Towards assisting the decision-making process for content creators in cinematic virtual reality through the analysis of movie cuts and their influence on viewers' behavior

Carlos Marañes[a], Diego Gutierrez[a] and Ana Serrano[a,*]

[a]*Universidad de Zaragoza, Spain*

Abstract

Virtual Reality (VR) is gaining popularity in recent years due to the commercialization of personal devices. VR is a new and exciting medium to tell stories, however, the development of Cinematic Virtual Reality (CVR) content is still in an exploratory phase. One of the main reasons is that in this medium the user has now total or partial control of the camera, therefore viewers create their own personal experiences by deciding what to see in every moment, which can potentially hinder the delivery of a pre-established narrative. In the particular case of transitions from one shot to another (movie cuts), viewers may not be aligned with the main elements of the scene placed by the content creator to convey the story. This can result in viewers missing key elements of the narrative. In this work, we explore recent studies that analyze viewers' behavior during cinematic cuts in VR videos, and we discuss guidelines and methods which can help filmmakers with the decision-making process when filming and editing their movies.

Keywords: Virtual Reality; Cinematography; Experimental results; Visual Attention

1. Introduction

Virtual Reality (VR) has gained popularity in the last decade due to the commercialization of personal devices. The most common uses of VR technology include training, education, and leisure among others. However, the development of cinematographic content is still under an exploratory phase since the use of VR presents new challenges that have to be addressed. In this medium, the audience has total or partial control over the camera. This allows for the audience to explore the scene in many different ways and, depending on what each user decides to watch, different unique experiences may be achieved. This implies that directors have to deal with very different behaviors among users since some of them may not watch the important elements of the scene that directors have intentionally placed to convey their narrative.

---

*Author to whom all correspondence should be addressed (e-mail: anase@unizar.es).

Since the first movie ever created, cinematography has changed over the years. Filmmakers have improved the way movies are recorded and edited by introducing techniques such as zooms, changes of perspective, close-ups, etc., leading to a well-established cinematographic language. However, these techniques polished over the years in traditional cinema do not always directly apply to Cinematic Virtual Reality (CVR) content. To achieve the same effects, they have to be reinvented or adapted at least. For example, a change of perspective is commonly used to put the audience in the actor's point of view. This change does not only produce a shift in perspective but also a relocation to another part of the scene. On the contrary, in CVR the audience may not feel comfortable with changes where the point of view is strongly modified. This may affect viewers' natural behavior and prevent them from following the intended narrative. In Figure 1 an example of this technique is shown.
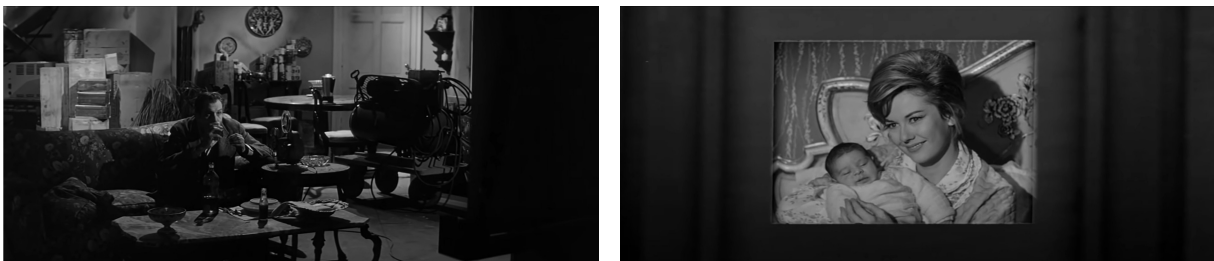

Fig. 1. In the movie *The Last Man on Earth*, the first shot before the cut (left) shows a man watching a video displayed by a projector, and in the next shot (right) the camera position changes to reveal the video for the audience, having a full understanding about what the character is perceiving. This example is difficult to carry out in a VR movie since the audience may need an adjustment period to understand the change of perspective.

Editing a VR movie is challenging, but there are recent works that propose different techniques to ease this task. Some suggested methodologies include greying-out uninteresting parts of the scene to make sure the user watches the intended part of the scene (Danieau et al., 2017) or catching users' attention using in-scene elements such as a firefly (Nielsen et al., 2016; Speicher et al., 2019; Kvisgaard et al., 2019), using auditory cues (Bala et al., 2019), or introducing different perceptual cues (Pillai and Verma, 2019a,b). Another way to help users follow the director's intentions is to rotate the virtual world (Stebbins and Ragan, 2019). This technique has been recently improved using machine learning in the work of Cha et al. (2020) in which they redirect users' attention through the desired path automatically. For a full analysis of current guidance techniques please refer to the work of Rothe et al. (2019).

Nevertheless, one of the main downsides of these techniques is that they may interfere with the users' experience because users may notice that their attention is being redirected. This also hinders the exploration of the 360º environment that VR offers, leading to a potentially negative effect on the experience. This is why a large body of recent work is focusing on studying how users explore the VR environments and applying the learned insights in a non-intrusive way.

When taking a look back in the history of cinematography, the usage of cuts represented a milestone in the way films are conceived, because they allow directors to film different shots and edit them in a way that gives rhythm to the conveyed narrative. In Hollywood films, shots have become shorter over the years. Nowadays the mean shot duration is around four seconds (Cutting et al., 2011). In these films, shots are usually taken at different times and locations, but viewers have no problem understanding

the changes between cuts, perceiving a coherent sequence of events. This is because shots are edited following the rules of continuity editing, making them look *invisible* to the viewer. Additionally, in traditional cinematography directors can decide how to frame their shots, focusing on the most relevant elements of the narrative and deciding for each scene what they want to show to the viewer. In contrast, when telling a story in VR, the viewer has control of the camera. Directors usually select parts of the scene that contribute to the narrative expecting that the audience will follow their intentions, e.g., two persons having a conversation or a person reading a letter aloud. These parts of the scene are usually called *regions of interest*. A common approach to increase the probability that the audience does not miss any relevant part of the story is to align these regions between cuts. This technique is shown in Figure 2, adapted from Jessica Brillhart (2016). The black dots represent the regions of interest and the different colored circles represent a 360º scene, separated by a cut. On the left part of the figure, dots are not aligned between cuts, so the audience may not follow them. A common approach to address this problem is to rotate each of the virtual scenes to make the dots aligned, as the right part of the figure shows.
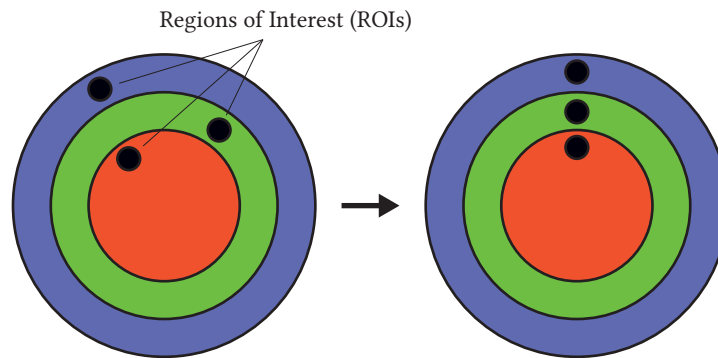
Fig. 2. Aligning Regions of Interest (ROIs) between cuts. The red, green, and blue circles represent a virtual world (scene), and each of the dots represents a ROI placed in each 360º world. If ROIs are not aligned between cuts (left) a common approach is to rotate the virtual worlds to make them aligned to avoid the viewer missing an important part of the scene (right). Example adapted from Jessica Brillhart (2016).

In this work, we focus our efforts on reviewing current state-of-the-art research related to the understanding of user's behavior through movie cuts. Understanding users' cognitive processes when visualizing a VR movie is key to deriving guidelines that can help in the process of decision-making for movie editing and directing users' attention as the narrative requires. The reviewed works analyze the viewers' behavior in short VR videos where only the visual component is studied and sound does not play an important role (see Martin et al. (2021) for a review of different modalities that can be leveraged for VR visualization).

First, since VR is a new paradigm it is crucial to investigate whether continuity editing is understood in the same way as in traditional cinema. We refer to the work of Serrano et al. (2017), which shows that the theory proposed by Magliano and Zacks (2011), which explains why cuts are well understood by the audience, also holds for CVR. We discuss these findings in Section 2. Then, we discuss tools to detect and measure the viewers' behavior through cuts, mainly extracted from the works of Serrano et al. (2017) and Marañes et al. (2020). We start discussing a possible parameterization for labeling a

cut depending on the content displayed before and after it, since different configurations may affect the behavior of the viewers (Section 3). Then we review visual tools that can help directors gain an intuition of where users are looking at (Section 4). Finally, we comment on some metrics to quantify this behavior and compare it (Section 5).

We hope that this review of existing tools and guidelines based on different studies of users' viewing behavior yields a comprehensive overview for supporting the decision-making process when filming and editing VR movies.

2. Continuity Editing and Event Segmentation

Since the first movie ever created by the Lumière brothers, *La Sortie de l'usine Lumière à Lyon* in 1895 filmmakers along history have introduced new techniques to tell stories. One of the groundbreaking changes established in cinematography has been the use of editing rules. In order to convey a smooth narrative over a series of shots and create a continuous flow of information, editing plays a strong role. This is achieved through *continuity editing* (Smith et al., 2016). Through a full movie, there are cuts with changes in location, time, action, etc. Despite these changes, the audience perceives these discontinuities as a coherent set of events, understanding the narrative behind the movie. For example, a very commonly used technique is the *match on action*. This technique is used to preserve spatial continuity and consists in following a given action through the cut, giving the sense of continuity. Figure 3 shows a real example of this technique.

Fig. 3. In the movie *Charade* a match on action is produced between cuts. On the left frame (before the cut), the woman is walking through a door, then on the right frame (after the cut), the camera has been moved to the new room where the woman is. This is a typical example of how continuity editing is applied in movies.

The higher-level cognitive processes that make continuity editing possible have been studied in traditional cinematography (Magliano and Zacks, 2011). The *event segmentation theory* (Kurby and Zacks, 2008) supports that segmentation is an automatic process that happens in the brain when we observe a series of events, and movie shots mimic this automatic process. In the experiment performed by Zacks et al. (2001), they showed that subjects were able to parse a series of activities in a movie into different, meaningful events. Through an fMRI, they also measured which parts of the brain were involved in the segmentation process. The authors suggest that brain processes that take part in event segmenta-

tion are automatic (Zacks and Swallow, 2007) and that event segmentation also helps the brain structure the information for retrieving memories (Zacks et al., 2006; Sargent et al., 2013; Flores et al., 2017), organizing internal memory, and relating different events (Reynolds et al., 2007). This discrete mental representation is used as a basis for predicting the immediate course of events. If a prediction is violated, a new event will occur. This mechanism is used to explain why continuity editing works, supporting the idea that the relation between cuts and perceived event boundaries leads to the perceived continuity (Magliano and Zacks, 2011).

In order to investigate whether cuts are perceived in the same way and whether continuity editing holds in CVR, a similar experiment was replicated in VR by Serrano et al. (2017). Their results suggest that viewers also understand the events in a similar way to traditional movies and therefore the rules of continuity editing still hold. Nevertheless, the content before and after the cut may affect users' experience and how they explore the VR scene, so this behavior should be studied. In the following sections, we overview different tools which can be helpful for directors when designing and editing their movies.

3. Editing in Cinematic Virtual Reality

Serrano et al. (2017) have shown that continuity editing holds in CVR. This suggests that the audience structures the events in their minds in a similar manner to a traditional movie. Content creators can use cuts to convey their stories in VR and viewers will perceive the continuous flow of information, parsing the events and storing them in an organized fashion. However, when designing scenes in CVR there should be awareness of how to place the different scene elements in order to increase the likelihood of the audience watching them, since they may provide crucial information to understand the narrative. These parts are usually referred to as Regions of Interest (ROIs) and a shot can have none or several ROIs. Notice that if a shot has more than one ROI users may experience difficulty when directing their attention.

When transitioning from one shot to another, directors introduce a cut. This implies that the audience experiences a change from one 360º world to another and they have to redirect their attention to the new ROI, if any. One of the challenges that the use of 360º scenes presents when conveying a story is that viewers now have control of the camera, preventing directors from framing the part of the scene that they would like to show to the audience. To help with the process of decision-making, several works analyze how viewers behave in 360º scenes and in VR movies, providing useful guidelines for filmmakers (Serrano et al., 2017; Marañes et al., 2020; Fearghail et al., 2018). Depending on how they design their *mise-en-scène* (the arrangement of actors and scenery on stage), the viewers' behavior can be influenced by the number of ROIs, the temporal changes between cuts, the alignment of the ROIs between cuts, etc. Knowing the elements that define a cut and a scene helps to identify potential viewing patterns. This can bring insights that facilitate the task of recording the movie by placing the elements in a certain manner and also when composing the different shots to create the final VR movie. Given the high dimensional parameter space that composes a cut, in this section we present different classifications of cuts based on their features (Serrano et al., 2017; Marañes et al., 2020).

Type of cut. Scenes between cuts can differ in space, time, and/or action (Smith et al., 2016). In the study performed by Magliano and Zacks (2011), traditional filmed movie cuts are classi-

fied in three types: Cuts which are discontinuous in space or time and discontinuous in action (action discontinuities); cuts that are discontinuous in space or time but continuous in action (spatial/temporal discontinuities); and cuts which are continuous in space, time, and action (continuity cuts). This last type of cut usually displays changes in viewpoint in the same scene and is rarely used in CVR (Serrano et al., 2017). However, there are works that use this type of cut to redirect the users' attention (Sassatelli et al., 2018), since it adds rhythm to the movie. This classification has been used by Serrano et al. (2017) to test continuity editing in CVR. In Figure 4 an example of each cut is shown. As can be seen in the action discontinuity cut example, the frames around the cut show different places without relation between them. The spatial/temporal discontinuity cut example is a part of a sequence where the same element is shown from different locations. In the continuity cut, the same scene is shown from different viewpoints, matching space, time and action.

Number of ROIs.   Directors can decide to include a different number of ROIs to convey their narrative. When a cut is produced, the number of ROIs can vary between shots, and users' behavior may be affected. When going from a scene with a ROI and then in the following scene no ROIs appear, the audience may search for the previous ROI. Depending on the number of ROIs before and after the cut, a number of possibilities arise. The number of ROIs before and after the cut is an influencing parameter in audience's behavior (Serrano et al., 2017; Marañes et al., 2020). Figure 5 shows an example of scenes with different numbers of ROIs.

Alignment of ROIs.   When a cut is produced in traditional cinematography, users experience a transition from a frame to another frame, but in CVR the audience experiences a transition from a VR world to another. Directors can decide how to place different ROIs in order to direct viewers' attention, taking into account how this will affect the transition between different scenes. A common approach to address this issue is to rotate these 360º worlds in order to make ROIs aligned through cuts. However, this technique does not leverage all the potential that VR offers since users are focused on a small region of the scene. Another possible cut classification arises by taking into account different misalignments of the ROIs between cuts (Serrano et al., 2017).

Editing techniques to support decision-making.   The previous taxonomy is a plausible way to classify movie cuts. Due to the high dimensional parameter space, previous work has limited the research space by taking into account some subset of cuts from the previous classification (Serrano et al., 2017; Marañes et al., 2020). This classification gives directors a solid starting point: they can label their scenes according to recent studies which take into account common tendencies when editing a VR movie. Once labeled, viewers' behavior can be analyzed by directly visualizing it (Section 4) or computing metrics (Section 5) to understand how viewers explore the VR environments. We expect follow-up research taking into account different scene features. Note that this classification does not account for sound, music, and/or soundtracks. In the studies reviewed in this work, the audio component of the analyzed movies and videos does not play an important role. However, recent research suggests that directional sound cues can influence users' attention while consuming CVR content (Masia et al., 2021).
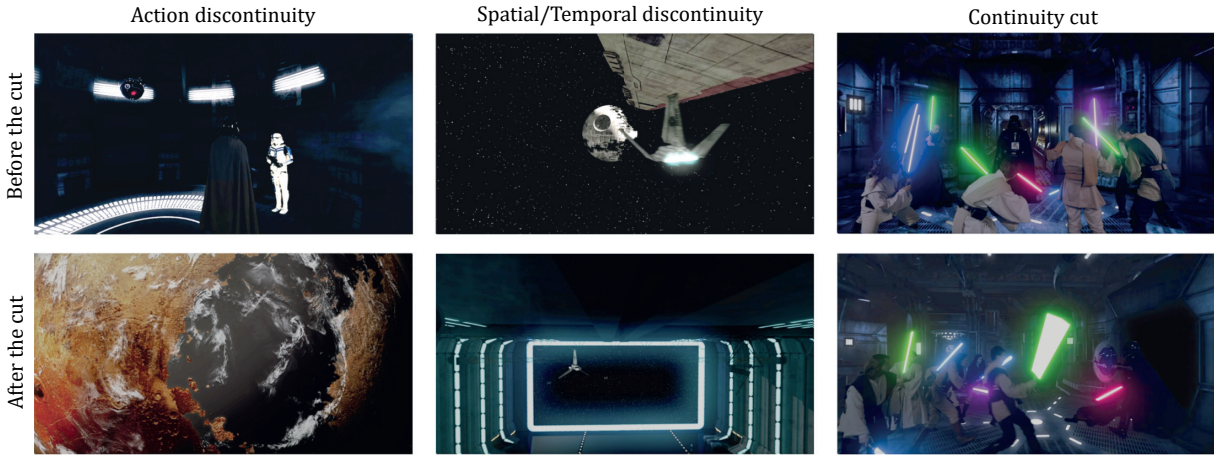
Fig. 4. Frames extracted from the VR movie *Star Wars - Hunting of the Fallen*, before and after the cut of the three types of cuts regarding continuity. **Left:** Example of an action discontinuity cut. **Center:** Example of spatial/temporal discontinuity cut. **Right:** Example of a continuity cut. Figure adapted from Serrano et al. (2017).

Fig. 5. From left to right, frames extracted from 360º clips showing examples of a scene with no ROIs, a scene with one ROI, and a scene with two ROIs. Frames adapted from Marañes et al. (2020) and Serrano et al. (2017).

Sound is always present across different fields of view, e.g., a user may not be looking to a specific region of interest at a given time, but sound will always be heard regardless of where the user is looking, offering a wide range of possibilities (Serafin et al., 2018). Although this work is not centered on analyzing the sound modality, researchers have proposed tools for leveraging audio in virtual environments (Naef et al., 2002) and guiding users' attention (Rothe et al., 2017).

4. Visualizing User Behavior

VR movies offer unique user experiences, allowing them to explore shots as they wish. This complicates the analysis of their behavior since viewers can be watching different regions of the scene at the same time. In order to understand how viewers consume VR content, their behavioral data needs to be analyzed.

In this section, we mainly discuss the results of different works that analyze the head orientation of the users. Although eye-tracking is becoming more affordable, arguably the most common consumer-level

HMDs (e.g., HTC Vive and Vive Pro, Oculus Quest, and Valve Index) do not have built-in eye-tracking available. Some devices, such as the HTC Vive Pro Eye which is targeted for business use, do include built-in eye-tracking but they are outside the scope of consumer-level devices. In most cases, an eye-tracking add-on has to be acquired separately from the HMD and it needs to be adjusted and calibrated by the user. However, an eye-tracker is not always necessary: in the work of Sitzmann et al. (2018) it is shown that head orientation is a robust proxy for estimating users' gaze. The most common technique to visualize all the viewers' information together is by computing saliency maps. A saliency map gives information about where most of the viewers are looking at. This is a powerful tool for directors since they can obtain feedback regarding how users are actually exploring the scenes. In Figure 6 on the left, there is a frame of a VR movie in which every single point represents the head orientation of a viewer. Due to the complexity of extracting valuable information from this raw representation, saliency maps provide an alternative representation of viewing behavior easier to interpret.

Fig. 6. **Left:** Frame of a VR movie in which each color point represents the head orientation of a viewer. **Right:** Saliency map computed taking into account the viewers' head orientation. Figure adapted from Marañes et al. (2020).

To compute a saliency map, a convolution is performed over the users' head orientation by using a Gaussian kernel (see Figure 6, right). However, not every sample has the same relevance in this process. To assess users' attention, two types of eye movements are typically considered: *fixations* and *saccades*. A fixation is considered when the user is paying attention by maintaining the gaze on a single fixed location. On the other hand, saccades are produced between fixations when viewers shift their gaze to a different part of the scene. Although these two behaviors are described for eye movements, Sitzmann et al. (2018) showed that they can also be estimated by leveraging the head velocity: While fixating, head velocity remains under a certain threshold. Due to the difficulty of estimating where viewers are looking at, recent work has focused their efforts on predicting saliency information (Martin et al., 2020; Ling et al., 2018; Chao et al., 2018; Chen et al., 2020; Zhu et al., 2019). Saliency prediction models are remarkably effective for content creators since they can obtain an accurate saliency map given only the stimuli they are working on.

From the saliency information, a useful metric to consider is the *Inter-Observer Congruency* (IOC) to understand the users' viewing agreement. This metric is unitless and its value ranges from 0 to 1, where a value of 0 indicates that users are observing different regions of the scene while a value of 1 indicates that users are watching exactly the same regions of the scene, at a given time. The metric is computed by using a leave-one-out-approach: the i_{th} subject is left and all the other users' fixations are aggregated by accumulating a time window; then the percentage of fixations of the i_{th} user that fall within the k% most salient regions predicted by the aggregated saliency map is computed. This process is repeated for all users and the mean value is computed.

Figure 7 shows an example of the IOC metric computed for $k \in [0\%..10\%]$ in 2.0 increments. This figure has been extracted from the work of Marañes et al. (2020), where only head orientation logged with the HMD has been used. This work observes users' behavioral patterns that content creators can leverage to achieve the desired behavior in their audience. The authors analyze the six first seconds of the scene after the cut for four different types of cuts, where the scene previous to the cut has a unique ROI or no ROIs at all and likewise the scene after the cut. Content creators can leverage this metric to observe the convergence to the main action after the cut and modify their shots to achieve their desired effect. In Figure 7 *a* and *b*, in both scenes there is a ROI, so users converge to a high congruency since the majority shows an interest in the ROI, watching the same part of the scene. However, in Figure 7 *a* the previous scene has a ROI and the ROIs before and after the cut are aligned, so the IOC is high in the starting seconds, while in Figure 7 *b* where there is no ROI in the scene previous to the cut, the users are scattered around the scene and they have to refocus their attention after the cut, so they converge more slowly. On the other hand, when there are no ROIs in the scene after the cut, users are scattered around the scene. It can be observed as a low IOC in Figure 7 *c* and *d*.
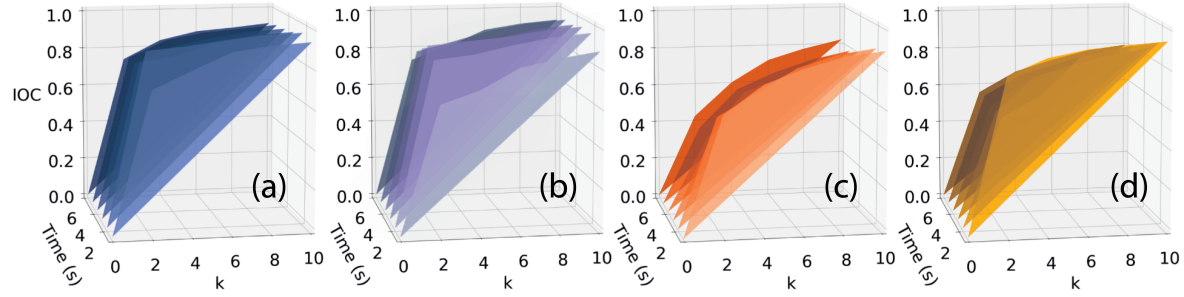

Fig. 7. IOC *(Inter-Observer Congruency)* computed for four different type of cuts. **(a):** The scene after the cut in which the metric is computed has a well-defined ROI so it presents a high metric value along the six seconds. Additionally, the scene previous to the cut also has a single ROI so it can be observed that the IOC quickly converges to a high value. **(b):** Similarly to the previous case, the figure shows how users consistently have a common behavior after the second two of the scene, indicating that there is a ROI after the cut however, at the beginning of the shot the IOC value is low since the scene previous to the cut has no ROI and users are scattered all over the scene. **(c):** The metric value remains low in every computed second, since there are no ROIs in neither scenes, before or after the cut. **(d):** Viewers have a consistent behavior at the beginning of the shot since they are directing their attention to a ROI before the cut, while after the cut viewers' behavior does not show that agreement since there is no ROI to direct their attention. Figure adapted from Marañes et al. (2020).

However, this metric may be hard to interpret due to its 3D nature, which makes it difficult to visualize. For this reason, the *Area Under the Curve* (AUC) can be computed by using the previous IOC metric values, simplifying the previous plot, and providing an easier interpretation. This metric is computed by integrating the area under the curve for each temporal interval. This metric is also unitless: an AUC of 0 indicates no congruency between users and an AUC of 100 indicates total congruency. Therefore, the AUC becomes a useful metric to gain an intuition about the users' behavior through cuts. Figure 8 illustrates an example of the AUC values for each cut of the previous IOC example. This figure also shows the displayed scenes and their associated saliency maps. In Figure 8 *a* and *b* the AUC remains high since these scenes display a ROI after the cut, while in Figure 8 *c* and *d* the AUC metric remains low since there is no ROI after the cut.
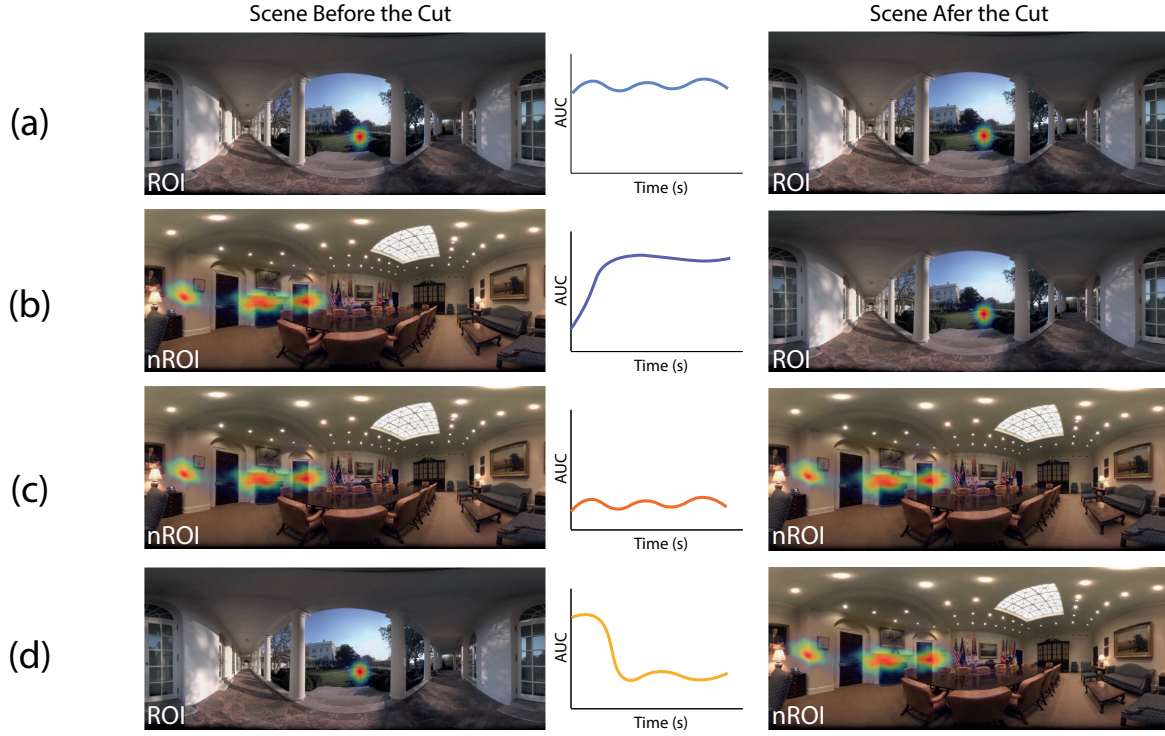
Fig. 8. Example of the AUC metric during the first six seconds after a cut for different combinations of ROI configurations before and after the cut. The frames and associated saliency maps in the figure are representative of the scene configurations before (left) and after (right) the cut. **(a):** When there is a well-defined ROI before and after the cut, users usually focus on it, therefore the AUC remains high along time. **(b):** When there is no ROI in the scene before the cut users tend to explore the scene so their attention is scattered. After the cut, it takes some time for users to refocus their attention on the ROI in the scene. This behavior results in low AUC values in the first seconds so after the users focus on the new ROI, this value increases. **(c):** There is no clear ROI before and after the cut, therefore users explore both scenes, resulting in a low AUC value along time. **(d):** Users are focused on the ROI of the first scene and when the second scene starts they start to scatter around the scene. This behavior results in a high AUC value at the beginning that then decreases and remains low.

Following this line of searching for users' similarities, Rossi et al. (2020) propose a novel metric based on a clique clustering algorithm (Rossi et al., 2019) called *User Affinity Index* (UAI). The key idea of their metric is to identify tendencies among users, detecting firstly users' clusters based on their agreement and then measuring the average of cluster popularity. These metrics report similar information to previous ones, making it a reliable alternative depending on the specific behavior to detect.

Visualization of user behavior to support decision-making. Saliency maps, and IOC and AUC metrics are powerful tools for content creators to obtain feedback from their productions. Metrics' results and previous knowledge about users' behavior can be used to jointly derive conclusions. In the work of Sitzmann et al. (2018) a bias is detected when users explore a static 360º panorama, the *equator bias*, meaning that users have a preference to explore the horizon line over the rest of the panorama.

Additionally, they suggest that users explore most of the scene after about 20 seconds. On the contrary, when observing VR movies instead of static panoramas, users may roam the scene in a different manner. In the work of Fearghail et al. (2018) they analyze how users behave in a short set of VR movies, both with and without cuts. Their results suggest that longer shots are preferred because users can freely explore the VR environment without worrying about missing relevant narrative aspects. Similarly, in the work of Marañes et al. (2020) a professionally edited movie produced by *Felix & Paul Studios* is analyzed. In their study, saliency maps are computed to understand how people explore a CVR movie. Figure 9 shows an example of two saliency maps for different cut combinations. In the left figure, it is shown a saliency map for a scene designed with a ROI. From that study, it is concluded that if the ROI is intended to contribute to the narrative, users will follow it with no problems because, as can be seen in the frame, users are mostly focused on the ROI. By contrast, in the right figure users are scattered through the scene. This is expected since there is no well-defined ROI. Depending on how the director is conveying the story, directors can detect where most users are looking in order to maximize the probability of redirecting the viewers' attention to the intended region of the scene (Fearghail et al., 2019).

In order to provide a more quantitative view of the users' behavior, IOC, AUC, and UAI metrics can be computed to detect if users show a similar behavior over time. For example, they can be used to know if a ROI will retain users' attention (Marañes et al., 2020). Rossi et al. (2020) have used their UAI metric to detect variations in users' behavior when watching VR clips with different devices, showing that HMDs lead to similar navigation patterns when there is a well-defined ROI. The computation time when deciding which metrics to compute should also be a decisive factor. For example, the IOC metric has a high computation time cost since each of the points that define the IOC curve grows exponentially with the number of users (leave-one-out-approach). The AUC curve needs previously the computation of the IOC metric, but once the IOC metric is computed the time cost of the AUC metric is negligible.

We expect that these metrics can be useful for assisting content creators to achieve their desired effect on the audience by providing feedback on users' expected behavior for their productions. A particular use case that leverages these metrics and visualization could be the following. A content creator has recorded footage for a VR documentary and has to edit it to compose the final content. In one of the scenes, there is a person sitting in a chair in a field talking about a farm (single ROI scene), and in another scene, there are some farm animals (nROI scene, without a clear single ROI). The creator would like to place first the scene with the person talking and then perform a cut and show the animals. Following the guidelines illustrated in Figure 8 *d*, the creator can expect a behavior similar to the yellow AUC line: users would first focus on the person and then each user would scatter around the scene, focusing on the different animals and therefore showing a low agreement. This needs to be taken into account when cutting to a new scene afterward: if users are scattered, they will need more time to converge to the main action after the cut.

5. Measuring User Behavior in Cinematic Virtual Reality

In Section 4 we have shown how to visualize users' behavior jointly through saliency maps. They provide content creators feedback about how users are actually exploring their scenes and help them detect repeated patterns among the audience. However, saliency maps are designed to be interpreted by humans,

Fig. 9. **Left:** Saliency map of a scene with only one ROI. **Right:** Saliency map of a scene with no clear ROIs defined. This feedback is useful for content creators. For example, if the saliency map of the left image resembles the right image, where users are scattered through the scene, this can suggest that the ROI is not engaging enough. Figure adapted from Marañes et al. (2020).

not offering quantitative measurements about users' behavior. In this section, we provide computable metrics for 360º scenes to measure different visualization patterns. These metrics are influenced by the scene configuration and cut combinations (see Section 3 for a description of scene configurations and cut classifications). In this section, we provide a set of metrics proposed by previous work (Serrano et al., 2017; Marañes et al., 2020) that measure user's behavior in a quantitative manner for different cut combinations. These metrics are usually computed in the scene after the cut in order to measure how the previous scene influences the viewing behavior after the cut.

Frames to reach a ROI. Directors place ROIs to convey their narrative along with the shots that compose the movie. When going through the movie, the audience usually redirects their attention to the different ROIs to follow the story. However, ROIs may not be aligned between cuts, so users have to redirect their attention to the new ROI each time a cut is produced. This metric is indicative of the time that it takes the viewer to converge to the ROI after the cut. A high metric value means that users experience more difficulty finding the ROI while a low value means that the audience finds easily the ROI. This metric is affected by the alignment between ROIs before and after the cut since if a large misalignment is present, users will not be aligned with respect to the new ROI after the cut and they will have to explore the scene until they find it (Serrano et al., 2017).

Percentage of total fixations inside the ROI. Designing a ROI to retain the audience's attention enough time to make sure they are understanding the narrative is a challenging task. This metric gives an indication of users' interest in the ROI. It is computed by taking into account the fixations that are inside the ROI divided by the total number of fixations. To avoid taking into account fixations produced when the user is searching for the ROI, the metric is only computed once the user has found the ROI. A high metric value means that the user is truly interested in the ROI while a low value is indicative that users are showing a poor interest in it.

Number of fixations. This metric gives an indication of the users' visual behavior. A high value indicates that the user is having a more fixating behavior whereas a low value indicates that the

user is showing a more exploratory behavior, performing more saccadic eye movements. This metric can be computed both when the user has found the ROI or not. If the user has fixated on a ROI this metric gives an indication of the user's behavior once he knows where the ROI is and the value is interpreted taking into account the ROI's influence. On the other hand, if the restriction of finding the ROI is relaxed, it indicates the overall behavior.

Total distance traveled. In a 360º scene, users can explore every degree that surrounds them. This metric is indicative of the users' desire to explore. It is computed by accumulating the or-thodromic distance (or great-circle distance) that the user has roamed while watching the scene. A high metric value means that users have roamed more whereas a low value indicates that users have explored less the scene. Notice that this metric is independent of the number of fixations since it is computed taking into account all gaze samples, and it can also be computed only with head-tracking data.

Percentage of the scene watched. A VR scene is usually projected onto a sphere that surrounds the user. This metric is an indicator of the sphere percentage that the viewer has actually seen. A part of the scene is considered viewed if the user has fixated on it. This metric gives an indication about the users' behavior when extracting information from a scene since it is computed only when fixating and the value does not change once he fixates again on a part of the scene already watched. A high value means that the user has seen more different parts of the scene whereas a low value means that the user has been revisiting the same regions. This metric does not necessarily need eye-tracking data and can be computed only with head-tracking data.

Metrics to support decision-making. In the process of decision-making when editing a VR movie, content creators do not know with certainty how the cut combinations will influence the users' behavior, and this can lead to undesired effects when conveying the story. By using these metrics creators can have quantitative measurements about how users are exploring their scenes and they can iterate the process of decision-making until achieving the desired behavior. In the work of Serrano et al. (2017) these metrics are used for analyzing users' behavior for different cut combinations, with different alignments between cuts and different configurations of the ROIs in the scene. They derive useful guidelines for content creators, e.g., for a fast-paced action movie, ROIs should be aligned across cuts, while to evoke a more exploratory behavior, misalignments are recommended. Regarding the ROIs position, we may suggest placing them in the equator line, since users are prone to explore more the parts of the scene that are situated in that region (Sitzmann et al., 2018). Besides, content creators should be aware that there seems to be an exploration peak at the beginning of the cut (Serrano et al., 2017). It is hypothesized that users require some time to adapt to the new visual content, so content creators should give the user some time before starting to tell a story through a ROI, because users may not be paying attention to it. This is in line with the work of Fearghail et al. (2018) in which users are more engaged in longer shots because they are not worried about missing any crucial element to understand the movie narrative. In the work of Marañes et al. (2020) an exploratory behavior is detected when there are no regions of interest in the scene. This is in line with the work of Fearghail et al. (2018) in which the authors also study a documentary movie and report an exploratory behavior. In that work, the director provided the desired scan-path about how

users should watch the movie but there were not any essential parts for the narrative. With the use of the previously mentioned metrics, in the work of Marañes et al. (2020) it is suggested that a scene with no ROIs followed by a scene with a ROI makes the user be more focused on the ROI of the second scene. This insight is useful for content creators since if they would like to increase the probability that a user is paying attention to the ROI, the director could place a scene that encourages exploration before the cut. In their work, it is also suggested that if users have been in a scene with a ROI, it does not matter if the scene after the cut contains or not a ROI because users will show a more exploratory behavior. This suggests that the exploration desire of the audience should not be restricted. However, these insights may vary in the presence of auditory cues. Recent research suggests that directional sound cues influence the viewing behavior (Masia et al., 2021): users converge faster to the ROI after the cut in the presence of directional sound cues. In that work, the authors also suggest that directional sound cues can be used to alert the viewer that there may be different ROIs in the same scene outside of the users' field of view.

We encourage filmmakers to use these metrics and insights to evaluate their scenes and we also expect follow-up research by providing new guidelines that help to consolidate the cinematographic language for VR.

6. Conclusion and Future Work

Making a VR movie entails decision-making by the directors when filming and also when editing it. While traditional cinema has relied on a well-established set of rules grouped under the term of *continuity editing*, virtual reality cinema is still in an exploratory phase. However, recent work described in Section 2 shows that continuity editing still holds in CVR, suggesting that viewers also build a mental model to represent the information perceived during the movie even through cuts. Directors can leverage this knowledge to think about cuts in a similar manner to traditional cinema.

At the scene design phase, the distribution of elements in the scene is of great importance to conveying the narrative, since viewers' attention can be drawn to different parts of the scene. Furthermore, when editing the different shots to compose the resulting movie, the nature of the scene before and after the cut influences viewers' behavior (Serrano et al., 2017; Marañes et al., 2020). The understanding of the cognitive processes that lead to a specific behavior is key to developing tools for assisting VR content creators and directors in the decision-making process when editing their VR movies. In order to help identify the types of scenes and cuts, we review in Section 3 a subset of scene features to take into account when this process takes place based on previous work. In order to have an overview of how viewers are behaving when watching the movie, in Section 4 we suggest the use of saliency maps and, to have a quantitative measurement about the agreement among users we also discuss the use of the metrics *Inter-Observer Congruency* and *Area Under the Curve*. In addition to these metrics, in Section 5 we review metrics proposed by previous works to have quantitative measurements of various types of viewing behavior. We hope that VR filmmakers can leverage these tools to obtain feedback from their audience and that this provides insights that can help them compose their movies to achieve the desired effect on their viewers.

Future Work.   The insights discussed in this work may vary depending on the type of content. Documentary movies, where the action is scarce and has a more paced rhythm, can lead to a very different viewing behavior than action movies, where a lot of movement is happening on stage. Additionally, VR environments do not only offer cinematic content but also interactive content where the viewer has an active role in the environment, responding to certain stimuli. Further research is needed in this direction in order to understand users' viewing behavior and attention in these different scenarios.

In this work, we have discussed different ways of analyzing viewers' behavior across movie cuts based on previous work. However, the transition between cuts can be performed in multiple formats, having a direct impact on the viewing behavior. For example, a *fade-to-black* technique can lead to different viewing patterns compared to a sharp cut, where viewers transition from one scene to another abruptly. Future work can consider the study of a wider range of transition techniques to quantify their potential effect on exploration patterns.

The number of possibilities regarding VR content is large, and for every possibility, viewing behavior can be influenced. Content can be presented supporting only three degrees-of-freedom (head rotations), which is usually the main format for VR cinematic content. However, six degrees-of-freedom (head rotations and translations) can be accomplished by using computer-generated content or alternatively by recording real content with specialized cameras (Overbeck et al., 2018) or using computational methods that allow for additional degrees of freedom from monocular content (Serrano et al., 2019; Attal et al., 2020; Richardt et al., 2020). Other factors such as the presence of sound and/or music, the exposure time to VR content, or the complexity of visual content may also influence viewing behavior. Given the highly dimensional parameter space of VR content, we hope that researchers and practitioners will continue to explore different scene configurations and new techniques towards developing a comprehensive language for storytelling in VR.

Acknowledgments

References

Attal, B., Ling, S., Gokaslan, A., Richardt, C., Tompkin, J., 2020. Matryodshka: Real-time 6dof video view synthesis using multi-sphere images. In *European Conference on Computer Vision*, Springer, pp. 441–459.

Bala, P., Masu, R., Nisi, V., Nunes, N., 2019. "When the Elephant Trumps" A Comparative Study on Spatial Audio for Orientation in 360º Videos. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13.

Brillhart, J., 2016. In the Blink of a Mind. https://medium.com/the-language-of-vr/in-the-blink-of-a-mind-prologue-7864c0474a29. Accessed: 2021-21-12.

Cha, S., Lee, J., Jeong, S., Kim, Y., Noh, J., 2020. Enhanced Interactive 360º Viewing via Automatic Guidance. *ACM Transactions on Graphics (TOG)*

Chao, F.Y., Zhang, L., Hamidouche, W., Deforges, O., 2018. Salgan360: Visual saliency prediction on 360 degree images with

generative adversarial networks. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, pp. 01–04.

Chen, D., Qing, C., Xu, X., Zhu, H., 2020. SalBiNet360: Saliency Prediction on 360º Images with Local-Global Bifurcated Deep Network. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, IEEE, pp. 92–100.

Cutting, J.E., Brunick, K.L., DeLong, J.E., Iricinschi, C., Candan, A., 2011. Quicker, faster, darker: Changes in Hollywood film over 75 years. *i-Perception* 2, 6, 569–576.

Danieau, F., Guillo, A., Doré, R., 2017. Attention guidance for immersive video content in head-mounted displays. In *2017 IEEE Virtual Reality (VR)*, IEEE, pp. 205–206.

Fearghail, C.O., Knorr, S., Smolic, A., 2019. Analysis of Intended Viewing Area vs Estimated Saliency on Narrative Plot Structures in VR Film. In *2019 International Conference on 3D Immersion (IC3D)*, IEEE, pp. 1–8.

Fearghail, C.O., Ozcinar, C., Knorr, S., Smolic, A., 2018. Director's Cut - Analysis of Aspects of Interactive Storytelling for VR Films. In *International Conference on Interactive Digital Storytelling*, Springer, pp. 308–322.

Flores, S., Bailey, H.R., Eisenberg, M.L., Zacks, J.M., 2017. Event segmentation improves event memory up to one month later. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 43, 8, 1183.

Kurby, C.A., Zacks, J.M., 2008. Segmentation in the perception and memory of events. *Trends in cognitive sciences* 12, 2, 72–79.

Kvisgaard, A., Klem, S.Ø., Nielsen, T.L., Rafferty, E.I., Nilsson, N.C., Høeg, E.R., Nordahl, R., 2019. Frames to Zones: Applying Mise-en-Scène Techniques in Cinematic Virtual Reality. In *2019 IEEE 5th Workshop on Everyday Virtual Reality (WEVR)*, IEEE, pp. 1–5.

Ling, J., Zhang, K., Zhang, Y., Yang, D., Chen, Z., 2018. A saliency prediction model on 360 degree images using color dictionary based sparse representation. *Signal Processing: Image Communication* 69, 60–68.

Magliano, J.P., Zacks, J.M., 2011. The Impact of Continuity Editing in Narrative Film on Event Segmentation. *Cognitive Science* 35, 8, 1489–1517.

Marañes, C., Gutierrez, D., Serrano, A., 2020. Exploring the impact of $360°$ movie cuts in users' attention. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*.

Martin, D., Malpica, S., Gutierrez, D., Masia, B., Serrano, A., 2021. Multimodality in VR: A survey. *ACM Computing Surveys*

Martin, D., Serrano, A., Masia, B., 2020. Panoramic convolutions for $360°$ single-image saliency prediction. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*.

Masia, B., Camon, J., Gutierrez, D., Serrano, A., 2021. Influence of directional sound cues on users exploration across $360°$ movie cuts. *IEEE Computer Graphics and Applications*. doi: 10.1109/MCG.2021.3064688.

Naef, M., Staadt, O., Gross, M., 2002. Spatialized audio rendering for immersive virtual environments. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pp. 65–72.

Nielsen, L.T., Møller, M.B., Hartmeyer, S.D., Ljung, T.C., Nilsson, N.C., Nordahl, R., Serafin, S., 2016. Missing the point: an exploration of how to guide user' attention during cinematic virtual reality. In *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, pp. 229–232.

Overbeck, R.S., Erickson, D., Evangelakos, D., Pharr, M., Debevec, P., 2018. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Transactions on Graphics (TOG)* 37, 6, 1–15.

Pillai, J.S., Verma, M., 2019a. Grammar of VR Storytelling: Analysis of Perceptual Cues in VR Cinema. In *European Conference on Visual Media Production*, pp. 1–10.

Pillai, J.S., Verma, M., 2019b. Grammar of VR Storytelling: Narrative Immersion and Experiential Fidelity in VR Cinema. In *The 17th International Conference on Virtual-Reality Continuum and its Applications in Industry*, pp. 1–6.

Reynolds, J.R., Zacks, J.M., Braver, T.S., 2007. A Computational Model of Event Segmentation From Perceptual Prediction. *Cognitive Science* 31, 4, 613–643.

Richardt, C., Tompkin, J., Wetzstein, G., 2020. Capture, reconstruction, and representation of the visual real world for virtual reality. In *Real VR–Immersive Digital Reality: How to Import the Real World into Head-Mounted Immersive Displays*. Springer International Publishing, pp. 3–32.

Rossi, S., De Simone, F., Frossard, P., Toni, L., 2019. Spherical Clustering of Users Navigating 360º content. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 4020–4024.

Rossi, S., Ozcinar, C., Smolic, A., Toni, L., 2020. Do Users Behave Similarly in VR? Investigation of the User Influence on the System Design. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2, 1–26.

Rothe, S., Buschek, D., Hußmann, H., 2019. Guidance in Cinematic Virtual Reality-Taxonomy, Research Status and Challenges.

Multimodal Technologies and Interaction 3, 1, 19.

Rothe, S., Hußmann, H., Allary, M., 2017. Diegetic cues for guiding the viewer in cinematic virtual reality. In *Proceedings of the 23rd ACM symposium on virtual reality software and technology*, pp. 1–2.

Sargent, J.Q., Zacks, J.M., Hambrick, D.Z., Zacks, R.T., Kurby, C.A., Bailey, H.R., Eisenberg, M.L., Beck, T.M., 2013. Event segmentation ability uniquely predicts event memory. *Cognition* 129, 2, 241–255.

Sassatelli, L., Pinna-Déry, A.M., Winckler, M., Dambra, S., Samela, G., Pighetti, R., Aparicio-Pardo, R., 2018. 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*, pp. 1–5.

Serafin, S., Geronazzo, M., Erkut, C., Nilsson, N.C., Nordahl, R., 2018. Sonic interactions in virtual reality: state of the art, current challenges, and future directions. *IEEE computer graphics and applications* 38, 2, 31–43.

Serrano, A., Kim, I., Chen, Z., DiVerdi, S., Gutierrez, D., Hertzmann, A., Masia, B., 2019. Motion parallax for 360° RGBD video. *IEEE Transactions on Visualization and Computer Graphics* 25, 5.

Serrano, A., Sitzmann, V., Ruiz-Borau, J., Wetzstein, G., Gutierrez, D., Masia, B., 2017. Movie Editing and Cognitive Event Segmentation in Virtual Reality Video. *ACM Transactions on Graphics (SIGGRAPH 2017)* 36, 4.

Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., Wetzstein, G., 2018. Saliency in VR: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics* 24, 4, 1633–1642.

Smith, J., Bordwell, D., Thompson, K., 2016. Film Art: An Introduction.

Speicher, M., Rosenberg, C., Degraen, D., Daiber, F., Krúger, A., 2019. Exploring Visual Guidance in 360-degree Videos. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 1–12.

Stebbins, T., Ragan, E.D., 2019. Redirecting View Rotation in Immersive Movies with Washout Filters. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, IEEE, pp. 377–385.

Zacks, J.M., Braver, T.S., Sheridan, M.A., Donaldson, D.I., Snyder, A.Z., Ollinger, J.M., Buckner, R.L., Raichle, M.E., 2001. Human brain activity time-locked to perceptual event boundaries. *Nature neuroscience* 4, 6, 651–655.

Zacks, J.M., Speer, N.K., Vettel, J.M., Jacoby, L.L., 2006. Event understanding and memory in healthy aging and dementia of the Alzheimer type. *Psychology and aging* 21, 3, 466.

Zacks, J.M., Swallow, K.M., 2007. Event Segmentation. *Current directions in psychological science* 16, 2, 80–84.

Zhu, Y., Zhai, G., Min, X., Zhou, J., 2019. The Prediction of Saliency Map for Head and Eye Movements in 360 Degree Images. *IEEE Transactions on Multimedia*