An Analytical Study of Visual Attention Behavior in Viewing Panoramic Video

Feilin Han*†
Department of Film and TV Technology
Beijing Film Academy
Beijing, China
hanfeilin@bfa.edu.cn

Ying Zhong*
Department of Film and TV Technology
Beijing Film Academy
Beijing, China
zhongyingkc@icloud.com

Ke-Ao Zhao
Department of Film and TV Technology
Beijing Film Academy
Beijing, China
zhaokeaoedu@outlook.com

Figure 1: Representative frames from panoramic video, produced to discuss how user explore the 360 motion pictures. The visualization of user attention saliency intuitively shows users' visual perception behavior and narrative cognition process. We collect a dataset to summarize attention distribution regulations and derive practical insights for panorama production.

ABSTRACT

Panoramic video offers an immersive viewing experience in which viewers can actively explore 360-degree motion pictures and engage with the narrative. Studying user visual attention behavior could help us to have a better understanding of video processing, semantic learning, and coding in 360-degree videos. In this paper, we developed two attention visualization toolkits, visual saliency map and semantic attention annotation, for collecting ROI data. The practice-based analytical methodology is employed to discuss user behavior while viewing panoramic shorts. We gathered viewing behavior data from 23 participants and visualized attention saliency to analyze the viewers' visual attention behavior and narrative cognition process. According to the collected data, we summarize attention distribution regulations and derive practical insights into the aspects of learning decision-making for panorama production.

CCS CONCEPTS

• **Applied computing** → **Media arts**; • **Human-centered computing** → *Visualization design and evaluation methods*; *Virtual reality*.

*Both authors contributed equally to this research.
†Corresponding authors

HCMA '23, November 2, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0272-3/23/11…$15.00
https://doi.org/10.1145/3606041.3618060

KEYWORDS

panoramic video, visual attention, immersive viewing behavior, attention visualization, virtual reality

ACM Reference Format:
Feilin Han, Ying Zhong, and Ke-Ao Zhao. 2023. An Analytical Study of Visual Attention Behavior in Viewing Panoramic Video. In *Proceedings of the 4th International Workshop on Human-centric Multimedia Analysis (HCMA '23), November 2, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3606041.3618060

1 INTRODUCTION

Panorama has been developed for decades, known as a powerful media with ultimate potential of immersive 360-degree information transmission. Many well-known film festivals, such as Venice, Cannes, Sundance, and the Beijing International Film Festival have established exhibitions for VR films [6]. With the spring up of award-winning VR films, panorama production draws researchers' attention that it is necessary to do further study on 360-degree video for supporting immersive content authoring. Thanks to the distinguishing characteristics of panorama, it makes it possible for users to explore the virtual environment freely, meanwhile which makes their visual attention and region of interest (ROI) out of control.

2D videos present visual messages within limited frames and fields of vision (FOV), where attention distributes in planar. However, in panoramic videos, visual attention is in a higher spatial complexity. An in-depth understanding of the relationship between visual attention and inter-frame spatiotemporal information could help us propose more effective coding and semantic understanding algorithms. To intuitively showcase the attention transition, we developed two attention visualization toolkits, visual saliency map and semantic attention annotation.

In this paper, we employ the practice-based analytical methodology to discuss user behavior while viewing panoramic shorts.

We gathered viewing behavior data from 23 participants and visualize attention saliency to analyze the viewers' visual perception behavior and narrative cognition process. To this end, we conduct a user study to comprehend user viewing preference and visual attention mode. Experimental datasets are collected by utilizing the saliency map to present cognitive processes. We also design an interactive comparative experiment to classify user active and passive attention behavior which infers subconscious perception-taking in viewing. According to the collected data, we summarize attention distribution regulations and derive practical insights into the aspects of learning decision-making for panorama production. Specifically, our contributions are:

- We developed two attention visualization toolkits, visual saliency map and semantic attention annotation, for collecting user ROI data to measure the narrative engagement.
- We produced a panoramic video, shot in a professional filmmaking process, for exploring the attention distribution regulations. We carry out a qualitative analysis to discuss the visual attention behavior and narrative cognition process in immersive viewing experience.
- In this paper, we derive practical insights on video processing, semantic learning, and coding in 360-degree videos.

2 RELATED WORK

Panoramic video offers 360-degree information that viewers can explore the surroundings in an immersive way while wearing the head-mount displays (HMDs). In general, the panoramic video needs to be projected into a rectangular plane first so that the normal video coding standards (such as AVC, HEVC, and AVS) could be used to compress the panorama. However, they drop off the depth and geometry information while compression, missing the panoramic features. IEEE 1857.9 Immersive Visual Content Coding Standard [3] and ISO/IEC MPEG Immersive Video (MIV) Standard [2] were designed for panoramic video compression. Researchers proposed a hybrid approach by embedding depth map transmission in the depth coding process [7]. To achieve high-efficiency video coding, cheon et.al [4] finds that high-solution and wide field of view (FOV) provide better visual experience [14], but attract more attention than fewer and low-resolution attentive regions. Further investigation of perceptual experience in immersive visual environments will be necessary for panoramic video coding. Inspired by these works, we hold that visual attention should be involved in deep video coding as well as panorama quality assessment in the future.

To annotate and capture visual attention data, we need specific tools for immersive viewing behavior analysis. Most of the existing immersive creation tools can be divided into two categories: professional-oriented (e.g. Unity, Unreal Engine) and amateur-oriented (e.g. VRChat, Tilt Brush) [10]. Academic researchers have developed a VR authoring system to represent rectilinear views for modifying color, contrast, and luminance information [12]. Inspired by this work, we use Unreal Engine and develop plug-in toolkits for attention distribution visualization. Our tool supports the panorama production pipeline and works in real-time and WYSIWYG (what you see is what you get) way.

In professional panorama production, visual-audio language principles are well-utilized for presenting a vivid story, such as mise-en-scène and montage [15]. A traditional authoring pipeline includes pre-preparation, shooting, and post-production. Due to the temporal-spatial complexity in panorama, the understanding of video content relies on the result of the multi-perception of visual, sensation, and engagement. Orduna et al. [11] proposed a complex methodology to assess the presence, empathy, attitude, and attention. Bindman et al. [1] give out a clear understanding of the perception of role with narrative engagement. To improve the amount of narrative information, it is crucial to adapt the visual principles to panorama production [17]. Han et al. [8] discuss the effect of mise-en-scène and Masia et al. [9] analyze the influence of attention guidence. Most of these works focus on guidance, with few discussions on exploring the relationship between visual attention and inter-frame spatiotemporal information.

3 ATTENTION VISUALIZATION TOOLKIT

In this work, we developed two attention visualization toolkits, visual saliency map and semantic attention annotation, for collecting user data to measure narrative engagement. The identification of visual attention regions has various modeling methods by embedding psychophysics, computational methods, and neurophysiology. Our attention visualization toolkit could be used as an alternative solution for ground-truth annotation.

3.1 Visual Saliency Map

To intuitively showcase the attention transition, we developed a visualization toolkit to generate visual saliency map. This toolkit is a plug-in in Unreal Engine, which could display and record HMDs orientation, field of view (FOV), time stamps as well as interactive action data. The HMDs orientation was recorded every 15 frames. We projected these data to the 2D position of UV coordinate and generate saliency map.

The visual attention saliency is calculated by the IoU, which is related to the overlap of viewpoints and narrative ROIs. There are two kinds of narrative ROIs in panorama, directorial-designed region, and semantic-attractive region. The directorial-designed regions are manually annotated by the creator, and the semantic-attractive regions could be generated by the semantic attention annotation toolkit. The bounding boxes of narrative ROIs are used to calculate IoU, shown as Figure 2.

Figure 2: The example of narrative ROI and attention saliency.

3.2 Semantic Attention Annotation

Considering semantic information is highly associated with scene and actor. Inspired by YOLO [13], we embed computer-vision-based classification and recognition methods for semantic region detection. Taking the use of intelligent and automatic methods, the sematic label and localization results could be recorded into metadata, in order to improve the efficiency of panorama understanding. The framework of semantic attention annotation is shown as Figure 3.

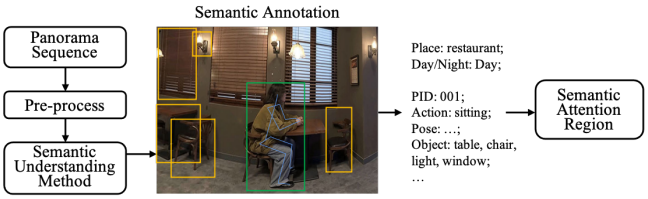

Figure 3: The framework of semantic attention annotation.

To detect semantic-attractive regions, we develop a metadata annotation system, which is used for semantic understanding and video content analysis. The interface of which is shown as Figure 4. This system could localize the narrative-related information, such as actor and performance as well as scene and object. To detect actor and performance, we employ the pre-trained RetinaFace [5] models for face detection and 3d pose estimation method for performance capture. To recognize object and scene, we use the pre-trained places365 scene recognition model [18] for place category classification and the YOLOv5 model [13] for object localization. After interviewing professional panorama creators, we list several essential semantic annotations. It should be mentioned that the labels introduced in this system could be extended and amended in terms of user demands.

4 PANORAMIC DATASET

To explore the user attention behavior in viewing panorama, we produced a short video, which is shot in a professional filmmaking process. The storyline could be summarized as

> *"A man comes to the bar, waiting for his girlfriend. The waiter hands him a glass of water. After knowing she won't come, he leaves, disappointed and heartbreaking."*

This panorama is 8K (7680×3840), omnidirectional, and 3 DOF (Degree of Freedom), which was shot by insta 360 Titan, stitched in Insta 360 Stitcher, and edited in Adobe Premiere. Examples of shotting device, filming on-set, and screenshot are shown as Figure 5. To avoid the dual influence of sound, which is well-known effective guidance in VR, no soundtrack was recorded and produced. This panoramic video is designed for collecting User Attention Data (UAD) to explore the attention distribution regulations. This dataset has enrolled 23 participants' viewing behavior. While watching the panoramic video, they wear head-mounted displays (HMD) in VR seated condition. We carry out a qualitative analysis to discuss the visual attention behavior and narrative cognition process in immersive viewing experience.

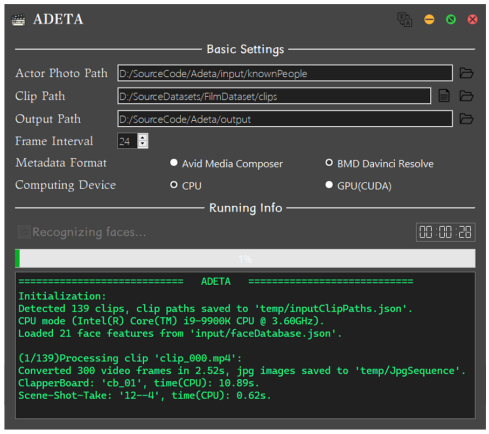

Figure 4: Diagram of our system user interface.

5 VISUAL ATTENTION BEHAVIOR ANALYSIS

In human visual system, the attention mechanism is selective and perceptional. The observable and significant objects are appealing to human vision interests, called salient region, which corresponds to attention. Utilizing data collected above, we analyze their attention transition and attention behavior. Visual attention behavior analysis is considered as an effective approach for salient region detection, which is useful for object-based image processing such as semantic understanding and learning-based video coding.

5.1 Attention Transition

To analyze the narrative engagement of the audience, we generate attention saliency maps 23 users. The saliency map is used as the estimations of participants' fixations since there is a strong correlation between user attention and head movement [16]. From the saliency map, we could see that most viewers are concentrating on the right ROI when the actors are performing or the story moves to the highlights. However, there are some obvious principles that could distract the audience.

To summarize these principles, we select 6 representative keyframes: $f003$, $f043$, $f061$, $f088$, $f105$, $f183$, shown as Figure 6. At the beginning and the end, users are more willing to explore the VR environment without any fixed ROI, such as in $f003$ and $f183$. When the storytelling is in a slow narrative rhythm or the story is not attractive enough, users are easy to lose their attention as in

Figure 5: Examples of the filming device, on-set, and participant viewing condition of panorama production.

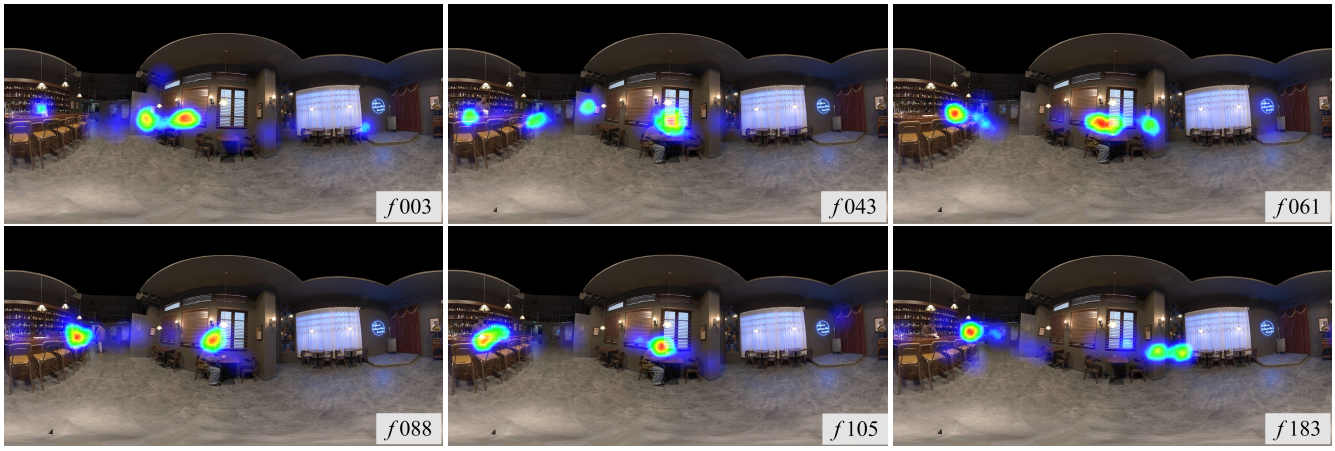

Figure 6: Example frames showing participants are missing narrative engagement and exploring the virtual environment.

$f043$ and $f088$. When the story suddenly changed or the performance started in another region without any guidance, users will be delayed to change their attention ROI, as in $f061$ and $f105$.

5.2 Attention Behavior

To infer the principle of viewer perception behavior in an immersive environment, we collect viewing behavior data from User Attention Dataset. All these data are utilized to demonstrate subconscious perception-taking in viewing. In order to further explore the mechanism of viewing attention, we recorded HMDs rotation data during participants engaging in the narrative. We compare it with the annotated narrative POI (point of interest) and viewing attention ground-truth, shown in Figure 7. In this figure, we utilize the Air-E Score as a measurement metric for quantitative analysis.

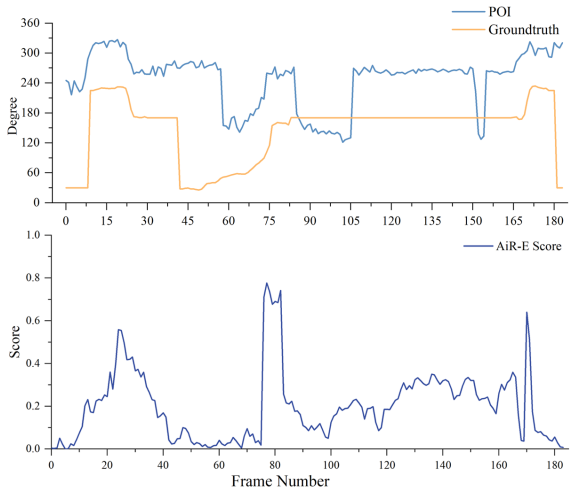

Figure 7: The attention POI and Air-E scores from User Attention Dataset.

We find that users spend more time in the first section and the last section, due to their adaption to controllers and exploration in panorama. It takes some time to focus and concentrate, which is a passive perception behavior, the same as the process of people getting used to a new place. So, at the beginning of a new story, a new scene, or a new narrative section, panorama should maintain information density as much as possible for viewers to observe the entire scene in detail.

According to Figure 7, we find that viewers change their perceptions along with the video sequence. When observable and significant objects appear, users change their ROI to catch up with the semantic salient region. When moving and narrative-related objects come, the audiences would like to predict story plots, so they usually shift their attention to the direction of the actor's movement in advance. We analyzed the visual saliency map sequence along with Figure 7, and summarize attention transition regulations as listed in Table 1.

Table 1: Visual attention transition regulations.

No.	Description
1	At the beginning of the video, attention explores the whole panorama.
2	New object appears, attention gradually focuses in latency.
3	When the scene shows no change, attention gradually turns to disperse and divergent.
4	The attention is predictable when objects are moving and disciplinary.
5	The attention is faster than the objects' movement.
6	Attention behaviors vary with individuals on the basis of their visual interests and personal preference.
7	Attention is a comprehensive result of active and passive perception. When the information density is higher, active attention weights greater, while the information density is lower, passive attention has a more significant proportion.

6 CONCLUSION AND INSIGHTS

Panoramic video is considered a great potential medium for the GLAM (Galleries, Libraries, Archives, and Museums) and Art industry. From the received questionnaires, the most mentioned keywords are narrative experience, immersion, more vivid, and active exploration. One of our participants wrote,

> *"The panorama shows a strong sense of presence and immersion but makes me easier to miss information."*

The 360-degree video shows more information, a strong sense of engagement and presence, and a multi-perception of immersion, interaction, multi-sensation, and self-determination. However, it is also easier to obscure and disregard important ROIs due to the spatial redundant information and temporal-spatial complexity.

How to annotate viewers' real attention is a challenging problem in immersive video processing as well as other visual content understanding. In our work, we use attention visualization toolkits to annotate users' regions of interest (ROI). The collected attention data is determined by the viewer's actual attention rather than the third-party annotators. Our work would like to offer a discussion on attention behavior, not only for attention learning but also for panorama authoring. Analyzing viewing behavior is also particularly useful for addressing the problems listed below.

- What kind of visual regions drive viewer attention most?
- Does all semantic regions have strong relations with attention?
- Where should be coded in minimized low compression to maintain the contextual information and comfortable viewing experience?
- How do we design visual attention guidelines for directing effective VR storytelling?

In this study, we aim to figure out the relationship between attention distribution and viewing behavior, which could help us to have a better understanding of video processing, semantic learning, and coding in 360-degree videos. In the future, we would like to employ proposed attention visualization and annotation toolkits to build up a panoramic attention ground-truth dataset for panoramic attention learning and video processing. We are willing to explore how visual-audio language principles, such as cinematography, cutting, and editing, influence visual attention. The proposed toolkits and guidelines could be applied to assist VR film directors in directorial design creation and production. We hope this work could also contribute to the VR film authoring and production.

7 ACKNOWLEDGMENTS

This work was supported by the National Social Science Fundation Art Project (No. 20BC040) and the Scientific Research Project of Beijing Educational Committee (No. KM202110050001). We thank the ACs, reviewers, and the participants of our dataset. We also thank our actors, Ji-An Yu and Dongyue Li, for their great performance in our produced panoramic short video.

REFERENCES

[1] Samantha W Bindman, Lisa M Castaneda, Mike Scanlon, and Anna Cechony. 2018. Am I a bunny? The impact of high and low immersion platforms and viewers' perceptions of role on presence, narrative engagement, and empathy during an animated 360 video. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–11.

[2] Jill M. Boyce, Renaud Doré, Adrian Dziembowski, Julien Fleureau, Joel Jung, Bart Kroon, Basel Salahieh, Vinod Kumar Malamal Vadakital, and Lu Yu. 2021. MPEG Immersive Video Coding Standard. *Proc. IEEE* 109, 9 (2021), 1521–1536. https://doi.org/10.1109/JPROC.2021.3062590

[3] Yangang Cai, Xufeng Li, Yueming Wang, and Ronggang Wang. 2022. An Overview of Panoramic Video Projection Schemes in the IEEE 1857.9 Standard for Immersive Visual Content Coding. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 9 (2022), 6400–6413. https://doi.org/10.1109/TCSVT.2022.3165878

[4] Manri Cheon and Jong-Seok Lee. 2018. Subjective and Objective Quality Assessment of Compressed 4K UHD Videos for Immersive Experience. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 7 (2018), 1467–1480. https://doi.org/10.1109/TCSVT.2017.2683504

[5] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5203–5212.

[6] Kath Dooley. 2021. *Cinematic virtual reality: A critical study of 21st century approaches and practices*. Springer.

[7] Patrick Garus, Félix Henry, Joel Jung, Thomas Maugey, and Christine Guillemot. 2022. Immersive Video Coding: Should Geometry Information Be Transmitted as Depth Maps? *IEEE Transactions on Circuits and Systems for Video Technology* 32, 5 (2022), 3250–3264. https://doi.org/10.1109/TCSVT.2021.3100006

[8] Feilin Han, Ying Zhong, and Minxi Zhou. 2022. Evaluating the Effect of Cinematography on the Viewing Experience in Immersive Environment. In *2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[9] Belen Masia, Javier Camon, Diego Gutierrez, and Ana Serrano. 2021. Influence of Directional Sound Cues on Users' Exploration Across 360° Movie Cuts. *IEEE Computer Graphics and Applications* 41, 4 (2021), 64–75. https://doi.org/10.1109/MCG.2021.3064688

[10] John T Murray. 2022. RealityFlow: Open-Source Multi-User Immersive Authoring. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 65–68.

[11] Marta Orduna. 2022. Quality, Presence, Empathy, Attitude, and Attention in 360-degree Videos for Immersive Communications. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.

[12] Tania Pouli and Thanh Hang Phung. 2018. VR Color Grading using Key Views. In *Proceedings of the Virtual Reality International Conference-Laval Virtual*. 1–8.

[13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 779–788.

[14] BT Series. 2020. The present state of ultra-high definition television.

[15] Raymond Spottiswoode. 2020. A Grammar of the Film. In *A Grammar of the Film*. University of california press.

[16] Tong Xue, Abdallah El Ali, Gangyi Ding, and Pablo Cesar. 2021. Investigating the relationship between momentary emotion self-reports and head and eye movements in hmd-based 360 vr video watching. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–8.

[17] Yu Zhang. 2020. *Developing a Cinematic Language for Virtual Reality Filmmaking*. Ph. D. Dissertation. Griffith University.

[18] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. 2016. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055* (2016).