference on Image Processing (ICIP), Beijing, 2017, pp. 1802-1806.

[7] M. Milesen and V. Lind, "Quality Assessment of VR Film - A Study on Spatial Features in VR Concert Experiences", Aalborg University, Copenhagen, 2017.

[8] F. Zotter and M. Frank, Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality. Springer, 2019.

[9] "P.913 Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment", Itu.int, 2016. [Online]. Available: https://www.itu.int/rec/T-REC-P.913/en. [Accessed: 28- May- 2020].

[10] C. Keighrey, R. Flynn, S. Murray, and N. Murray, “A QoE evaluation of immersive augmented and virtual reality speech & language assessment applications”, 2017 9th Int. Conf. Qual. Multimed. Exp. QoMEX 2017, 2017.

[11] A. Farina, "Angelo Farina's Home Page", Angelofarina.it, 2020. [Online]. Available: http://www.angelofarina.it/. [Accessed: 28- May- 2020].

[12] A. Farina, "Index of /Public", Angelofarina.it, 2020. [Online]. Available: http://www.angelofarina.it/Public/. [Accessed: 28- May- 2020].

[13] J. M. Rigby, S. J. J. Gould, D. P. Brumby, and A. L. Cox, Development of a questionnaire to measure immersion in video media: The Film IEQ, TVX 2019 - Proc. 2019 ACM Int. Conf. Interact. Exp. TV Online Video, pp. 35–46, 2019.

[14] U. C. Lab, “Sheet PRESENCE QUESTIONNAIRE(PQ),” 2004.

[15] Qualinet White Paper on Definitions of Quality of Experience (2012). European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Patrick Le Callet, Sebastian Möller and Andrew Perkis, eds., Lausanne, Switzerland, Version 1.2, March 2013.

[16] C. Keighrey, R. Flynn, S. Murray and N. Murray, "A Physiology-based QoE Comparison of Interactive Augmented Reality, Virtual Reality and Tablet-based Applications," in IEEE Transactions on Multimedia, 2020.

[17] E. Hynes, R. Flynn, B. Lee and N. Murray, "A Quality of Experience Evaluation Comparing Augmented Reality and Paper Based Instruction for Complex Task Assistance," 2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP), Kuala Lumpur, Malaysia, 2019, pp. 1-6.

[18] C. Keighrey, R. Flynn, S. Murray and N. Murray,, "Comparing user QoE of AR and VR applications using physiological and interaction measurements," in 25th ACM International Conference on Multimedia (ACM MM 2017), Thematic Workshop, Oct 2017.

[19] “IBM SPSS - IBM Analytics,” IBM, [Online]. Available: https://www.ibm.com/analytics/us/en/technology/spss/. [Accessed 28 May 2020].