Exploring the impact of 360° movie cuts in users' attention

Carlos Marañes* Diego Gutierrez† Ana Serrano‡

Universidad de Zaragoza

Figure 1: Example frame from the movie *The People's House* produced by *Felix & Paul Studios* used to study how people explore professionally edited 360° movies. For analyzing how users behave across movie cuts, we start from head orientation data (left), and we take into account visual attention priors in order to compute saliency maps (right) describing users' gaze.

ABSTRACT

Virtual Reality (VR) has grown since the first devices for personal use became available on the market. However, the production of cinematographic content in this new medium is still in an early exploratory phase. The main reason is that cinematographic language in VR is still under development, and we still need to learn how to tell stories effectively. A key element in traditional film editing is the use of different cutting techniques, in order to transition seamlessly from one sequence to another. A fundamental aspect of these techniques is the placement and control over the camera. However, VR content creators do not have full control of the camera. Instead, users in VR can freely explore the 360° of the scene around them, which potentially leads to very different experiences. While this is desirable in certain applications such as VR games, it may hinder the experience in narrative VR. In this work, we perform a systematic analysis of users' viewing behavior across cut boundaries while watching professionally edited, narrative 360° videos. We extend previous metrics for quantifying user behavior in order to support more complex and realistic footage, and we introduce two new metrics that allow us to measure users' exploration in a variety of different complex scenarios. From this analysis, (i) we confirm that previous insights derived for simple content hold for professionally edited content, and (ii) we derive new insights that could potentially influence VR content creation, informing creators about the impact of different cuts in the audience's behavior.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality;

1 INTRODUCTION

Virtual Reality (VR) offers a new medium to tell stories, with unprecedented immersion capabilities. With the recent technical developments in immersive video technologies (such as better and more affordable capture devices), 360° video is becoming an increasingly popular format for storytelling. However, little is known about user behavior and expectations in this new environment; while traditional cinematography has been in constant development for over a century, leading to a well established cinematographic language, storytelling in VR is still undergoing an initial process of experimentation, in which both content creators and researchers are trying to create a new narrative language that can be effective and leverages all the potential of the medium.

In traditional cinema the director chooses how to compose the different shots and edits, and which parts of the scene are going to be shown to the viewer. However, in VR viewers can fully and freely explore the 360° of the scene that surrounds them, so they may not follow the filmmaker's intentions. As a result, key narrative aspects may not be perceived. Directing and retaining users' attention to ensure that all important parts of the narrative are being watched is a difficult task. Some attempts have been made to ensure they observe the important areas of the scene at the right time, but they are usually overly intrusive. Common approaches may grey-out uninteresting areas (interfering with the viewer's free immersion [4]), or trigger events only when the user is watching (interfering with the flow of the narrative). Moreover, the process of creating situational continuity across movie cuts (*continuity editing*) differs from traditional cinema. When applying continuity editing techniques, scenes are edited in such a way that suggests to the viewer a sequence of events that have a narrative continuity [2]. In VR, this may be hindered by the additional freedom of users to create their own individual experience by controlling the interaction with the camera in 360º.

Given the rapid democratization of VR, it is crucial to understand how editing techniques in VR affect the audience's ability to follow a given narrative. In order to understand how viewers consume VR films and whether continuity editing is still possible in VR, previous works have focused on analyzing users' behavior in narrative 360° footage. Recently, Serrano et al. [32] showed that continuity editing techniques seem to hold in VR scenarios, and proposed the first attempt at a systematic analysis of viewing behavior across movie cuts and perceived continuity in narrative VR content. Although a valuable contribution, their stimuli consists of simple videos specifically designed for their intended analysis, limiting the scope and applicability of their results. Such videos lack the complexity of real footage edited by professional filmmakers, and were not designed for storytelling.

In our work, we address these issues and propose the first analysis

*e-mail: maranes@unizar.es
†e-mail: diegog@unizar.es
‡e-mail: anase@unizar.es

based on a large-scale collection of user behavioral data watching professionally edited 360° video content distributed in commercial platforms (such as GearVR, Oculus Rift, or Daydream). We investigate user behavior across movie cut boundaries for 18 scenes from the VR movie *The People's House*, created by *Felix & Paul Studios*[1] in 2017 (an example frame is shown in Fig. 1). In the film, the former US president Barack Obama and his wife Michelle Obama guide the audience in a tour of the White House. These scenes are significantly more complex than those of Serrano et al. since they have been created for storytelling by professional filmmakers, and users watched the scenes outside laboratory conditions (in an unconstrained manner, and without any particular task). Specifically, our contributions are as follows:

- We analyze complex scenes edited by professional filmmakers, which have been viewed by 3259 users in an unconstrained manner, outside laboratory conditions. This is two orders of magnitude larger than previous studies. We carry out an in-depth analysis to validate previous findings under significantly more complex scenarios.
- We extend the type of scenes used in previous works and include in our analysis scenes without a clear region of interest (ROI). We propose to use the Inter-Observer Congruency [13] for identifying different types of scenes and cuts. Identifying the different nature of scenes (designed with strong ROIs, or just exploratory scenes, without any explicit ROI) allows us to measure behavior in complex artistic footage rather than simple hand-crafted content [32].
- We adapt a number of user behavior metrics introduced in previous work to measure attention without the explicit need of eye-tracking data. This can be useful for future research since eye-tracking is not usually available and head orientation is easier to gather, even for commercial applications. Additionally, we introduce two metrics for measuring users' behavior in exploratory scenes.
- Based on our analysis, we derive insights with potential implications to 360° cinematic content creation.

We believe that our work is the first to attempt a precise analysis of professionally edited 360° movies based on a large-scale samples of users' behavioral data. Our analyses reveal some findings that can be relevant for VR content creators and editors. For instance, how the nature of the scene previous to a cut influences the user's behavior after such cut. We have found patterns in users' behavior for different types of cuts which in turn may help to identify them automatically, without the need to manually label scenes. In addition to this, we give quantitative measurements of users' exploration of a scene.

2 Related work

Attention in VR. Static omnidirectional panoramas are one of the most widespread distribution formats for 360° content. Several works have attempted to analyze [17,26] and predict [1,10,21,39,40] user behavior and gaze scanpaths in such content. In the absence of eye-tracked gaze data, which is not typically accessible in VR devices, other works have analyzed head movements instead, showing that head orientation is a valuable proxy of gaze behavior [34, 38]. We leverage these works and use head orientation for analyzing our videos, which allows to gather large amounts of user data without relying on explicit eye-tracking information.

The way in which users explore 360° video may be fundamentally different from the way they explore static imagery. Therefore, recent research has focused on predicting saliency and visual scanpaths for 360° videos [3, 41]. To further support this line of research, several tools have been proposed for improving user behavior visualization [29, 31]. Gutierrez et al. [8] propose a VR platform to evaluate and compare the performance of different saliency and prediction models of user behavior, both for static and dynamic content. Rossi et al. [27] propose a graph-based method for classifying users who pay attention to the same regions of the scene for a long time. One of the main implications in VR video content is that, since the field of view is limited and users do not explore the elements of the scene with the same probability, users may miss the most relevant elements of the video. To address this issue, Tang et al. [37] propose a joint video stabilization and redirection, in which smooth camera motion is introduced in order to reorient important events to the front view of the user. Pavel et al. [24] propose a reorientation technique so that viewers can choose to quickly reorient to the main area of interest of the film and not loose detail about the narrative. Stebbins et al. [36] proposed a technique that automatically rotates the virtual world in seated conditions to help redirect the viewer's physical rotation while viewing immersive narrative experiences and reducing the amount of necessary physical turning.

There are several works that focus on investigating different techniques for directing users' attention. Speicher et al. [35] gather insights on several visual guidance methods, in particular diegetic (including cues that are part of the narrative), and non-diegetic (including external cues, such as blurring the unimportant regions of the scene). Nielsen et al. [22] analyze the effect of directing users' attention by encouraging them to follow a firefly with their gaze, aiming for a less intrusive method than using forced rotation. Rothe et al. [28] study the effect of using flickering in cinematic VR to guide users' attention, concluding that it improves the recall of details but decreases the enjoyment. The use of lights, movements and sounds has been also studied for directing users' attention, concluding that diegetic cues are useful for guiding the attention of viewers in cinematic VR [30].

Datasets. To support research in these directions, during the last years several works have introduced datasets consisting of recorded data from users watching 360° footage [5, 7, 12, 15, 25]. Unfortunately, all these datasets are limited to the order of 50 users per video, and typically under controlled laboratory conditions. In contrast, the scenes analyzed in our work have been visualized by 3259 users, which allow us to extend the generality of our insights.

Cinematic VR. The filmmaking process has evolved over the years. Since the first cinematographic productions, different editing techniques have been developed and refined in order to contribute to the creation of a robust cinematographic language (for a compilation and analysis of different techniques we refer to the reader to the work of Henderson [9]). In order to bridge the gap between traditional and 360° cinematography, Mateer [18] discusses how different existing film directing techniques could be applied in this new emerging field. Focusing on the field of cinematography for 360° content, Knorr et al. [12] and Fearghail et al. [23] analyze users' behavior while watching 360° movies by taking into account the intended viewing orientation designed by the director, to verify whether users follow this intended path when watching the film. Fearghail et al. [6] use the same dataset to identify the scene elements attracting users' attention. They analyze how cuts affect storytelling and derive some insights about how to redirect users' attention to the main action of the scene, such as using graphical elements and matching the orientation of interesting parts of the story across transitions. Closer to our approach is the work of Serrano et al. [32], who introduce a set of metrics for quantifying user behavior in the context of continuity editing in VR content. Despite their valuable insights, the authors use very simple video scenes, specifically recorded for their work. These videos lack the complexity of real footage edited by professional filmmakers, and thus they are not designed for actual storytelling. In contrast, we propose a systematic analysis of user

[1] https://www.felixandpaul.com/

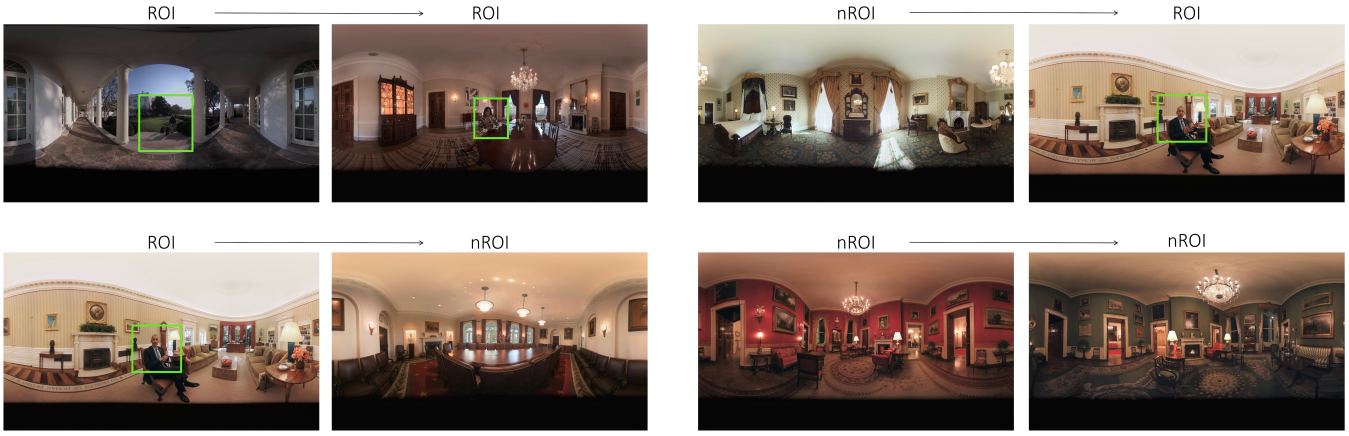

Figure 2: Examples of each cut type in the form $S_b \rightarrow S_a$. For each cut, we include a frame of the scene previous to the cut and a frame of the scene after the cut. Scenes with no explicit ROI elicit to a exploratory behavior while scenes with a defined ROI make user fixate on it.

behavior in story-based, professionally edited narrative VR, and derive insights with potentially direct implications in 360° cinematic content creation.

3 Data Set

Our data has been gathered from 3259 users watching the VR movie *The People's House*, created by *Felix & Paul Studios*. In the film, the former US president Barack Obama and his wife Michelle Obama guide the audience in a tour of the White House (Fig. 2 shows some representative frames[2]). Data was anonymously collected *in the wild*, from users watching the movie at their own personal setups, therefore, they could be either sitting or standing. There were no particular goals or instructions, other than viewing the provided content. Our data contains users watching the movie in two different systems (computer and mobile), and four different devices (Oculus Rift, Oculus Go, Gear VR, and Daydream). Demographic information was not collected. Every ten frames, we record, for each user, the 2D position in the panorama (UV coordinates) that such user is visualizing. This position is obtained by intersecting the forward camera vector, which can be directly obtained from the HMD tracking system, with the spherical geometry in which the panorama is projected for 360 visualization. Then, this per-user information is aggregated for each scene, which facilitates the analysis. In order to be able to consistently analyze user behavior along time, we only take into account those users who fully watched all the scenes analyzed. Note that since the experience contains recorded content, only three degrees-of-freedom (head rotation, and not translation) are supported. Since the movie is played at 60 *frames per second*, these head orientation points are then interpolated to obtain per-frame information.

Given that our data only consists of head orientations, we do not have any information about the actual eye gaze of users. We leverage the existing strong correlation between head movement and gaze behavior in order to obtain an estimation of fixations. In particular, it has been recently shown that eye fixations usually occur with low longitudinal head velocities (under 19.6°/s) [34], while saccadic movements between fixations correspond to higher head velocities, therefore we use this threshold to estimate fixations in our data. Then, we create saliency maps by blurring these estimated fixations with a Gaussian kernel of 11.7° of visual angle, to take into account the mean eye offset while fixating [34], since gaze points are likely to fall within this region for a given head orientation. We use this information for our analysis in Section 4.

Similar to Serrano et al.'s work [32], we manually label the region of interest (ROI) in each scene as the area in the 360° frame in which the action takes place, usually in the form of Barack or Michelle Obama speaking to the user, or an item that stands out from the scene, such as the *White House* (example frames of all labeled ROIs can be found in the supplementary material). According to this, scenes with a ROI are tagged as *ROI* scenes. Scenes without an obvious ROI are tagged as *nROI* scenes, usually consisting of outdoor or indoor scenes which the user explores freely, without any main character addressing the viewer. Since we are estimating gaze from head orientations, we follow a conservative methodology and consider that a user is fixating on a given ROI if it overlaps with the estimated saliency map.

Our goal is to analyze user behavior across movie cut boundaries; we thus introduce two variables to classify the scenes: $S_b = \{ROI, nROI\}$ for the scene before the cut, and $S_a = \{ROI, nROI\}$ for the scene after the cut. Each cut can thus be expressed as $S_b \rightarrow S_a$, which yields four possible types; see Fig. 2 for example frames before (S_b) and after (S_a) the cut. Following the previous methodology introduced by Serrano et al., we consider for our analysis the six seconds previous to the cut boundary, and the six seconds after such boundary. The movie has a total of 27 cuts, from which we select the 18 most representative of our proposed parametrization, distributed as follows: $ROI \rightarrow ROI$: five cuts, $nROI \rightarrow ROI$: four cuts, $nROI \rightarrow nROI$: five cuts, $ROI \rightarrow nROI$: four cuts.

4 Analyzing the Influence of Cuts

We first compute saliency maps for each scene by taking into account estimated fixations and the mean eye offset while fixating, as described in Section 3 (Fig. 4 shows an example). In the rest of this section, we first analyze users' viewing congruency for our two types of scenes $\{ROI, nROI\}$, then we analyze users' behavior across cuts for all four possible combinations $S_b \rightarrow S_a$.

4.1 Analyzing users' congruency

From the saliency information, we first analyze the consistency between users' viewing behavior by computing the *Inter-Observer Congruency* (IOC) [14]. In order to understand the influence of the cut in users' viewing consistency, we compute this metric considering six seconds after the cut boundary [32], for each of our cuts. Following previous work, we use a leave-one-out-approach: we leave out the i_{th} subject and aggregate all other users' fixations by accumulating one-second windows; then we compute the percentage of fixations of the i_{th} user that fall within the $k\%$ most salient regions

[2]Frames displayed for illustrative purposes with permission from the creators.

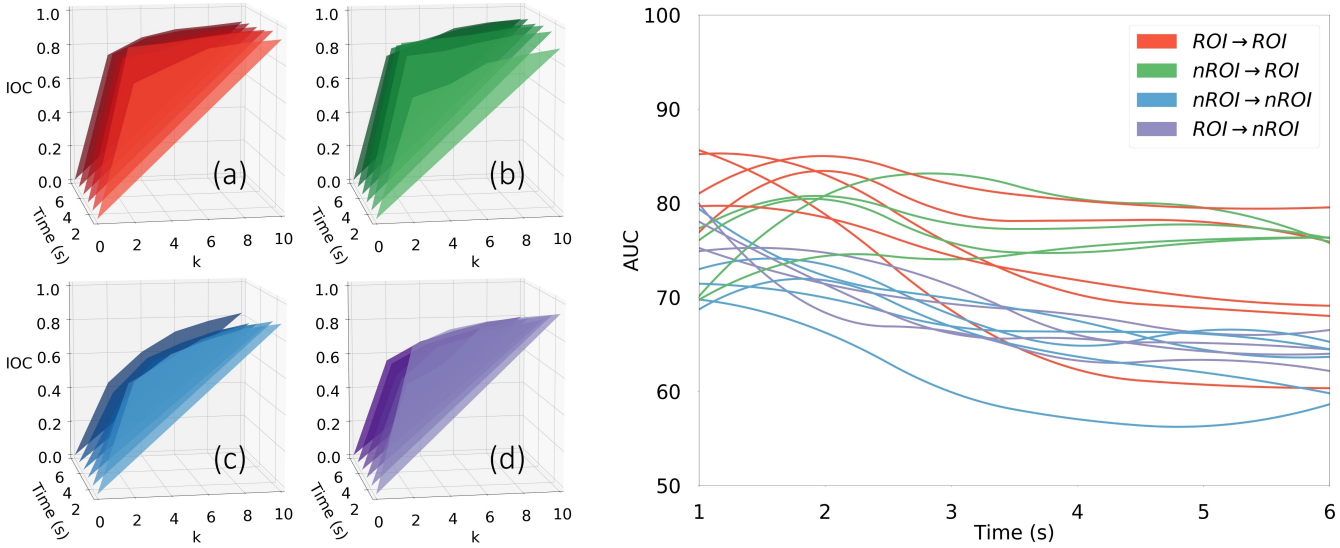

Figure 3: **Left**: IOC (*Inter-Observer Congruency*) computed for each of the four types of cuts. It shows the evolution of the IOC value along time, by varying the k percent of most salient regions of the scene. Intuitively, this metric can be seen as an estimation of users' agreement on the most salient regions of the scene. A high IOC value (1 being perfect agreement) indicates similar viewing behavior across users. **(a)**: IOC for a $ROI \rightarrow ROI$ cut. This configuration presents a high and constant IOC due to the single ROI present in the scene before and after the cut. **(b)**: IOC for a $nROI \rightarrow ROI$ cut. During the first few seconds there is a low IOC since users were scattered across the scene before the cut; once users have fixated on the new ROI, the IOC increases. **(c)**: IOC for a $nROI \rightarrow nROI$ cut. This configuration shows a low and constant IOC because users are scattered across the scene due to the lack of a ROI. **(d)**: IOC for a $ROI \rightarrow nROI$ cut. During the first few seconds there is a high IOC because users come from a ROI scene before the cut and are concentrated in the same regions, however, when the scene after cut starts, users start scattering due to the lack of ROI after the cut, and the IOC decreases. **Right**: Temporal evolution of the Area Under the Curve (AUC) computed from the IOC curves for each analyzed cut. Changes in the IOC value along time are easily seen using the AUC.

Figure 4: Example saliency map of a 360° frame.

predicted by the aggregated saliency map, varying $k \in [0\%..10\%]$ in 2.0 increments. We repeat this process for all users and compute the mean value, which is an indicator of users' viewing behavior agreement for a given cut. Intuitively, the IOC gives an estimation of how well other users' data approximate the behavior of the i_{th} user. A high value of this metric indicates that most of the users are viewing the same region of the scene, while a low value indicates that users are scattered all over the scene watching different regions of it. Fig. 3 (left) illustrates results for a cut of each type: $ROI \rightarrow ROI$, $nROI \rightarrow ROI$, $nROI \rightarrow nROI$, and $ROI \rightarrow nROI$ (IOC curves for all analyzed cuts can be found in the supplementary material). In the presence of a ROI in the scene after the cut, the IOC rapidly approaches a high value. This is clearly not the case for *nROI* scenes after the cut. This is a consequence of the free-exploration behavior that a *nROI* scene after the cut elicits. We can also observe this behavior in the saliency map of the scene (see Fig. 5 for an example), where *ROI* scenes after the cut retain users' attention, while *nROI* scenes lead to a more exploratory behavior. To provide an easier interpretation of the evolution of the IOC along time, we have additionally computed the Area Under the Curve (AUC) comprised under the previously computed IOC curves. Following the interpretation of the IOC curves, the AUC takes values between 0 (no congruency between users) and 100 (total congruency). The resulting curves are shown in Fig. 3 (right). It can be seen how *nROI* scenes after the cut display lower AUC values than *ROI* scenes, specially in the first few seconds after the cut, indicating that post-cut *nROI* scenes clearly affect users' congruency. This is to be expected, since there is no explicit ROI that draws users' attention, exploration patterns differ in a more pronounced way.

One of the drawbacks of using this metric for analyzing the temporal evolution of users' congruency is that computing saliency maps using a leave-one-out-approach for every user is very costly for large datasets like ours. We have observed that similar results can be obtained by relying on the entropy of the saliency maps. We first compute the entropy as the Normalized Shannon Entropy [34] as follows:

$$H(t) = -\sum_{i=1}^{N} \frac{S_{\Delta t} \log(S_{\Delta t})}{\log(N)} \quad (1)$$

where N is the number of pixels of the panorama, and $S_{\Delta t}$ corresponds to the saliency map computed by aggregating all users' saliency in a temporal window $\Delta t = 1$ *second*. For facilitating interpretation, we compute the *reverse entropy* such that $Hr(t) = 1 - H(t)$ (see Fig. 7), in order to match the behavior of this curve to the AUC curves. A low reverse entropy indicates that there are a large number of similarly salient objects distributed throughout the scene, causing users' fixations to be scattered all over the scene; a high reverse entropy results from a few salient objects that capture all the viewer's attention. From these curves, we can observe several interesting behaviors. First, we can confirm that this metric also supports the

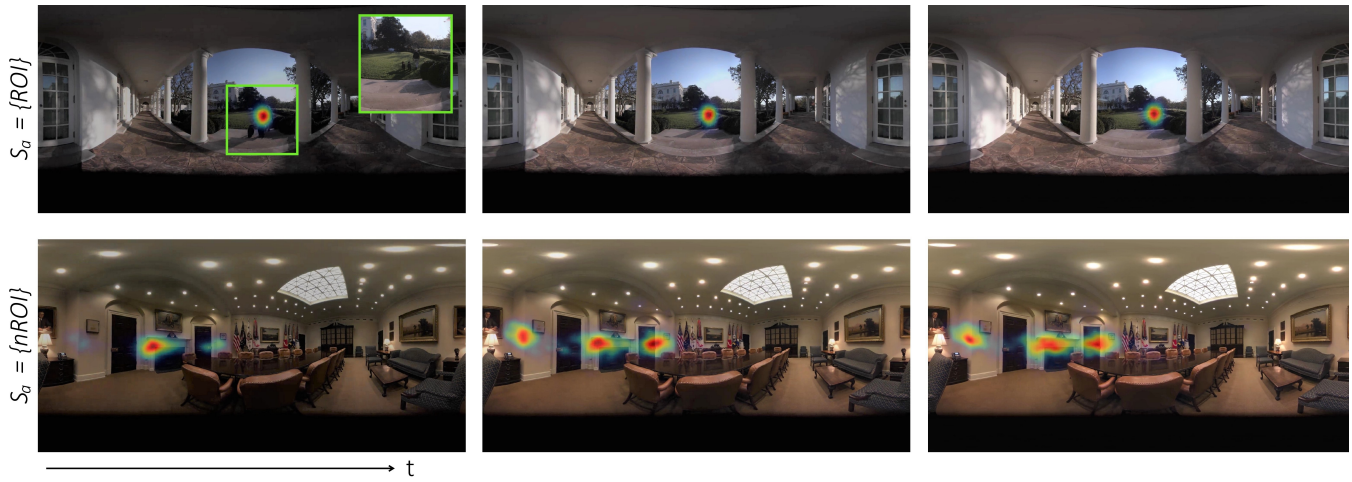

Figure 5: Evolution of users' gaze through a *ROI* and *nROI* scene, represented as a saliency map. **Top**: The ROI, a person playing with a dogs, is able to retain the users attention along the scene. **Bottom**: Since there is no explicit ROI, users explore the different parts of the room.

Figure 6: Example of users' behavior (head orientation) for a scene with a static ROI (a speaker's podium). **Top**: The ROI quickly attracts users' attention at the beginning of the scene. **Bottom**: After a few seconds, gaze scatters leading to low congruency between users.

insights derived from the AUC curves. Second, we have included in our dataset several scenes where the ROI is not a human speaking to the user, but a salient static feature in the form of a speaker's *podium* or the *White House*. In these cases, we have noticed that the ROI catches users' attention at the beginning, displaying a high reverse entropy, however, reverse entropy drops significantly after a few seconds, indicating that users' attention gets scattered, as the corresponding $ROI \rightarrow ROI$ curves in Fig. 7 shows. Fig. 6 shows two representative frames of the *podium* scene, illustrating this effect. We can further distinguish between two types of static ROIs, those which contain a considerable amount of details to be explored by users (Fig. 7, *White House* inset), and those which can be considered as containing less interesting features (Fig. 7, *podium* inset). In order to quantify the attention of users to the ROI over time we propose the following metric, which we term *Attention Retention*, computed as the normalized integral of the $Hr(t)$ curve along time:

$$Attention\ Retention = \frac{\int_1^T Hr(t)dt}{T-1} \quad (2)$$

The *Attention Retention* metric takes values between 0 (the ROI does not consistently catch all users' attention) and 1 (the ROI retains all users' attention). We have computed the 6 first seconds after the cut ($T = 6$) for those cuts which meet $S_a = ROI$ (Fig. 7 bottom). Our results suggest that static ROIs with fewer details catch briefly the users' attention, while ROIs with more interesting features are able to retain users' attention for longer times. Finally, some of our ROIs include dynamic content, i.e., they featured moving actors relevant to the narrative, instead of static important elements of the scene. An example of this is shown in Fig. 7 (top), where we can see Obama talking to the audience. We can observe that these dynamic ROIs are able to retain users' attention more consistently along time.

4.2 Analyzing users' behavior

In order to analyze users' behavior in a quantitative manner, we make use of five different metrics:

Frames to reach a ROI (framesToROI). Number of frames after the cut before the user fixates on a ROI. Intuitively, this metric is an estimation on how long it takes users to converge to the main action after a cut.

Percentage of total fixations inside the ROI (percFixInside). This metric computes the percentage of fixations inside the ROI with respect to the total amount of fixations (inside or outside the ROI) after the user finds the ROI after the cut (i.e., it is independent of *framesToROI*). It gives an estimation of the viewer's interest in the ROI.

These two metrics were introduced by Serrano et al. [32], and are limited to scenes in which a ROI can be explicitly defined (*ROI* scenes in our terminology); therefore they can only analyze cuts in the form $\{ROI, nROI\} \rightarrow ROI$. To solve this problem, we have adapted another metric proposed by Serrano et al. (*nFix*) and we have introduced two additional novel metrics (*traveledDistance* and *percSceneWatched*) that capture the user behavior for cuts in which

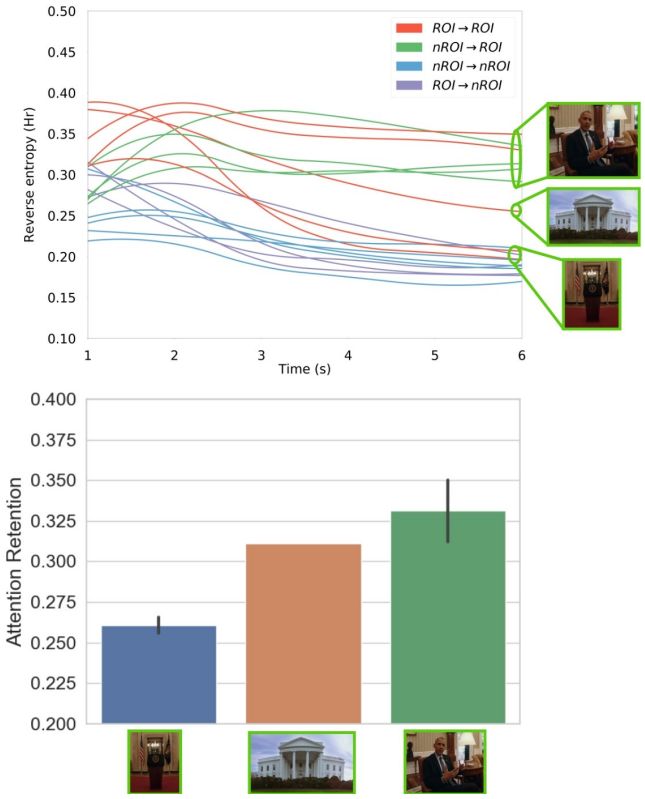

Figure 7: **Top**: Reverse entropy for each of our analyzed cuts, computed during six seconds after the cut. The behavior of this metric is similar to the AUC metric: when there is a strong ROI present in the scene, the metric value increases indicating that most users' fixate on the ROI; when there is not an explicit ROI, users gaze is scattered across the scene, resulting in a low metric value. Insets show representative frames of different types of ROIs which meet $S_a = ROI$. **Bottom**: Mean of the Attention Retention metric computed for cuts that meet $S_a = ROI$. Static ROIs with few interesting details fail to retain users' attention, while ROIs presenting more interesting details are able to retain users' attention longer. ROIs that catch attention for a more prolonged time are those which are dynamic and contribute to the film narrative. Error bars correspond to a 95% confidence interval.

the scene after the cut does not have an explicit ROI ($\{ROI, nROI\} \rightarrow nROI$):

Number of fixations (nFix). In this metric we compute the ratio between the number of samples that correspond to estimated fixations, and the total number of samples after the cut (as opposed to Serrano et al. that only considered the time after the user had fixated on the ROI, we consider all the six seconds after the cut). A higher value indicates that the user has been fixating most of the time while a low value suggests a more exploratory behavior, meaning that the user has been performing saccades.

Total distance traveled (traveledDistance). This metric measures the accumulated orthodromic distance (or great-circle distance) traveled after the cut (refer to Appendix A). It is an indicative of how much users have roamed through the scene, and it is independent of the number of fixations, since it is computed taking into account all samples.

Percentage of the scene watched (percSceneWatched). This metric computes the percentage of the 360° environment watched. Note that a region of the scene is considered watched if the user has fixated on it. A high value indicates that a user has been observing different parts of the scene, and thus can be used as an indicative of how much of the scene content the user has actually explored.

Since we can not assume that our observations are independent, and our data does not follow a normal distribution for any of our metrics ($p < 0.001$ for the Smirnov-Kolmogorov normality test for large data samples) we employ a Generalized Linear Mixed Model in which we model the influence of each particular user as a random effect. Then, for each of our metrics, we choose the distribution that better fits our data (poisson distribution for *framesToROI*, gamma distribution for *percFixInside*, *nFix*, *traveledDistance* and *percSceneWatched*). Since we have categorical variables among our predictors, we re-code them to dummy variables for the regression. We establish our significance level to $p = 0.01$ for all our analyses. Note that for our first two metrics (*framesToROI* and *percFixInside*) we can not analyze the influence of S_a since they can not handle scenes without an explicit ROI ($S_a = nROI$ scenes). In the rest of the section we describe the main findings of our analyses by taking into account 6 seconds of the scene after the cut [32].

Influence of S_b in framesToROI. For analyzing the dependent variable *framesToROI* we include in the regression our factor S_b as a predictor. Our analysis suggests that it takes users significantly longer ($p < 0.001$) to reach the ROI in the scene after the cut when the scene before the cut does not have an explicit ROI ($S_b = nROI$). This can be seen in Fig. 8a. We hypothesize that, since the scene before the cut does not have an explicit ROI (i.e., it is an scene intended for exploration), users are scattered looking at different locations of the scene when the cut occurs, and therefore it takes them longer to converge to the main action after the cut. This behavior seems consistent with the findings of Serrano et al., in which users took longer to find the ROI when ROIs were misaligned before and after the cut. The notion of aligning ROIs for facilitating transitions across cuts is also in line with previous research on editing in cinematic VR [11], and the technique of *match on attention* advocated by practitioners such Jessica Brillhart [3].

Influence of S_b in percFixInside. For analyzing the dependent variable *percFixInside* we include in the regression our factor S_b as a predictor. Interestingly, there is a significantly ($p < 0.001$) larger percentage of fixations inside the ROI (after the ROI is found after the cut) when the scene before the cut does not have an explicit ROI ($S_b = nROI$). This can be seen in Fig. 8b. This effect is not highly pronounced, however the mean difference is enough to suggest that introducing a scene of a more exploratory nature (*nROI*) before the cut leads to an increasing interest in the ROI after the cut, while a ROI before the cut elicits a more exploratory behavior after it.

Influence of S_b and S_a in nFix, traveledDistance *and* percSceneWatched. In order to analyze these dependent variables, we have included in the regression the factors S_a and S_b and their interaction ($S_b * S_a$) as predictors. The first effect we notice is that, as expected, a scene without an explicit ROI after the cut ($S_a = nROI$) increases the traveled distance (Fig. 8e) and the percentage of scene watched (Fig. 8g), and decreases the number of fixations (Fig. 8c). This is to be expected, since users do not have a clear area of interest to fixate into. We have found that a scene with a ROI before the cut ($S_b = ROI$) seems to elicit a more exploratory behavior after the cut: users perform less fixations (Fig. 8d) and roam more (Fig. 8f). However, this does not have a strong effect in the percentage of scene watched (Fig. 8h). This is in accordance with the previous two metrics: even though users roam more through the scene, they perform less fixations, so they do not necessarily fixate on more regions of the scene. We have also found a significant effect of the interaction $S_b * S_a$: When the scene after the cut does not have an explicit ROI ($S_a = nROI$) it will elicit an exploratory behavior regardless of the type of scene before the cut (S_b). However, when

[3] `https://medium.com/the-language-of-vr`

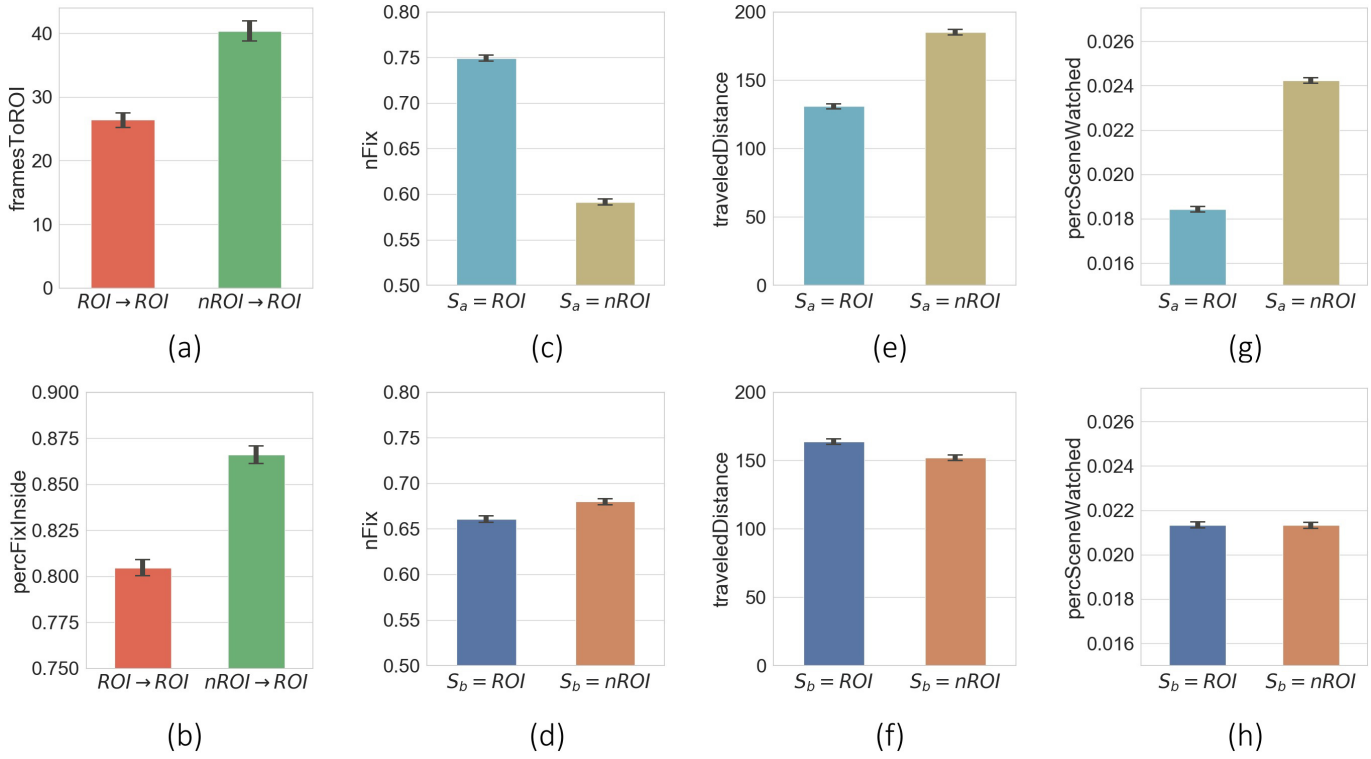

Figure 8: **(a)**: Mean *framesToROI* for the cuts $\{ROI, nROI\} \rightarrow ROI$. Users require more time to fixate when they have been exploring in the previous scene. **(b)**: Mean *percFixInside* for the cuts $\{ROI, nROI\} \rightarrow ROI$. The ROI in the scene after the cut captures the attention of the users more strongly if there is no ROI before the cut. **(c)**: Mean *nFix* grouped by type of scene after the cut (S_a). Users perform more fixations when there is a ROI in the scene. **(d)**: Mean *nFix* grouped by type of scene before the cut (S_b). Users perform more fixations if there is a *nROI* scene before the cut. **(e)**: Mean *traveledDistance* grouped by type of scene after the cut (S_a). As expected, users roam more if the scene after the cut has no ROI. **(f)**: Mean *traveledDistance* grouped by type of scene before the cut (S_b). Users roam more if they come from a scene in which there is a ROI. **(g)**: Mean *percSceneWatched* grouped by type of scene after the cut (S_a). When there is no explicit ROI in the scene, users observe more different parts of the scene than if there is a ROI on it, because it catches the users' attention limiting their exploration. **(h)**: Mean *percSceneWatched* grouped by type of scene before the cut (S_b). Users approximately explore the same proportion of the scene regardless of the scene before the cut. Error bars correspond to a 95% confidence interval.

the scene after the cut has a clear ROI ($S_a = ROI$), the type of scene before the cut (S_b) becomes relevant: in these cases our results suggest that it is more effective to place an exploratory scene before the cut ($S_b = nROI$) in order to retain users' attention in a single ROI after the cut. This insight is supported by our three metrics, users perform more fixations (Fig. 9a), roam less (Fig. 9b) and explore less in different parts of the scenes (Fig. 9c) when the scene before the cut corresponds to an exploratory scene ($Sb = nROI$) and the scene after the cut is a ROI scene ($S_a = ROI$).

Additional analyses. Our data has been collected from users watching the footage in two different systems: computer (Oculus Rift), and mobile (Oculus Go, Gear VR, and Daydream). We have repeated our analysis by including the *system* as a factor (see Section E in the supplementary material). Even though there are significant differences across devices, all the insights described in this section hold. A detailed analysis on the influence of the device has been proposed by previous works [16, 33], and it is outside the scope of this paper. Additionally, there are some brief moments during the footage in which the narrator suggests viewers to watch a part of the scene through gestures or auditory cues. To ensure the robustness of our results, we have also repeated our analysis eliminating all cuts exhibiting any minimal gestural or auditory cue that may alter user behavior, resulting in a subset of 11 cuts. We provide this analysis in Section D in the supplementary material, showing that our insights hold.

5 Discussions and Conclusions

To our knowledge, our work is the first to attempt a systematic analysis of professionally edited, narrative 360° video. Our analyses are based on a large-scale collection of 3259 users' behavioral data, which is orders of magnitude larger than existing datasets. Analyzing professionally edited videos is very challenging due to two main reasons: (i) this footage is hard to parametrize since it was not created for research purposes; and (ii) users have watched the videos in an unconstrained manner, outside of laboratory conditions, and without any particular task or instructions. We have adapted existing metrics for quantifying user behavior under more complex and realistic footage, and without the need of eye-tracking. Additionally, we have introduced two new metrics that allow us to measure the degree of users' exploration without the explicit need of defining potential regions of interest. Finally, we have also shown how both the Inter-Observer Congruency (IOC), and the entropy of the saliency map could be leveraged to classify scenes and cuts automatically, without the need for manual labeling, and we have explored the possibility of using these metrics to measure how much a ROI can retain users' attention over time, showing that ROIs that engage with the user (such as a character narrating the story) retain more attention than static ROIs.

Our results are consistent with previous works on simpler scenes, suggesting that certain behaviors are shared across users regardless of the complexity of the content and the conditions in which this

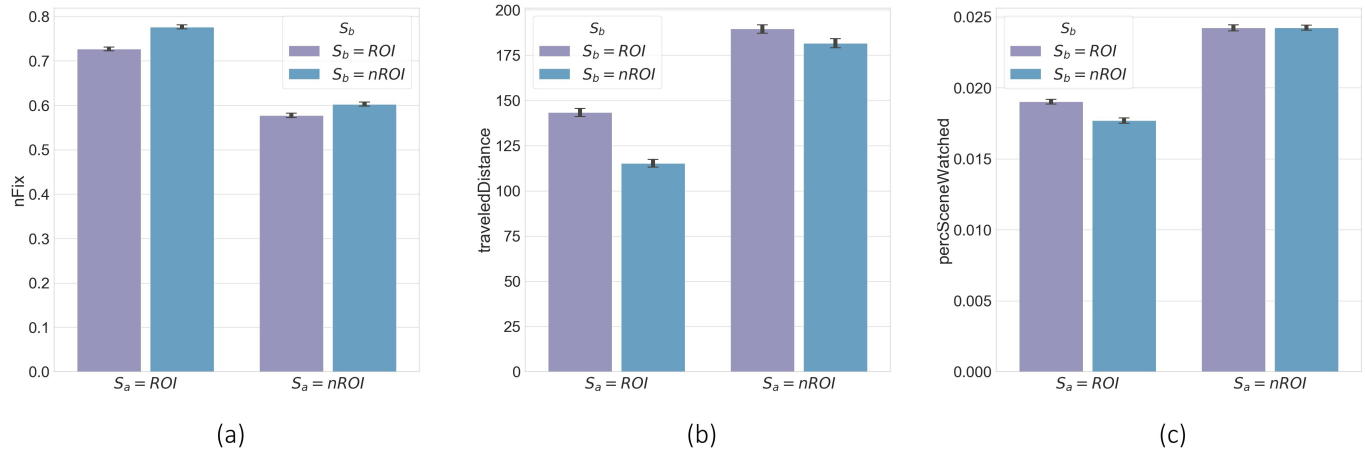

Figure 9: **(a)**: Mean *nFix* metric for each type of cut. Users perform slightly more fixations in a *ROI* scene after the cut ($S_a = ROI$) if there is no explicit ROI before the cut ($S_b = nROI$). **(b)**: Mean *traveledDistance* metric for each type of cut. When the scene after the cut has a ROI on it ($S_a = ROI$), users roam less if the previous scene encourages exploration ($S_b = nROI$). **(c)**: Mean *percSceneWatched* for each type of cut. For scenes with a ROI after the cut ($S_a = ROI$), a lower percentage of the scene is explored if there was no single ROI previous to the cut ($S_b = nROI$). Error bars correspond to a 95% confidence interval.

content is visualized. For example, Serrano et al. [32] analyzed the influence of the misalignment of ROIs between cuts, concluding that users require some time to adapt to the new visual content once a misalignment has been introduced. Our analyses seem to point in the same direction: when the scene before the cut is an exploratory scene (*nROI*) users take longer to converge to the main ROI after the cut, since they are scattered and potentially misaligned with respect to the new ROI appearing after the cut. Quantitatively, we have observed that users require more time to reach the main ROI after the cut: this could be due to the increased complexity of our videos. In the work of Nasrabadi et al. [20], users' attention is grouped in clusters [27], concluding that in scenes where there is an interesting ROI, there are less clusters because most users are watching the ROI. We have also observed this effect in the analysis of the IOC, in which for scenes without an explicit ROI ($S_a = nROI$), users do not seem to converge to the same regions of the video. We have also observed new effects that could potentially influence VR content creation, informing creators about the impact of different cuts in the audience's behavior. For instance, if the content creator wants to better attract the user's attention when there is an important narrative element (a ROI), one option would be to design the previous scene as an exploration scene (without an explicit ROI). This scene configuration seems to make the user more focused when an interesting element appears after the cut. This configuration would match with the establishing shot technique used in traditional cinema where the scene before the cut gives a context of the environment, and then the main narrative and ROIs are resumed after the cut.

Limitations and future work. Similarly to other studies of the same nature, although we have analyzed a large-collection of user data under an unconstrained scenario, our results may not extrapolate to conditions outside of our study, since the footage and the editing techniques studied are not representative of all existing 360° movies. We have analyzed a film of documentary nature, which implies that the action is scarce: there are no conversations between characters and there is little movement on stage. It would be interesting to analyze how our insights would extrapolate to content of other nature. In this footage, fade-to-black is used for transitions between scenes. In the context of teleportation, Moghadam et al. [19] did not find significant differences between instant and fade-to-black transitions in terms of spatial awareness and sickness. However, the technique used for transitioning between scenes may have an impact in the context of narrative VR and remains to be further explored. We have analyzed user behavior for 360° footage that only supports three degrees-of-freedom (head rotations), which is the main distribution format for VR cinematic content. Computer Generated (CG) experiences that allow for six degrees-of-freedom (head rotations and translation) favor a more interactive behavior, and therefore our insights may not apply in such cases. Many other parameters could be analyzed in future work, such as more complex editing techniques, or the influence in users' attention when the ROIs are dynamically moving through the scene. Moreover, our work is targeted towards the development of a cinematographic language through the establishment of editing techniques, however, more experimental techniques leveraging VR immersive capabilities could be investigated. This could include existing techniques in related fields such as immersive theater, or narrative-based videogames (ranging from simply placing viewers at the center of the action to techniques where users' actions and decisions play a role in the story, such as polychronic narratives). This intrinsic increased interaction may incur in different viewing behaviors and engagement levels, and could be an interesting avenue for future work. We have manually classified our scenes in two types, scenes either containing a strong region of interest (*ROI*), or scenes without an explicit region of interest (*nROI*). Our analysis of the AUC and entropy curves along time seems to support this as a reasonable choice, nevertheless, other parametrizations could be possible. We have created two new metrics (*traveledDistance* and *percSceneWatched*) that can measure user exploration without the explicit need of defining ROIs. However, these metrics are limited to quantifying exploration assuming that the potential regions of interest remain relatively fixed in the field of view: for ROIs moving across the scene, high values of these metrics could be potentially due to the user following the ROI, instead of exploring. These metrics could be easily adapted to such cases by only taking into account the metric value when the user is not fixating inside the ROI. In the future, more general metrics for quantifying user behavior could be potentially explored and analyzed in order to detect consistent changes in attention patterns.

We believe that our findings are one step forward towards building a cinematographic language for VR. We hope that our work helps guiding some design decisions for content creators, and expect follow-up research to continue exploring this emerging field of narrative VR.

6 Acknowledgements

We would like to thank *Felix & Paul Studios* for invaluable discussions, and for generously providing their collected data for our analyses. We would also like to thank the anonymous reviewers for their encouraging and insightful feedback on the manuscript. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (CHAMELEON project, grant agreement No 682080), and the Spanish Ministry of Economy and Competitiveness (project TIN2016-78753-P).

References

[1] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2331–2338, 2017.

[2] D. Bordwell, K. Thompson, and J. Smith. *Film Art: An Introduction*. McGraw-Hill Education, 11 ed., 2016.

[3] H. Cheng, C. Chao, J. Dong, H. Wen, T. Liu, and M. Sun. Cube Padding for Weakly-Supervised Saliency Prediction in 360° Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] F. Danieau, A. Guillo, and R. Doré. Attention guidance for immersive video content in head-mounted displays. In *2017 IEEE Virtual Reality (VR)*, pp. 205–206, March 2017. doi: 10.1109/VR.2017.7892248

[5] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet. A Dataset of Head and Eye Movements for 360° Videos. In *Proceedings of the 9th ACM Multimedia Systems Conference*, MMSys '18, pp. 432–437. ACM, New York, NY, USA, 2018. doi: 10.1145/3204949.3208139

[6] C. O. Fearghail, C. Ozcinar, S. Knorr, and A. Smolic. Director's Cut - Analysis of Aspects of Interactive Storytelling for VR Films. In *International Conference for Interactive Digital Storytelling (ICIDS) 2018*, 2018.

[7] S. Fremerey, A. Singla, K. Meseberg, and A. Raake. AVtrack360: an open dataset and software recording people's head rotations watching 360° videos on an HMD. In *Proceedings of the 9th ACM Multimedia Systems Conference*, pp. 403–408. ACM, 2018.

[8] J. Gutiérrez, E. J. David, A. Coutrot, M. P. Da Silva, and P. Le Callet. Introducing UN Salient360! Benchmark: A platform for evaluating visual attention models for 360º contents. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–3. IEEE, 2018.

[9] B. Henderson. THE LONG TAKE. *Film Comment*, 7(2):6–11, 1971.

[10] Z. Hu, C. Zhang, S. Li, G. Wang, and D. Manocha. SGaze: A Data-Driven Eye-Head Coordination Model for Realtime Gaze Prediction. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):2002–2010, May 2019. doi: 10.1109/TVCG.2019.2899187

[11] 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*, VRST '17, pp. 4:1–4:4. ACM, New York, NY, USA, 2017. doi: 10.1145/3139131.3139166

[12] S. Knorr, C. Ozcinar, C. O. Fearghail, and A. Smolic. Director's Cut - A Combined Dataset for Visual Attention Analysis in Cinematic VR Content. In *The 15th ACM SIGGRAPH European Conference on Visual Media Production*, 2018.

[13] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, Mar 2013. doi: 10.3758/s13428-012-0226-9

[14] O. Le Meur, T. Baccino, and A. Roumy. Prediction of the Inter-Observer Visual Congruency (IOVC) and Application to Image Ranking. pp. 373–382, 11 2011. doi: 10.1145/2072298.2072347

[15] W. Lo, C. Fan, J. Lee, C. Huang, K. Chen, and C. Hsu. 360 video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 211–216. ACM, 2017.

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

[17] G. Marmitt and A. T. Duchowski. *Modeling visual attention in VR: Measuring the accuracy of predicted scanpaths*. PhD thesis, Clemson University, 2002.

[18] J. Mateer. Directing for Cinematic Virtual Reality : how traditional film director's craft applies to immersive environments and notions of presence. *Journal of Media Practice*, 18, 05 2017. doi: 10.1080/14682753.2017.1305838

[19] K. Moghadam, C. Banigan, and E. Ragan. Scene Transitions and Teleportation in Virtual Reality and the Implications for Spatial Awareness and Sickness. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 11 2018. doi: 10.1109/TVCG.2018.2884468

[20] A. T. Nasrabadi, A. Samiei, A. Mahzari, R. P. McMahan, R. Prakash, M. C. Farias, and M. M. Carvalho. A Taxonomy and Dataset for 360° Videos. In *Proceedings of the 10th ACM Multimedia Systems Conference*, pp. 273–278. ACM, 2019.

[21] A. Nguyen, Z. Yan, and K. Nahrstedt. Your Attention is Unique: Detecting 360-Degree Video Saliency in Head-Mounted Display for Head Movement Prediction. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pp. 1190–1198. ACM, New York, NY, USA, 2018. doi: 10.1145/3240508.3240669

[22] 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

[23] C. O Fearghail, C. Ozcinar, S. Knorr, and A. Smolic. Director's Cut - Analysis of Aspects of Interactive Storytelling for VR Films. 12 2018.

[24] A. Pavel, B. Hartmann, and M. Agrawala. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*, UIST '17, pp. 289–297. ACM, New York, NY, USA, 2017. doi: 10.1145/3126594.3126636

[25] Y. Rai, J. Gutiérrez, and P. Le Callet. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 205–210. ACM, 2017.

[26] Y. Rai, P. Le Callet, and P. Guillotel. Which saliency weighting for omni directional image quality assessment? In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6. IEEE, 2017.

[27] S. Rossi, F. De Simone, P. Frossard, and L. Toni. Spherical Clustering of Users Navigating 360º Content. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4020–4024, May 2019. doi: 10.1109/ICASSP.2019.8683854

[28] S. Rothe, F. Althammer, and M. Khamis. GazeRecall: Using Gaze Direction to Increase Recall of Details in Cinematic Virtual Reality. pp. 115–119, 11 2018. doi: 10.1145/3282894.3282903

[29] S. Rothe, T. Höllerer, and H. Hußmann. CVR-Analyzer: A Tool for Analyzing Cinematic Virtual Reality Viewing Patterns. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, MUM 2018, pp. 127–137. ACM, New York, NY, USA, 2018. doi: 10.1145/3282894.3282896

[30] S. Rothe and H. Hussmann. *Guiding the Viewer in Cinematic Virtual Reality by Diegetic Cues*, pp. 101–117. 06 2018. doi: 10.1007/978-3-319-95270-3_7

[31] S. Rothe and H. Hußmann. Spatial statistics for analyzing data in cinematic virtual reality. pp. 1–3, 05 2018. doi: 10.1145/3206505.3206561

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

[33] S. Sharples, S. Cobb, A. Moody, and J. R. Wilson. Virtual reality induced symptoms and effects (VRISE): Comparison of head mounted display (HMD), desktop and projection display systems. *Displays*, 29(2):58–69, 2008.

[34] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, 36(4), 2018.

[35] M. Speicher, C. Rosenberg, D. Degraen, F. Daiber, and A. Krúger. Exploring Visual Guidance in 360-degree Videos. In *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '19, pp. 1–12. ACM, New York, NY, USA, 2019. doi: 10.1145/3317697.3323350

[36] T. Stebbins and E. D. Ragan. Redirecting View Rotation in Immersive Movies with Washout Filters. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 377–385, March 2019. doi: 10.1109/VR.2019.8797994

[37] C. Tang, O. Wang, F. Liu, and P. Tan. Joint stabilization and direction of 360° videos. *ACM Trans. Graph.*, 38(2):18:1–18:13, Mar. 2019. doi: 10.1145/3211889

[38] E. Upenik and T. Ebrahimi. A simple method to obtain visual attention data in head mounted virtual reality. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 73–78. IEEE, 2017.

[39] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[40] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze Prediction in Dynamic 360° Immersive Videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5333–5342, June 2018. doi: 10.1109/CVPR.2018.00559

[41] Z. Zhang, Y. Xu, J. Yu, and S. Gao. Saliency Detection in 360° Videos. In *The European Conference on Computer Vision (ECCV)*, September 2018.

A COMPUTING FIXATIONS

To compute fixations giving only head orientation data in equirectangular coordinates (u and v), we first transform our coordinate space to spherical coordinates, in order to have correctly measured distances between points (instead of computing distances in image space, where they may be distorted to the equirectangular projection). In this new coordinate space, head orientation is represented as latitude and longitude. We compute this by using the Equation 3, where equirectangular coordinates are normalized (values between 0 and 1).

$$\begin{aligned} latitude &= v * \pi - \frac{\pi}{2} \\ longitude &= u * 2\pi - \pi \end{aligned} \tag{3}$$

Once we have head orientation defined in spherical coordinates, we can extract the distance between two points using the great circle distance as described in Equation 4. In this equation, Δd is the distance for a sphere with unitary radius; ϕ is the latitude, λ is the longitude and $\Delta\lambda = \lambda_q - \lambda_p$.

$$\Delta d = arccos(sin\phi_p sin\phi_q + cos\phi_p cos\phi_q cos(\Delta\lambda)) \tag{4}$$

To compute the longitudinal head velocity that is later used to determine whether users are fixating or not, we just divide Δd by the time passed. When this velocity is under 19.6°/s, as Sitzmann et al. [34] indicate, we can assume that the user is fixating. According to this work, the fixation will likely fall within the neighboring area of diameter 11.7° around the head orientation point, therefore, to create saliency maps from our estimated fixations, we take into account this area by convolving fixation points with a Gaussian corresponding to 11.7° of visual angle.