AUDIO-VISUAL PERCEPTION OF OMNIDIRECTIONAL VIDEO FOR VIRTUAL REALITY APPLICATIONS

Fang-Yi Chao[†], *Cagri Ozcinar*[‡], *Chen Wang*[‡], *Emin Zerman*[‡], *Lu Zhang*[†],
Wassim Hamidouche[†], *Olivier Deforges*[†], *Aljosa Smolic*[‡]
[†]Univ Rennes, INSA Rennes, CNRS, IETR - UMR 6164, F-35000 Rennes, France,
[‡]V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Ireland.

ABSTRACT

Ambisonics, which constructs a sound distribution over the full viewing sphere, improves immersive experience in omnidirectional video (ODV) by enabling observers to perceive the sound directions. Thus, human attention could be guided by audio and visual stimuli simultaneously. Numerous datasets have been proposed to investigate human visual attention by collecting eye fixations of observers navigating ODV with head-mounted displays (HMD). However, there is no such dataset analyzing the impact of audio information. In this paper, we establish a new audio-visual attention dataset for ODV with mute, mono, and ambisonics. The user behavior including visual attention corresponding to sound source locations, viewing navigation congruence between observers and fixations distributions in these three audio modalities is studied based on video and audio content. From our statistical analysis, we preliminarily found that, compared to only perceiving visual cues, perceiving visual cues with salient object sound (*i.e.*, human voice, siren of ambulance) could draw more visual attention to the objects making sound and guide viewing behaviour when such objects are not in the current field of view. The more in-depth interactive effects between audio and visual cues in mute, mono and ambisonics still require further comprehensive study. The dataset and developed testbed in this initial work will be publicly available with the paper to foster future research on audio-visual attention for ODV.

***Index Terms*—** Ambisonics, omnidirectional video, virtual reality (VR), visual attention, audio-visual saliency.

1. INTRODUCTION

With recent technological advancements in virtual reality (VR) systems, omnidirectional video (ODV), also known as 360° video, is an increasingly important multimedia representation to provide a high-quality immersive VR experience. The audio-visual representation of ODV is typically captured with omnidirectional microphone and camera systems. The audio part of ODV can be represented by spatial audio, *e.g.*, ambisonics, which is a description of a 3D spatial audio scene. The ambisonics format encodes the directional properties of the sound field to four or more fixed audio channels. The visual part of the ODV signal is typically stored in 2D planar representations such as equirectangular projection (ERP) to be compatible with the existing video technology systems. Thanks to its immersive and interactive nature, ODV can be used in different applications such as entertainment and education.

Although technical aspects of ODV have been widely investigated for different applications, many research questions are still open in the context of audio-visual perception of ODV. The need for understanding and anticipating human behavior while watching ODVs in VR is essential for optimizing VR systems, such as streaming [1] and rendering [2]. Towards this aim, recent visual attention/saliency research activities and studies for ODV have set a fundamental background for understanding users' behavior in VR systems. Most research works have investigated users' behavior with subjective experiments and developed algorithms for predicting users' visual attention. However, they have focused on visual cues only. Specifically, audio-visual perception of ODVs is highly overlooked in the literature. Creating immersive VR experiences requires full spherical audio-visual representation of ODV. In particular, the spatial aspect of audio might also play an important role in informing the viewers about the location of objects in the 360° environment [3], guiding visual attention in ODV films [4], and achieving presence with head-mounted displays (HMDs). To this end, in spite of the existing evidence on the correlation between audio and visual cues and their joint contribution to our perception [5], to date, most user behavior studies and algorithms for prediction of visual attention neglect audio cues, and consider visual cues as the only source of attention. The lack of understanding of the audio-visual perception of ODV rises interesting research questions to the multimedia community, such as *How does ODV with and without audio affect users' attention?*

To understand the auditory and visual perception of ODV, in this work, we investigated users' audio-visual attention using ODV with three different audio modalities, namely, mute, mono, and ambisonics. We first designed a testbed for gathering users' viewport center trajectories (VCTs), created a dataset with a diverse set of audio-visual ODVs, and conducted subjective experiments for each ODV with mute, mono, and ambisonics modalities. We analyzed visual attention in ODV with mute modality and audio-visual attention in ODV with mono and ambisonics modalities by investigating the correlation of visual attention and sound source locations, the consistency of viewing paths between observers, and distribution of visual attention in the three audio modalities. An ODV with ambisonics provides not only auditory cues but also the direction of sound sources, while mono only provides the magnitude of auditory cues. Users only perceive the loudness of the audio without audio direction in mono modality. Our new dataset includes VCTs and visual attention maps from 45 participants (15 for each audio modality), and our developed testbed will be available with this paper[1]. To the best of our knowledge, this dataset with such audio-visual analysis is the first to address the problem of audio-visual perception of ODV. We expect that this initial study will be beneficial for future research on understanding and anticipating human behavior in VR.

The rest of the paper is organized as follows. Section 2 discusses the related literature on visual attention studies for ODV. Section 3

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/27760.

[1]https://v-sense.scss.tcd.ie/research/360audiovisualperception/

978-1-7281-1331-9/20/$31.00 ©2020 European Union

describes the technical details of subjective experiments and post-processing, and Section 4 presents our analysis. Finally, Section 5 concludes the paper.

2. RELATED WORK

Although visual attention has been widely investigated for ODV in recent years, audio-visual perception of ODV has not been studied much for VR. Here, we briefly review recent ODV perception research, in particular, visual attention/saliency studies and algorithms for modeling the visual attention of ODV. For a comprehensive literature review on the analysis of ODV visual attention, we refer the reader to [6].

Analysis of visual attention of ODV based on eye and head tracking datasets aims to identify the most salient regions of ODVs. In particular, eye or head movements determine the areas of ODVs which are salient for users. For instance, David *et al.* [7] established a dataset for head and eye movements to understand how users consume ODV. Their study investigates the impact of the longitudinal starting position when watching ODV. Several algorithms have been proposed for modeling visual attention of ODV. In particular, the Salient360! Grand challenges at ICME 2017-2018 fostered the development of saliency prediction models for ODV by providing benchmark platforms and datasets [7]. Also, Zhang *et al.* [8] presented a large-scale eye-tracking dataset using only sport-related ODV. Their analysis demonstrates that salient objects (*e.g.*, appearance and motion of the object) easily attract the viewer's attention. Ozcinar and Smolic [9] analyzed content consumption of ODV viewed in HMDs. Their results prove that the quantity of fixations depends on the motion complexity of ODV. Furthermore, Nasrabadi *et al.* [10] investigated the impact of the type of camera motion and the number of moving objects in ODV using HMD navigation trajectories. Their results also reveal that users tend to look at moving targets in ODV. In the studies mentioned above, the audio signal was discarded from ODVs during subjective experiments. For realism and presence in VR, experiences should be multi-modal, but none of the above-mentioned perception studies proposes an audio-visual ODV dataset nor performs an analysis of human behavior on audio-visual ODV.

Recently, there has been increasing interest in the audio-visual aspects of ODV. For example, Rana *et al.* [11] and Morgado *et al.* [12] focused on generating ambisonics for ODV using different modeling strategies. They concentrate on utilizing texture and mono audio of ODV, predicting the location of the audio sources and encoding ambisonics. Also, Senocak *et al.* [13] proposed a unified end-to-end deep convolutional neural network for predicting the location of sound sources using an attention mechanism that is guided by sound information. Furthermore, a recent work conducted by Tavakoli *et al.* [14] proposed *DAVE* to investigate the applicability of audio cues in conjunction with visual ones in predicting saliency maps for standard 2D video using deep neural networks. Min *et al.* [15] proposed a multi-modal framework which fuses spatial, temporal and audio saliency maps for standard 2D video with high audio-visual correspondence. Their results show that the audio signal contributes significantly to standard video saliency prediction. However, to the best of our knowledge, neither audio-visual perception analysis nor audio-visual saliency prediction algorithms exist for ODV.

3. SUBJECTIVE EXPERIMENTS AND POST-PROCESSING

3.1. Design of testbed

We developed a JavaScript-based testbed that allows us to play ODVs with three different modalities (*i.e.*, mute, mono, and ambisonics) while recording VCTs of participants for the whole duration of the experiment. The testbed was implemented using three JavaScript libraries, namely `three.js` [16], `WebXR` [17], and `JSAmbisonics` [18]. The libraries of `three.js` and `WebXR` enable the creation of fully immersive ODV experiences in a browser, allowing us to use an HMD with a web browser. The `JSAmbisonics` facilitated spatial audio experiences for ODVs with its real-time spatial audio processing functions (*i.e.*, non-individual head-related transfer functions based on spatially oriented format for acoustics). The developed testbed can record VCTs without the need for eye tracking devices, which is an adequate use-case for many VR applications.As shown in Fig. 1, the developed testbed records participants' VCTs with the current time-stamp, name of ODV, and audio modality. At the front-end of the testbed, a `.json` file of a given set of ODVs is first loaded as the playlist file, and a given video is played while the recorded data is stored at the back-end of the testbed with the refresh rate of the device's graphics card. The HTTP server was implemented at the back-end using an Apache web server with the MySQL database, where 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 are stored in the database.

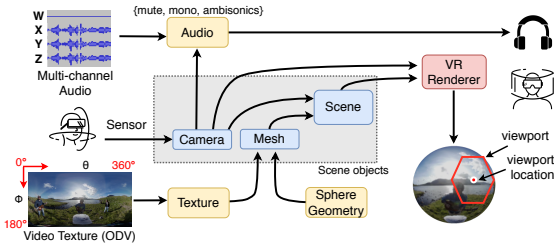

Fig. 1: Schematic diagram of the designed testbed.

3.2. Methodology

To equalize the number of VCTs per audio modality for each ODV, and to ensure that each participant watches each ODV content only once, three playlists were prepared. Each playlist included a training and four test ODVs per audio modality, so there were three training and twelve ODVs for testing. The ODVs with three different audio modalities, namely, mute, mono, and ambisonics, and three content categories were allocated to three playlists, respectively, and equal numbers of participants were distributed to the three playlists. The playing order of the test ODVs for each playlist was randomized before starting each subjective test.

Task-free viewing sessions were performed in our subjective experiments. All the participants were wearing an HMD, sitting in a swivel chair, and asked to explore the ODVs without any specific intention. In the experiments, we used an Oculus Rift consumer version as HMD, Bose QuietComfort noise-canceling headphones, and Firefox Nightly as web browser. During the test, VCTs were recorded as coordinates of longitude ($0^\circ \leq \theta < 360^\circ$) and latitude ($0^\circ \leq \Phi \leq 180^\circ$) in a viewing sphere. We fixed the starting position of each viewing as the center point ($\theta = 180^\circ$ and $\Phi = 90^\circ$) in the beginning of every ODV display. A 5-second rest period showing a gray screen was included between two successive ODVs to avoid eye fatigue and motion sickness. The total duration of the experiments was about 10 minutes. During experiments, participants were alone in the environment to avoid any influence by the presence of instructor.

3.3. Materials

Our dataset contains 15 monoscopic ODVs (three training and 12 testing) with first-order ambisonics in 4-channel B-format (W, X, Y,

Table 1: Description of the ODVs in our dataset.

	Dataset ID	ODV Name	Fps	YouTube ID	Selected Segment
Conversation	**Train**	*VoiceComic*	24	5h95uTtPeck	00:30:10 – 00:55:10
	01	*TelephoneTech*	30	idLVnagjI_s	00:32:00 – 00:57:00
	02	*Interview*	50	ey9J7w98wII	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	**Train**	*Chiaras*	30	Bvu9m_ZX60	00:12:15 – 00:37:15
	05	*Philarmonic*	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	**Train**	*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

and Z) collected from YouTube. In our experiment, ODVs in mute modality were produced by removing all audio channels, and ODVs in mono modality were produced by mixing four audio channels into one channel which can be distributed equally in left and right headphones. They are all 4K resolution (3840×1920) in ERP format, and 25 sec. segment length each. We divided ODVs into three categories, namely, *Conversation*, *Music*, and *Environment*, depending on their audio-visual cues in a pilot test with two experts. The category of *Conversation* presents a person or several people talking, the category *Music* features people singing or playing instruments, while the category *Environment* includes background sound such as noise of crowds, vehicle engines and horns on the streets. Table 1 summarizes the main characteristics of ODVs used in our dataset, where *Train* denotes the training set in each category, and Fig. 2 presents examples of each ODV. Also, Fig. 3 illustrates the visual diversity of each ODV in terms of spatial and temporal information measures [19], SI and TI, respectively. Each ODV is re-projected to cubic faces for computation of SI and TI to prevent effects from serious geometric distortion along latitude in ERP, as suggested by De Simone *et al.* [20].

3.4. Participants

Forty-five participants were recruited in this subjective experiment. Each ODV with each modality was viewed by 15 participants, and each participant viewed each ODV only once. These participants were aged between 21 and 40 years with an average of 27.3 years, and sixteen of them were female. Eight of them were familiar with VR, and the others were naïve viewers. All were screened and reported normal or corrected-to-normal visual and audio acuity, and 24 participants wore glasses during the experiment.

3.5. Post-processing

We recorded the VCTs of each participant in our subjective tests. As human eyes tend to look straight ahead [21] and head movement follows eye movement to preserve the eye resting position, similar as previous works [9, 10, 22, 23], we also consider the viewport center as an approximate gaze position for visual attention estimation. As observers do not see the complete 360° view at a glance, but only the content in the viewport, the information about the VCT is an important information for ODV applications, such as streaming [1, 22]. For analyzing visual attention, only fixations in raw scan-path data collected from VCTs are identified with a density-based spatial clustering (DBSCAN) algorithm [24]. We define a fixation as a particular location where successive gaze positions remain almost unmoved for at least 200 *ms*. To ignore the minor involuntary head movement and to reduce the sensitivity to noise, similar to previous ODV visual attention studies [9, 23, 25, 26], we utilized the DBSCAN algorithm to filter noisy fixation points.

Fig. 2: Examples for each ODV used in subjective experiments. Rows from top to bottom respectively belong to category: Conversation; Music; Environment.

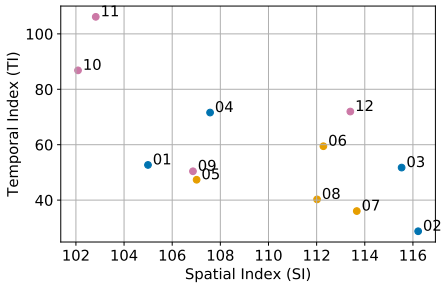

Fig. 3: SI and TI [19, 20] for each ODV used in subjective experiments. Each color visualizes each category: Conversation; Music; Environment.

After detecting all fixations of each ODV, we estimated a fixation map for t-th sec by gathering fixations of all the participants of a given ODV. Then, a dynamic visual attention map (*i.e.*, saliency map) of each ODV was generated by applying a Gaussian filter to its corresponding fixation map sequence. A Gaussian filter with σ visual angle was used to spread the fixation points to account for the gradually decreasing acuity from the foveal vision towards the peripheral vision. Based on the fact that gaze shifts smaller than 10° can occur without the corresponding head movement [21], we set σ to 5° according to the 68–95–99.7 rule in Gaussian distribution [23]. As only the viewport plane is displayed to an observer in HMD, we applied the Gaussian filter on the projected viewport plane rather than the entire ERP image.

4. ANALYSIS

In the following, we analyze the observer behavior while watching ODVs with three audio modalities (mute, mono, and ambisonics) in three content categories.

4.1. Do audio source locations attract attention of users?

To analyze the effect of audio information on visual attention when audio and visual stimuli are presented simultaneously, we measure how far visual attention corresponds to areas with audio sources under three audio modalities. We generate an audio energy map (AEM), representing the audio energy distribution with a frame-by-frame heat map. In AEM, the energy distribution is calculated with the help of given audio directions in four channels (W, X, Y, Z) in ambisonics as proposed in [12]. We then estimate normalized scan-path saliency (NSS) [27] to quantify the number of fixations that overlap with the distribution of audio energy via AEM. NSS is a widely-used saliency evaluation metric. It is sensitive to false positives and relative differences in saliency across the image [28].

Fig. 4 illustrates the mean and 95% confidence intervals computed by bootstrapping of NSS per user for each modality of ODVs. A higher NSS score indicates more fixations are attracted to areas of audio source locations with AEM, and negative scores indicate

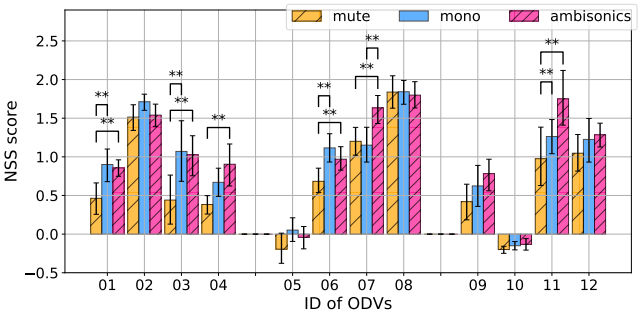

Fig. 4: Mean and 95% confidence interval of Normalized Scanpath Saliency (NSS) of fixations falling in sound sources areas under three audio modalities. ** marks statistically significant difference (SSD) between two modalities.

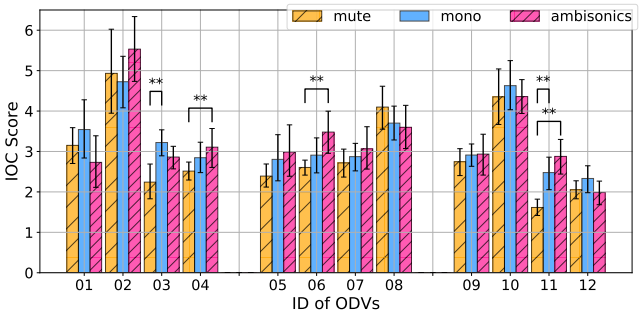

Fig. 5: Mean and 95% confidence interval of IOC based on NSS of each ODV with three audio modalities. ** marks statistically significant difference (SSD) between two modalities.

most fixations are not corresponding to areas of audio source locations. Numerical results show that either ambisonics or mono case has a greater NSS score than mute case. From Fig. 4, we observe that users may tend to follow audio stimuli (especially human voice) in categories conversation and music while they tend to look around in general regardless of the background sound in category environment. Notably, the two ODVs (*ODV 06, 07*) in the category music feature singing humans, while the others (*ODV 05, 08*) contain humans playing instruments. However, in category conversation, *ODV 02* obtains almost equal NSS scores in three audio modalities, which shows that visual attention could also be affected by the interaction of visual stimuli and audio stimuli depending on contents. In the category environment, *ODV 10* and *ODV 12* have similar NSS scores, while *ODV 09* and *ODV 11* have some difference. It appears that only *ODV 11*, which has an ambulance driving through with siren, obtains much higher NSS in ambisonics and mono than mute. It shows that hearing the siren and the sound direction of siren catches more attention than only seeing the ambulance.

To understand the significance of the NSS results, we performed a statistical analysis with a Kruskal-Wallis H Test following a Shapiro-Wilks normality test which rejects the hypothesis of normality of variables. Statistically significant difference (SSD) between two modalities is detected by the Dunn-Bonferroni non-parametric post hoc method. The pairs with a SSD are marked with ** in Fig. 4, which shows that there are three ODVs in category conversation, two ODVs in category music, and one ODV in category environment obtain SSD between mute and mono, and mute and ambisonics. The statistical significance analysis results are in line with our observations above. Furthermore, only one ODV has SSD between mono and ambisonics, which demonstrates that perceiving the direction of sound (*i.e.*, ambisonics) might not catch more attention than only perceiving the loudness of sound without directions (*i.e.*, mono) in most of ODVs.

For a visual comparison, Fig. 6 presents AEMs and fixations of two ODVs for each category. In this example, we show an ODV for each category (*ODV 04, 06, 11*) that receives statistically significantly higher NSS in ambisonics, and the other (*ODV 02, 05, 10*) receives almost equal NSS or negative NSS under three modalities. Looking at the figures, we can see that fixations are widely distributed along the horizon under mute modality and are more concentrated in AEMs under ambisonics modality. We can see in *ODV 04, 06, 11*, which obtain higher NSS in mono and ambisonics, feature talking or singing people or ambulance with siren outside the center field of view that can attract visual attention by object audio cues. However, in *ODV 02, 05, 10*, we observe that visual cues (*e.g.*, human faces, moving objects and, fast-moving camera) have more effect than audio cues on the distribution of fixations. For example, as seen in *ODV 02*, three human faces are very close to one another in the center of the ODV, and the users focused on the area of faces in all three modalities. In *ODV 05*, a moving object, which is a conductor in the center of an orchestra, has a more substantial contribution to visual attention than audio cues. Furthermore, in *ODV 10*, we see that the participants paid attention to the direction of camera motion regardless of the sound source location.

4.2. Do observers have similar viewing behavior in mute, mono, and ambisonics?

Observers' viewing behavior could exhibit considerable variance when consuming ODVs. Viewing trajectories might be more consistent to one another when observers perceive audio (*i.e.*, mono) or audio direction (*i.e.*, ambisonics). To investigate this, we estimate inter-observer congruence (IOC) [29], which is a characterization of fixation dispersion between observers viewing the same content. A higher IOC score represents lower dispersion implying higher viewing concurrency. NSS is used here to compute IOC as suggested in [30] to compare the fixations of each individual with the rest of observers. Statistical analysis was also conducted with the same methods as mentioned in Section 4.1.

Fig. 5 illustrates mean and 95% confidence intervals of IOC scores of each ODV in three modalities and SSD is marked as **. From the figure, it is shown that there is a significant difference between without sound (*i.e.,* mute) and with sound (*i.e.,* mono or ambisonics) cases. In particular, only in 4 out of 12 cases we observe a statistically significant difference between the two different cases, without sound and with sound. Moreover, we observe that visual attention is guided by object's sound to look for that object when observers do not see it in current field of view. For example, in category conversation, *ODV 03, and 04* featuring talking people in the back of viewing center receive significantly higher IOC between mute and mono, or mute and ambisonics, while the other two ODVs (*ODV 01, 02*) featuring talking people in the front that can be seen in the beginning of ODV display have no significant differences between three audio modalities. Similarly, in category music, *ODV 06* featuring people taking turns singing around the viewing center receives significantly higher IOC in ambisonics as it informs the direction of singing person unseen in the current filed of view to observers. However, in *ODV 07* which has singing people in the front and *ODV 05 and 08* featuring playing of instruments receive no significant IOC in three audio modalities. In category environment, *ODV 11* featuring an ambulance with siren driving from right to left obtains significantly higher IOC between mute and mono, and mute and ambisonics, while in other ODVs *ODV 09, 10, 12* having background sound from vehicle engines or crowds on the street obtain no

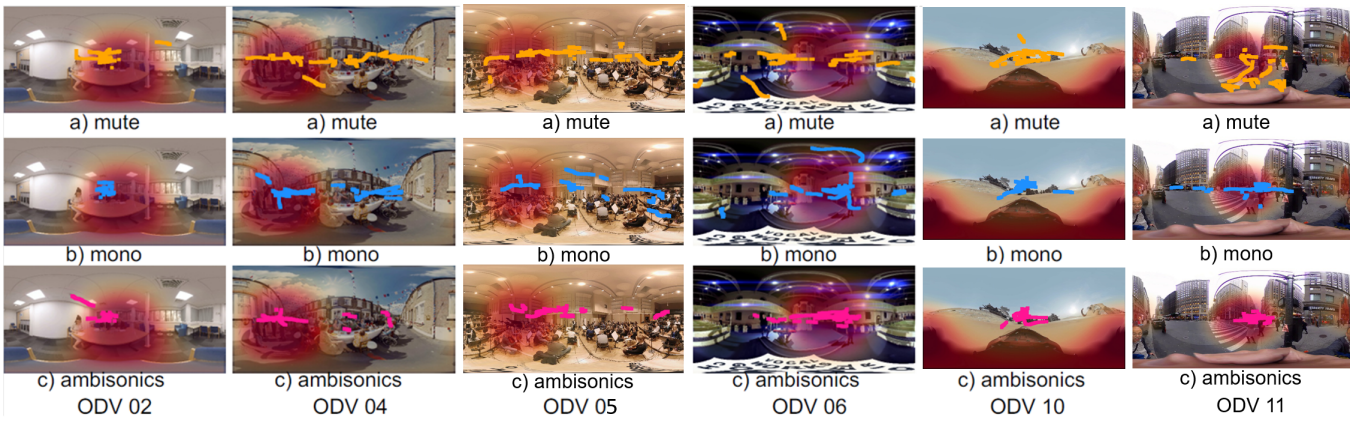

Fig. 6: A sample thumbnail frame with its AEM and fixations for each ODV, where the red represents AEM and the orange, blue, and pink denotes fixations recorded under none, mono, and ambisonics modality, respectively. A frame for each ODV ID from left to right: *02*, *04*, *06*, *08*, *09*, and *10*.

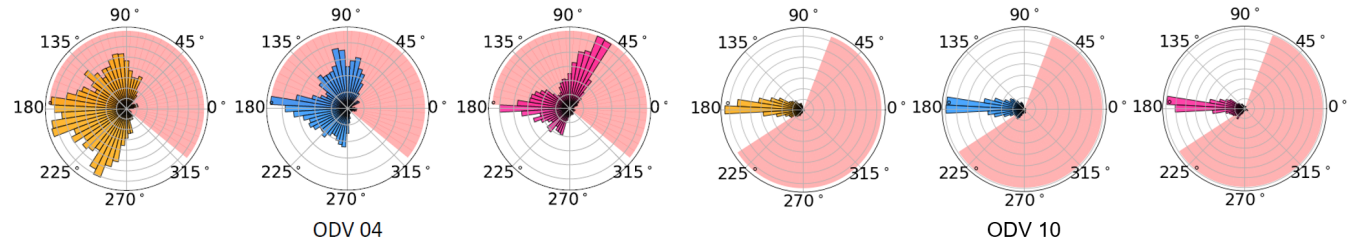

Fig. 7: Distribution of fixations and AEM in longitude of ODV 04 and 10. The orange, blue, and pink denotes fixations recorded under none, mono, and ambisonics modality, respectively. In each polar sub-figure, the longitude value of the ERP and its number of fixations (normalized) are respectively represented by the angle and the radius of the polar plot. Distribution of AEM is represented with red.

significant differences between three audio modalities. This demonstrates that perceiving object audio cues and the corresponding direction guides visual attention and increases consistency of viewing patterns between observers, when that object is not in the current field of view. Comparing the IOC scores between mono and ambisonics, we can see that the latter does not always receive higher scores in our subjective experiments. It shows that hearing the direction of sound (*i.e.*, ambisonics) does not certainly increase consistency of viewing patterns between observers, compared to only hearing the loudness of sound (*i.e.*, mono).

4.3. Does sound affect observers' navigation?

To study the impact of perception of audio (*i.e.*, mono) and audio direction (*i.e.*, ambisonics) to visual attention, we estimated the overall fixation distributions and overall AEM of all the frames. In most of the cases as shown in Fig. 6, the distribution of fixations for the ODVs with ambisonics modality is more concentrated. Fig. 7 shows the distribution of fixations and AEM in longitude of *ODV 04, 10* with three modalities. This figure shows that, in the *ODV 04*, the participants follow the direction of object audio with ambisonics case in a crowded scenario, where the main actors talking in the back side of observers are attracting visual direction in the crowded scene. In contrast, in *ODV 10*, the fixation distributions of three modalities are similar to each other and unrelated to the audio information. This is due to visual saliency of the fast moving camera, where most of visual attention corresponds to the direction of camera motion.

From our analyses in Sections 4.1, 4.2, and 4.3, we can generally conclude that when salient audio (*i.e.*, human voice and siren) is presented, it catches visual attention more than if only visual cues are presented. On the other hand, in some cases having salient visual cues (*i.e.*, human faces, moving objects, and moving camera), audio and visual information interactively affect visual attention. In addition, perceiving sounds and sound directions of salient objects can guide visual attention and achieve higher IOC, if these objects are not in the current field of view.

Although this study reveal several initial findings, more studies are required to support the open research questions raised with this work. In particular, "does the directions of sounds lead higher viewing congruence than mono sound?" and "does the directions of sounds guide visual attention more than mono sound?" are still not confirmed due to limited number of participants. For this purpose, we plan to conduct more comprehensive subjective experiments (in terms of number of participants and diverse ODVs), and we plan to further investigate these questions with statistical tests.

5. CONCLUSION

This paper studied audio-visual perception of ODVs in mute, mono, and ambisonics modalities. First, we developed a testbed that can play ODVs with multiple audio modalities while recording users' VCTs at the same time, and created a new audio-visual dataset containing 12 ODVs with different audio-visual complexity. Next, we collected users' VCTs in subjective experiments, where each ODV had three different audio modalities. Finally, we statistically analyzed the viewing behavior of participants while consuming ODVs. This is, to the best of our knowledge, the first user behavior analysis for ODV viewing with mute, mono, ambisonics.

Our results show that in most of cases visual attention disperses widely when viewing ODVs without sound (*i.e.*, mute), and concentrates on salient regions when viewing ODVs with sound (*i.e.*, mono and ambisonics). In particular, salient audio cues, such as human voices and sirens, and salient visual cues, such as human faces, moving objects, and fast-moving cameras, have more impact on visual attention of participants. Regarding audio cues, the nature of the sound (*i.e.*, informative content, frequency changing, performance

timing, audio ensemble) may also play a role in how it gets noticed. We will leave the aforementioned as future work to further foster the study of audio-visual attention in ODV. We expect this initial work which provides a testbed, a dataset from subjective experiments, and an analysis of user behavior could contribute the community and arouse more in-depth research in the future.

6. REFERENCES

[1] Cagri Ozcinar, Julian Cabrera, and Aljosa Smolic, "Visual attention-aware omnidirectional video streaming using optimal tiles for virtual reality," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 1, March 2019.

[2] Konrad Tollmar, Pietro Lungaro, Alfredo Fanghella Valero, and Ashutosh Mittal, "Beyond foveal rendering: smart eye-tracking enabled networking (SEEN)," in *ACM SIGGRAPH 2017 Talks*. 2017.

[3] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng, "Scene-aware audio for 360 videos," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, 2018.

[4] Colm O Fearghail, Cagri Ozcinar, Sebastian Knorr, and Aljosa Smolic, "Director's cut - Analysis of aspects of interactive storytelling for VR films," in *International Conference for Interactive Digital Storytelling (ICIDS) 2018*, 2018.

[5] Erik Van der Burg, Christian NL Olivers, Adelbert W Bronkhorst, and Jan Theeuwes, "Audiovisual events capture attention: Evidence from temporal order judgments," *Journal of Vision*, vol. 8, no. 5, 2008.

[6] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–1, 2020.

[7] Erwan J. David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet, "A dataset of head and eye movements for 360° videos," in *Proceedings of the 9th ACM Multimedia Systems Conference*, 2018, MMSys '18.

[8] Ziheng Zhang, Yanyu Xu, Jingyi Yu, and Shenghua Gao, "Saliency detection in 360 videos," in *Proceedings of the European Conference on Computer Vision (ECCV '18)*, 2018.

[9] Cagri Ozcinar and Aljosa Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018.

[10] Afshin Taghavi Nasrabadi, Aliehsan Samiei, Anahita Mahzari, Ryan P. McMahan, Ravi Prakash, Mylène C. Q. Farias, and Marcelo M. Carvalho, "A taxonomy and dataset for 360° videos," in *Proceedings of the 10th ACM Multimedia Systems Conference*, 2019, MMSys '19.

[11] Aakanksha Rana, Cagri Ozcinar, and Aljosa Smolic, "Towards generating ambisonics using audio-visual cue for virtual reality," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019.

[12] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang, "Self-supervised generation of spatial audio for 360 video," in *Advances in Neural Information Processing Systems*, 2018.

[13] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon, "Learning to localize sound sources in visual scenes: Analysis and applications," *arXiv preprint arXiv:1911.09649*, 2019.

[14] Hamed R. Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala, "DAVE: A deep audio-visual embedding for dynamic saliency prediction," *CoRR*, vol. abs/1905.10693, 2019.

[15] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Transactions on Image Processing*, vol. 29, pp. 3805–3819, 2020.

[16] "JavaScript 3D library. https://threejs.org/," Jan 2020.

[17] "WebXR device api specification," https://github.com/immersive-web/webxr, Jan 2020.

[18] "Jsambisonics," https://github.com/polarch/JSAmbisonics, Jan 2020.

[19] ITU-T, "Subjective video quality assessment methods for multimedia applications," ITU-T Recommendation P.910, Apr 2008.

[20] Francesca De Simone, Jesús Gutiérrez, and Patrick Le Callet, "Complexity measurement and characterization of 360-degree content," in *Electronic Imaging, Human Vision and Electronic Imaging*, 2019.

[21] Otto-Joachim Grüsser and Ursula Grüsser-Cornehls, "The sense of sight," in *Human Physiology*, Robert F. Schmidt and Gerhard Thews, Eds. Springer Berlin Heidelberg, Berlin, Heidelberg, 1983.

[22] Xavier Corbillon, Francesca De Simone, and Gwendal Simon, "360 degreee video head movement dataset," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, 2017, MMSys'17.

[23] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic, "Look around you: Saliency maps for omnidirectional images in VR applications," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2017.

[24] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996, AAAI Press.

[25] 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*, 2018.

[26] Anh Nguyen and Zhisheng Yan, "A saliency dataset for 360-degree videos," in *Proceedings of the 10th ACM Multimedia Systems Conference*, New York, NY, USA, 2019, MMSys '19, Association for Computing Machinery.

[27] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, 2005.

[28] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Fredo Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, Mar. 2019.

[29] Olivier Le Meur, Thierry Baccino, and Aline Roumy, "Prediction of the inter-observer visual congruency (IOVC) and application to image ranking," in *Proceedings of the 19th ACM International Conference on Multimedia*, New York, NY, USA, 2011, MM '11, Association for Computing Machinery.

[30] Alexandre Bruckert, Yat Hong Lam, Marc Christie, and Olivier Le Meur, "Deep learning for inter-observer congruency prediction," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 3766–3770.