Frames to Zones: Applying Mise-en-Scène Techniques in Cinematic Virtual Reality

Adam Kvisgaard Sune Øllgaard Klem Thomas Lund Nielsen Eoin Ivan Rafferty*
Niels Christian Nilsson Emil Rosenlund Høeg Rolf Nordahl†

Aalborg University Copenhagen

ABSTRACT

Virtual reality (VR) is becoming increasingly commonplace and consumers, among other things, use the technology to access cinematic VR experiences. However, cinematic VR limits filmmakers' ability to effectively guide the audience's attention. This paper describes a between-groups study (n=60) exploring the use of a zone-division system to incorporate mise-en-scène in VR to guide attention. Participants were exposed to a cinematic VR experience including five points of interest (POIs). Half of the participants experienced a version containing visual cues arranged in the scene based on the zone-division system and the other half experienced no such guidance. The effectiveness of the intervention was assessed by measuring the angle between users' head orientation and the relevant POIs. Additionally, a questionnaire was used to determine whether the participants recalled the POIs. The results showed that participants exposed to additional visual cues orientated themselves significantly closer to the POIs than those in the other condition. However the questionnaire showed no significant differences between conditions.

Index Terms: I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual Reality;

1 INTRODUCTION

The advent of affordable virtual reality (VR) displays and 360° video cameras has sparked an interest in bringing cinematic experiences from the screen and into VR, and such experiences are becoming increasingly accessible to consumers. In traditional screen media, directors have great power in directing their audience's attention within the frame. Two of the main techniques filmmakers use to harness this power are *cinematography* and *mise-en-scène* [1]. Cinematography refers to the process of capturing a scene using a camera and includes factors such as camera placement and movement. Mise-en-scène refers to everything presented in front of the camera. By controlling elements such as the movement of actors, the placement and use of props, and lighting of a scene, filmmakers can guide the viewer's attention precisely where they want it to be [1]. The filmmaker knows that these elements are likely to be seen by the viewer because the cinematography ensures control of the audience's point of view.

However, in *cinematic virtual reality* (VR) – cinematic content displayed using technologically immersive displays – viewers generally have control over where in the scene they wish to look. Additionally, there are often constraints on field-of-view and depth-of-field caused by both the hardware and heuristic guides for minimizing simulator sickness [3]. The combination of these factors takes away much of the power granted to filmmakers through cinematography, as the user now controls camera movement, and other camera controls are limited. Without a director controlling cinematography, it is possible that a viewer could entirely miss key moments of the narrative. Thus, content creators must strike a difficult balance between allowing viewers to experience the freedom that comes with VR, while ensuring that they witness key narrative moments.

There has been a surge in popularity and accessibility of VR in recent years and cinematic content for VR is reaching new levels of mainstream recognition. Research on cinematic VR has explored different approaches to controlling playback of 360° video [11], dynamic placement of subtitles [15], and the effects of varying display types [7], viewpoint transformation [8, 9], and editing [4, 16]. Despite an increase in the amount of research on cinematic VR, there is very little in the way of a consensus on how to approach designing content, and research on how to guide viewers' attention remains relatively scarce.

It is possible to distinguish between cues for guiding viewers' attention depending on whether they qualify as *diegetic* (they are part of the story world) or *non-diegetic* (they are only perceivable to the user) [10]. While non-diegetic cues (e.g., forced rotation of the user's viewpoint, colored arrows, or vignette effects) may be effective, the optimal choice of technique depends on the content being viewed and the preferences of the viewer [6], and such techniques may be difficult to apply in a non-intrusive manner. Moreover, it has been suggested that non-diegetic techniques may limit presence because they draw attention to the mediated nature of the experience [10]. It is well-known from classical cinema that digetic manipulation of the mise-en-scéne can guide the viewer's eyes, and diegetic cues also appear provide an effective way of guiding the attention in VR [12–14, 17].

This paper introduces a heuristic framework for applying classic mise-en-scène techniques in virtual environments (VEs) to guide the viewers' attention. Moreover, the paper presents a user study exploring whether application of the framework affects participants' focus of attention during exposure to cinematic VR.

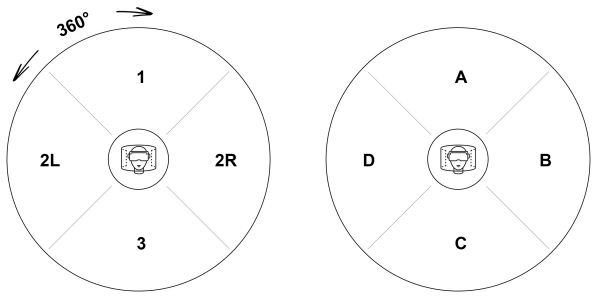

Figure 1: POI Zones (left) rotate around the user, so that the POI is always in Zone 1, while UC Zones (right) refer to the local space of the scene around the user.

*e-mails: {akvisg15, sklem15, tlni15, eraffe15}@student.aau.dk
†e-mails: {ncn, erh, rn}@create.aau.dk

2 Four Zones of Mise-en-Scène

The term mise-en-scène, which literally translated from French means "putting on stage", refers to all the elements in front of the camera that the filmmaker can control. Bordwell et al. [1] distinguish between four elements of mise-en-scène: (1) *setting*, which includes environments, objects, and props, (2) *costumes and makeup*, (3) *lighting*, and (4) *staging* which includes the placement, movement, and performance of characters and objects. By manipulating these elements, filmmakers can guide the viewer's attention to the desired part of the frame. For example, lighting and, by extension, shadows can be used to highlight or obfuscate; careful staging of characters and objects may frame or conceal an important element of the mise-en-scène; and a character's gaze can be used to draw the gaze of the viewer. As such, mise-en-scène techniques are powerful tools for guiding the viewer's eyes toward important, and away from non-essential, parts of the frame. However, rather than guiding viewers' eyes on a two-dimensional plane, creators of cinematic VR need to shepherd the viewer's gaze to *points of interest* (POIs) in three-dimensional worlds where some POIs are likely to be outside the viewer's field of view.

Our framework is inspired by the work of a Los Angeles-based company, *Visionary VR*, who try to ensure that viewers do not miss important content by dividing the scene into four zones [5]. In this system, the primary zone would show the main content of the narrative. The borders between zones are visually represented using non-diegetic lines, and slow-motion effects and pausing of the main content are used to ensure that the viewer does not miss narrative events. However, because such non-diegetic manipulations are likely to diminish the plausibility illusion and in turn presence [18], our approach focuses exclusively on diegetic manipulation of the mise-en-scène. Specifically, we propose that the process of deploying traditional mise-en-scène techniques in VR can be eased by dividing the virtual space into zones.

First, we divide the virtual space surrounding the viewer into four discrete *POI Zones* [Figure 1(left)]. The four POI Zones, which are defined independently of the user's current viewing direction, are: the primary zone where the POI is located (*Zone 1*), two secondary zones to the left and right of the user (*Zone 2R* and *2L*), and the tertiary zone behind the user (Zone 3). In other words, the zones rotate around the user, so that the POI is always in Zone 1. When creating the mise-en-scène, the developer can treat each zone as a traditional "frame" and seek to draw the user's gaze away from the secondary and tertiary zones toward the POI in Zone 1. For example, light and shadow can be used to emphasize the POI in Zone 1, characters and objects can block parts of Zone 2R and Zone 2L, and characters located in all zones can direct attention toward the POI using their gaze and other actions.

In addition, we distinguish between four *User-Centered (UC) Zones* that divide the local space around the user: Zone A, B, C, and D representing the areas in front, to the left, to the right, and behind the user, respectively [Figure 1(right)]. As such, the aim is to ensure that POI Zone 1 and UC Zone A are aligned. Combining these two zone systems allows the developer to clearly explain the scene in relation to the both the main POI, and in relation to the user's position in the environment.

3 User Study

The study relied on a between-groups design comparing two conditions: A cinematic VR experience where the viewers' attention was guided using visual cues, set up in the scene based on the developed zonal framework (*Condition A*) and the same cinematic VR experience devoid of visual guidance (*Condition B*).

3.1 Participants

A total of 60 participants took part in the study. Participants were recruited from the student body at Aalborg University Copenhagen and randomly assigned to one of the two conditions (n = 30). All participants gave written informed consent prior to participation and were unfamiliar with the purpose of the study. When asked if they had prior experience with VR 15 and 17 reported having prior experience with VR for condition A and B, respectively. Only eight participants reported minor symptoms of simulator sickness post exposure, as assessed using a single questionnaire items asking whether or not the participants experienced any symptoms of simulator sickness.

3.2 Narrative and Visual Cues

All participants were exposed to the same cinematic VR experience with the only difference being the presence or absence of visual cues. The narrative forming the basis for the scenario experienced by the participants was inspired by the film *Apocalypse Now* (Francis Ford Coppola, 1979) and can be summarized as follows:

The main character (the user) is on a research expedition together with his colleague John and a translator. John, who had been unhinged, went missing recently. Accompanied by the translator, a local guide, and a dog, the main character travels into the wilderness to search for John. They travel down the shallow river on a small raft and eventually finds John's boat. They learn that their colleague has become indoctrinated into a Lovecraftian cult. In a violent struggle they kill their colleague. They return to civilization, plagued by the same nightmares which haunted their colleague.

The cinematic VR experience revolves around the journey down the river. The translator and the guide converse throughout the journey. The primary purpose of this dialogue was to provide the user with necessary background information (e.g., the translator suggests that John's disappearance may be connected to passages in his journal describing strange dreams he has had recently). The dialogue was deliberately written so that it did not reveal or point towards any specific objects in the VE, and no dialogue was presented when the

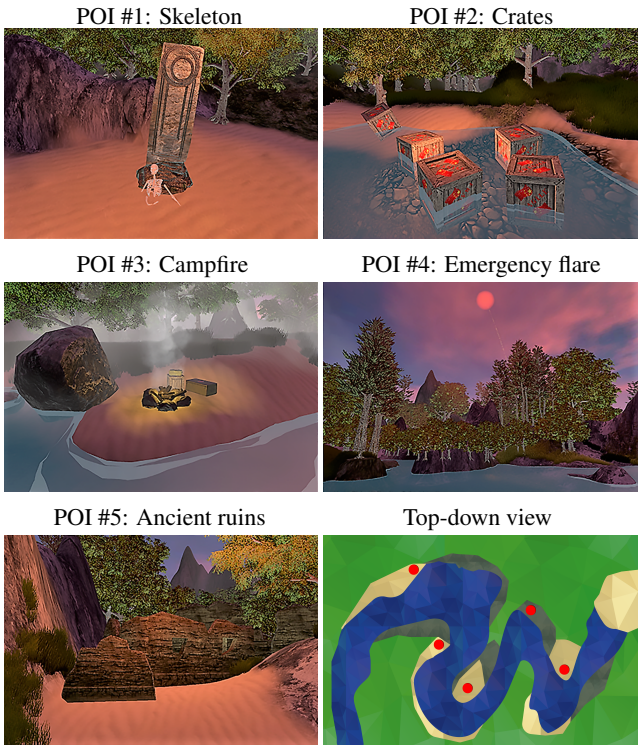

Figure 2: The five POIs and a schematic drawing illustrating a top-down view of the VE with the five POIs highlighted with red.

Table 1: Summary of how elements of mise-en-scéne were used to guide attention across the five POIs. The elements were used to highlight POIs (h), direct attention towards POIs (d), and to mask unimportant areas of the VE (m).

POI	Zone 1	Zone 2R	Zone 2L	Zone 3
1	Intensified sunlight on POI #1 (h) and the interpreter blocks irrelevant parts of Zone 1 (m).	The dog is looking into Zone 1 (d), and the glare from the sun makes it difficult to see when facing toward this zone (m).	Light from Zone 1 is visible (d), shadow cast from the mountains limit visibility in this zone (m) and the guide blocks parts of this zone (m).	Shadow cast from the mountains limit visibility in this zone (m).
2	Intensified sunlight on POI #2 (h) and vegitation and the interpreter blocks irrelevant parts of Zone 1 (m)	The guide looks in the direction of zone 1 (d) and blocks parts of this zone (m), and the glare from the sun makes it difficult to see when facing toward this zone (m).	The glare from the sun makes it difficult to see when facing toward this zone (m)	Shadow cast from the mountains limit visibility in this zone (m).
3	Intensified sunlight on POI #3 (h) , the interpreter walks through the zone right before POI #3 appears (d), and the dog barks in the direction of the POI (d).	Birds are flying toward Zone 1 (d) and the glare from the sun makes it difficult to see when facing toward this zone (m).	Birds are flying toward Zone 1 (d) and the interpreter and guide blocks parts of this zone (m).	Shadows cast by vegetation limits visibility (m) and the interpreter and guide blocks parts of this zone (m).
4	Intensified sunlight on POI #4 (h).	The guide is looking into Zone 1 (d) and blocks parts of this zone (m). The glare from the sun and the mist makes it difficult to see when facing toward this zone (m).	The local blocks parts of this zone (m) and the mist limits (m).	Shadow cast from the mountains and the mist limit visibility in this zone (m).
5	Intensified sunlight on POI #5 (h) and the mist and vegetation blocks irrelevant parts of the zone (m).	The dog walks into Zone 1 while barking (d), leaves are soaring into Zone 1 (d), and the glare from the sun makes it difficult to see when facing toward this zone (m) which also is blocked by the guide (m).	Leaves are soaring into Zone 1 (d), the local blocks parts of this zone (m), and visibility is low due to the mist (m).	Leaves are soaring into Zone 1 (d), and shadow cast from the mountains makes and the mist limit visibility in this zone (m).

VE included items relevant to the plot. Specifically, five POIs were gradually presented along the river banks. The POIs provided clues regarding what had happened to John and the presence of the cult. The five POIs were: (1) A *skeleton* lying against a standing stone with occult engravings. (2) Some washed up *crates*, possibly from John's boat. (3) A recently extinguished *campfire*. (4) An *emergency flare* being shot in the distance. (5) An *ancient ruin* at the shore where John's boat is found (see Figure 2).

The VE used for the two conditions was identical with exception of the visual cues used to guide viewer's attention in condition A. That is, during encounters with the five POIs in condition A, the mise-en-scène was manipulated based on the zonal framework (Section 2). Specifically, in Zone 1 POIs were emphasized using additional sun light; characters' gaze and movement were used to direct attention toward Zone 1; characters and objects were used to block irrelevant parts of the VE across all zones; and shadows, mist, and glares from sunlight were used to limit visibility outside of Zone 1. Table 1 presents a summary of how elements of mise-en-scéne were used to guide attention across the five POIs.

Condition B did not include highlights of POIs, strong shadows, mist, and glares, and character movement was randomized to avoid blocking, gazing, and movements guiding the user's attention. Figure 3 shows a part of Zone 1 as it appeared to participants when they encountered POI #1 during condition A and B.

3.3 Setup and Procedure

The VE used for the study was developed in Unity 3D and presented using a HTC Vive head-mounted display and a pair of circumaural headphones. The participants were seated on a non-swiveling chair throughout the experience. After a short introduction, the participants were exposed to a 6 minute cinematic VR experience, and they subsequently filled out a short questionnaire pertaining to their experience. The study lasted for approximately 10 minutes, including the introduction and post-test questionnaire.

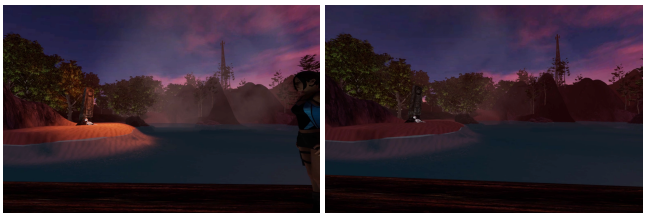

Figure 3: POI #1 as it appeared during condition A (top) and B (bottom).

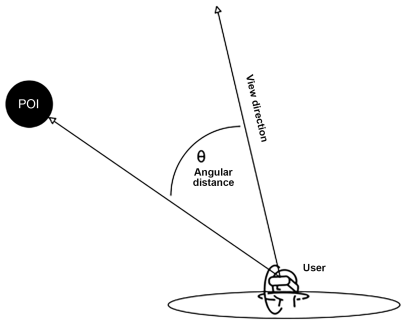

Figure 4: Visualization of how the angular distance was derived from the viewing direction, the position of the user's viewpoint and the POI.

3.4 Measures

To determine whether application of the zonal framework guided participants' attention, we measured the *angular distance* between a participant's viewing direction and the vector formed by the virtual viewpoint and the POIs (Figure 4). The angular distance was measured when a POI was visible, resulting in 39 data points per POI per participant.

Additionally, each participant was asked to fill out a questionnaire after completing the VR experience. This questionnaire consisted of six sections. Each section would first list a number of objects and require the participant to mark which of those objects, if any, they saw in the VE. They were also asked if they could elaborate on why they noticed those objects. Each section mentioned one of the POIs and multiple objects which were not present in the scene. The sixth section only mentioned objects which were not present in the VE. The results from the questionnaire were primarily meant to triangulate data and give an impression of whether the intervention affected how well users recalled seeing the POIs.

4 Results

The aim of the study was to determine if application of the framework had an overall effect on the participants' attention during exposure to cinematic VR. Thus, the analysis of the data related to angular distance focused on the *mean angular distance* (MAD) across all five POIs for each participant. Figure 5 visualizes the results pertaining to MAD in terms of means and 95% confidence intervals (CIs), as well as the corresponding results for each POI. The latter were not subjected to statistical analysis and are solely included to give an impression of the angular distances that contributed to the MAD (the rightmost bars in Figure 5). The assumption of homogeneity of variances was violated, as assessed by Levene's test for equality of variances ($p = .012$); the data contained no outliers, as assessed by inspection of boxplots; and MAD for both groups were normally distributed, as assessed by Shapiro-Wilk's test ($p > .05$). Thus, a Welch t-test was performed to determine if there were differences in MAD between condition A and B, The results showed that participants' orientation towards the POIs in condition A ($M = 48.6, SD = 12.3$) was significantly closer than that of participants exposed to condition B ($M = 59.0.6, SD = 21.6$), with a medium effect size ($t = -2.29, p = 0.013, r = 0.32$).

The data obtained from the questionnaire was considered binary nominal data, where each POI was either "seen" or "not seen" by the participants. Figure 6 shows the number of participants who recalled seeing each of the five POIs. A χ^2-test to compare two proportions showed that there was no significant difference between the reported sightings of the POIs in the two conditions ($p = 0.289$).

5 Discussion

The significant result from the test comparing MAD across condition A and B indicated that application of the framework was successful in causing users to orient themselves closer towards POIs in the created scene. This is in line with previous work suggesting that implicit diegetic cues can be used to guide viewers' attention during exposure to cinematic VR. Particularly, previous work has indicated that the movements of a virtual agent (a firefly) can be used to affect the viewer's gaze [10], and the combination of sound and movement within the VE may also influence what areas of the scene the viewer attends to, whereas static lighting was not found to have and effect [13]. The later finding may seem to contradict the results of the current study where static light (i.e., sunlight) was used to highlight POIs. However, it is worth stressing that the application of the zonal framework involved the concurrent manipulation of several cues guiding the viewer's attention toward POIs and away from other objects in the scene. Thus, it is not possibly to conclude wither the static lightning would be sufficient to guide the viewers' attention if applied in isolation.

The lack of significant results from the questionnaire suggests that condition A did not have any notable impact on whether or not the participants recalled seeing the POIs after exposure to the VE. It is possible to offer at least two potential explanations for this difference between the results of the two measures. First, we cannot rule out the possibility that the cues may not have affected whether the participants saw the intended POI. In other words, the mise-en-scène in the secondary and tertiary zones may have helped guide the viewer's attention toward the primary zone, but the cues in the primary zone did not affect whether the participants noticed the POI or not. This might indicate that the cues deployed in the primary zone were too subtle. However, it seems equally likely

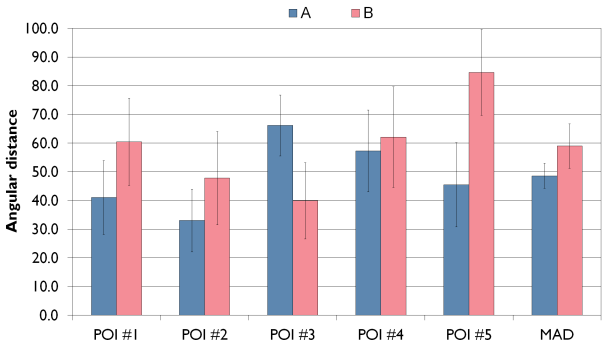

Figure 5: Angular distances across all POIs and MAD for condition A and B. Error bars represent 95% CIs

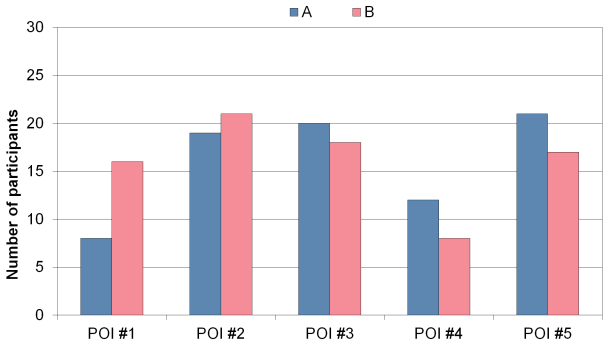

Figure 6: Number of participants per condition (n=30) who recalled seeing each of the five POIs.

that the cues in the primary zone had a limited effect because the POIs themselves were quite salient. Second, it is possible that the questionnaire design may have affected the results to some degree. Specifically, the item designed to determine whether the participants had seen POI #1 included the option "some skeletons". As only one skelton was present in the scene, participants may have decided not to choose this option. Finally, it is worth highlighting that the viewers' ability to recall specific POIs need not affect the degree to which they enjoyed the experience. That is, even if viewers do not recall all POIs, they may have found the narrative coherent, as the sensation of narrative closure does not always necessitate narrative intelligibility [2]. The discrepancy between the results related to MAD and recollection of POIs suggests the need for future work exploring how implicit guidance in cinematic VR affects not only visual attention and recall but also the experience of enjoyment, narrative coherence, and intelligibility.

It is interesting to note that the angular distances associated with POI #3 was higher for condition A. This may be viewed as an indication that the specific manipulation of the mise-en-scéne in that instance distracted the viewer from POI #3. During that moment the dog barks while standing on the border of Zone 1 and Zone 2R looking at POI #3, and the translator walks through Zone 1 and into Zone 2R (see Figure 7). Both of these elements of mise-en-scéne could have caused the viewers to shift their gaze rightward and away from POI #3; thus producing a higher angular distances. However, the events occurred while POI #3 had been in view for some time and approximately two thirds of the participants recalled seeing POI #3 in both conditions (20 and 18 in condition A and B, respectively). Thus, it is unlikely that the (potential) distraction had a notable effect on the participants experience. Nevertheless, it highlights the potential risk of deploying misguiding cues and the need for prototyping cinematic experiences during early stages of production. Moreover, it suggests that angular distances measured for the entire time a POI is visible only provide a partial picture of shifts in viewers' attention. Alternative measures relying on eye tracking would help mitigate such problems.

6 Conclusion

This paper detailed a user study exploring the application of a zone-division system to incorporate mise-en-scène in VR to guide user's attention. The results showed that participants exposed to visual guidance orientated themselves significantly closer to the points of interest than those in the condition devoid of guidance. However, no significant differences were found in relation to the participants' recollection of the points of interest. Further studies are needed to determine exactly what elements of mise-en-scéne are the most effective and the least intrusive in the context of varying VEs, different narratives, and when POIs are less salient. Nevertheless, we regard the results as an indication that mise-en-scéne techniques, deployed based on the framework, may affect viewer's attention. Moreover, we believe the proposed framework to serve as a useful tool for filmmakers aspiring to use elements of mise-en-scéne to guide viewers' attention during exposure to cinematic VR.

Figure 7: The border between Zone 1 and 2R as it appeared to the participants during condition A (top) and B (bottom).

References

[1] D. Bordwell, K. Thompson, and J. Smith. *Film art : an introduction*. McGraw-Hill higher education. Mcgraw Hill, Boston, 8. ed. ed., 2008.

[2] L. E. Bruni and S. Baceviciute. Narrative intelligibility and closure in interactive systems. In *International Conference on Interactive Digital Storytelling*, pp. 13–24. Springer, 2013.

[3] J. Jerald. *The VR book : human-centered design for virtual reality*. San Rafael : Morgan & Claypool Publishers, 2016.

[4] T. Kjær, C. B. Lillelund, M. Moth-Poulsen, N. C. Nilsson, R. Nordahl, and S. Serafin. Can you cut it: an exploration of the effects of editing in cinematic virtual reality. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, p. 4. ACM, 2017.

[5] B. Lang. Visionary vr is reinventing filmmaking's most fundamental concept to tell stories in virtual reality. 2015.

[6] Y.-C. Lin, Y.-J. Chang, H.-N. Hu, H.-T. Cheng, C.-W. Huang, and M. Sun. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2535–2545. ACM, 2017.

[7] A. MacQuarrie and A. Steed. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *Virtual Reality (VR), 2017 IEEE*, pp. 45–54. IEEE, 2017.

[8] L. Men, N. Bryan-Kinns, A. S. Hassard, and Z. Ma. The impact of transitions on user experience in virtual reality. In *Virtual Reality (VR), 2017 IEEE*, pp. 285–286. IEEE, 2017.

[9] K. R. Moghadam and E. D. Ragan. Towards understanding scene transition techniques in immersive 360 movies and cinematic experiences. In *Virtual Reality (VR), 2017 IEEE*, pp. 375–376. IEEE, 2017.

[10] L. T. Nielsen, M. B. Møller, S. D. Hartmeyer, T. C. M. Ljung, N. C. Nilsson, R. Nordahl, and S. Serafin. Missing the point: An exploration of how to guide users' attention during cinematic virtual reality. In *Proceedings of the 22Nd ACM Conference on Virtual Reality Software and Technology*, VRST '16, pp. 229–232. ACM, New York, NY, USA, 2016. doi: 10.1145/2993369.2993405

[11] T. Pakkanen, J. Hakulinen, T. Jokela, I. Rakkolainen, J. Kangas, P. Piippo, R. Raisamo, and M. Salmimaa. Interaction with webvr 360° video player: Comparing three interaction paradigms. In *Virtual Reality (VR), 2017 IEEE*, pp. 279–280. IEEE, 2017.

[12] R. Pausch, J. Snoddy, R. Taylor, S. Watson, and E. Haseltine. Disney's aladdin: first steps toward storytelling in virtual reality. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 193–203. ACM, 1996.

[13] S. Rothe and H. Hußmann. Guiding the viewer in cinematic virtual reality by diegetic cues. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*, pp. 101–117. Springer, 2018.

[14] S. Rothe, H. Hußmann, and M. Allary. Diegetic cues for guiding the viewer in cinematic virtual reality. In *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*, p. 54. ACM, 2017.

[15] S. Rothe, K. Tran, and H. Hußmann. Dynamic subtitles in cinematic virtual reality. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 209–214. ACM, 2018.

[16] L. Sassatelli, A.-M. Pinna-Déry, M. Winckler, S. Dambra, G. Samela, R. Pighetti, and R. Aparicio-Pardo. Snap-changes: a dynamic editing strategy for directing viewer's attention in streaming virtual reality videos. In *Proceedings of the 2018 International Conference on Advanced Visual Interfaces*, p. 46. ACM, 2018.

[17] A. Sheikh, A. Brown, Z. Watson, and M. Evans. Directing attention in 360-degree video. In *international Broadcasting Conference 2016*, pp. 29—-38. IET, 2016.

[18] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1535):3549–3557, 2009.