Director's Cut - A Combined Dataset for Visual Attention Analysis in Cinematic VR Content

Sebastian Knorr
Communication Systems Group, TU Berlin, Germany and
V-SENSE, School of Computer Science and Statistics,
Trinity College Dublin, The University of Dublin, Ireland
knorr@nue.tu-berlin.de

Cagri Ozcinar
V-SENSE, School of Computer Science and Statistics,
Trinity College Dublin, The University of Dublin, Ireland
ozcinarc@scss.tcd.ie

Colm O Fearghail
V-SENSE, School of Computer Science and Statistics,
Trinity College Dublin, The University of Dublin, Ireland
ofearghc@scss.tcd.ie

Aljosa Smolic
V-SENSE, School of Computer Science and Statistics,
Trinity College Dublin, The University of Dublin, Ireland
smolica@scss.tcd.ie

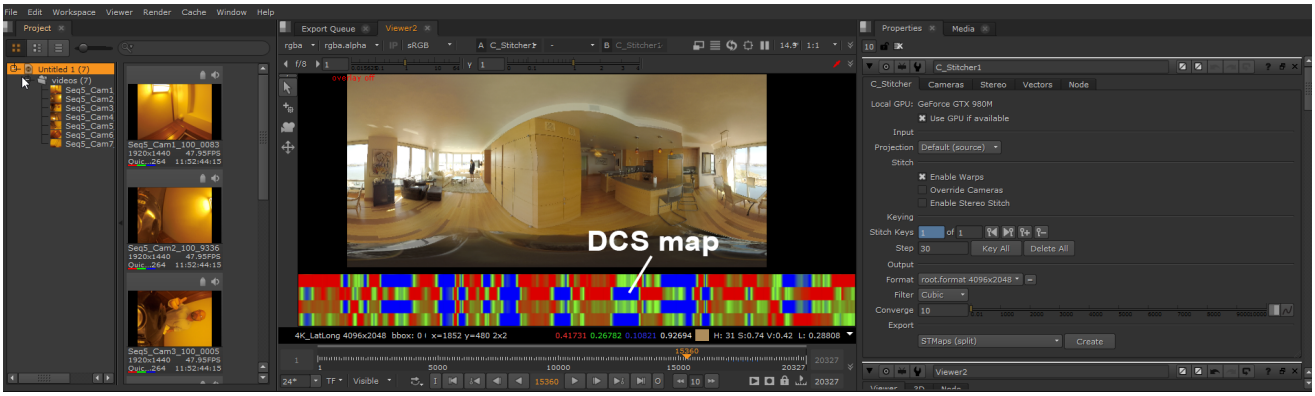


Figure 1: Example of visualization of predicted visual attention as color-coded timeline during editing in post.

ABSTRACT

Methods of storytelling in cinema have well established conventions that have been built over the course of its history and the development of the format. In 360° film many of the techniques that have formed part of this cinematic language or visual narrative are not easily applied or are not applicable due to the nature of the format i.e. not contained the border of the screen. In this paper, we analyze how end-users view 360° video in the presence of directional cues and evaluate if they are able to follow the actual story of narrative 360° films. We first let filmmakers create an intended scan-path, the so called director's cut, by setting position markers in the equirectangular representation of the omnidirectional content for eight short 360° films. Alongside this the filmmakers provided additional information regarding directional cues and plot points. Then, we performed a subjective test with 20 participants watching the films with a head-mounted display and recorded the center position of the viewports. The resulting scan-paths of the participants are then compared against the director's cut using different scan-path similarity measures. In order to better visualize the similarity between the scan-paths, we introduce a new metric which measures and visualizes the viewport overlap between the participants' scan-paths and the director's cut. Finally, the entire dataset, i.e. the director's cuts including the directional cues and plot points as well as the scan-paths of the test subjects, is publicly available with this paper.

KEYWORDS

360° video, virtual reality, visual attention, scan-path metrics, storytelling

ACM Reference Format:
Sebastian Knorr, Cagri Ozcinar, Colm O Fearghail, and Aljosa Smolic. 2018. Director's Cut - A Combined Dataset for Visual Attention Analysis in Cinematic VR Content. In *Proceedings of The 15th ACM SIGGRAPH European Conference Visual Media Production (CVMP 2018).* ACM, New York, NY, USA, 10 pages.

1 INTRODUCTION

With the increasing commercialization of virtual reality (VR) as a medium, one of the main factors that drive the uptake of devices is content. One of the most popular formats to deliver this content is 360° film, which is also called cinematic VR or live-action VR. In contrast to traditional cinema, where the viewer perceives the

CVMP 2018, Dec. 13–14, London, UK
2018.

world through a window i.e. the cinema screen, cinematic VR allows a person to be present within the world by wearing a head-mounted display (HMD) [42].

This poses a new challenge for filmmakers as it necessitates the expansion of cinematic language to a border-less format. It also poses technical challenges in the entire production, post-production and delivery chain. 360° video needs resolutions of above 8K in order to reach a similar quality as current ultra-high-definition television because only a fraction of the video, the so-called viewport, is visible for the end-user. Thus, new ways of compressing and streaming of cinematic VR content are required to get quality content to the final consumers. An example of how a director's input might optimize the bandwidth and decoding power requirement of 360° video streaming can be found in the report of the ISO/IEC working group MPEG [43] where the concept of an additionally streamed 'director's view' with higher bitrates is introduced. If the user is looking into the direction of the director's view, then the viewport pixels are rendered from this view instead of the bitstream of the full sphere, which has lower quality due to bandwidth limitations. However, visual attention modelling [8, 11, 22] and saliency prediction [30, 33, 38] are crucial in order to predict where users will eventually look. With this information in place, video streaming quality can be enhanced [31, 34, 35] or labor-intensive quality control can be automated in post-production [14]. Finally, understanding visual attention also supports the storytelling process for 360° video [4, 37, 40], e.g. through automatic prediction of visual attention of end-users directly during the editing process in post. Figure 1 shows an example with a graphical user interface where the predicted visual attention is visualized with a director's cut similarity (DCS) map (see Section 4 for details).

In this paper, we analyze how users view 360° video in the presence of directional cues and plot points, and evaluate if they can follow the actual story of narrative 360° films. We first let filmmakers create an intended scan-path, the so called director's cut (DC), by setting position markers in the equirectangular format of eight short 360° films. Alongside this, the filmmakers provided additional information regarding directional cues and plot points for their own films. Then, we performed a subjective test with 20 participants watching the films with an HMD and recorded the users' head orientation.

In order to better visualize the comparison of the viewers and the directors preferred viewing, we introduce a new metric which measures and visualizes the viewport overlap between the participants' scan-paths and the director's cut, which is one of our contributions.

Finally, the entire dataset, i.e. the director's cuts including the directional cues and plot points as well as the scan-paths of the participants, are publicly available with this paper, which is the main contribution of the paper alongside with the presented analysis[1]. To our knowledge, it is the first time that director's cuts of professional VR filmmakers combined with subjective data have been provided to the scientific community. This dataset can be seen as a first important step to contribute to streaming concepts like the one introduced by MPEG [43]. Moreover, it can further be used to develop and test saliency prediction approaches, which can then be integrated into post-production applications, and new streaming solutions by creating and utilizing the additional director's cut, which is an important contribution for the multimedia community.

The remainder of the paper is organized as follows. In Section 2, we review related work. In Section 3, we describe the proposed metrics which measure and visualize the similarities between the users' scan-paths and the director's cut. The methodology and evaluation of the subjective experiment are presented in Sections 4 and 5, respectively. In Section 6, we conclude the paper with a discussion and further outlook.

2 RELATED WORK

Storytelling. Four techniques that have traditionally formed the 'tools' that filmmakers rely on to tell their stories are: cinematography, mise-en-scene, sound and editing [46]. The expansion of these tools into VR, however, requires each to be re-evaluated as the viewer is free to look in any direction of the 360° film without the direct control of the filmmaker. Spatial audio can be an effective tool to guide the viewer to another area of the scene as are directional cues by the actors present in it, new concepts such a presence need to be also taken into account in [20].

One of the most central ideas to the notion that continuity-led film grammars [7] are also applicable to the cinematic VR is the ability of the director to predict and indirectly control the user's viewport [29]. For instance, Serrano et al. [39] investigated continuity editing in VR video in the context of segmentation theory [25]. Their findings include that continuity of action across cuts by aligning the region of interest (ROI) between them is best suited to fast-paced action while misaligning these regions of interest or action discontinuity between cuts leads to more exploratory behavior from the viewer. In [23], a survey was carried out which aimed to measure the effect of cut frequency on viewers disorientation and their ability to follow a story. Their findings suggested that if the ROI remains consistent across cuts, a high frequency does not increase disorientation or affect the ability to follow the story.

Nielsen et al. [32] investigated two methods of directing the viewer in a 360° narrative short; one where the orientation of the virtual body was faced in the ROI, the other where the viewers' attention was guided by the use of implicit diegetic guidance, in this case a firefly (which is story-centric sense in the context of the scene). They found that the viewers preferred the firefly method of guiding attention and that forcing the viewer's attention by orientating the virtual body increased visual discomfort. A similar approach to non-narrative 360° videos can be found in [27]. Blur was also evaluated as a method to direct the viewer within a virtual environment in [21] and in a 360° video in [15].

Scan-path metrics and visualization. Scan-path metrics have been developed for the processing and analyzing of eye movement such as string edit algorithms [9, 13, 26], probabilistic approaches [24], and geometric vector based comparisons [17, 19]. For instance, Shepherd et al. developed several methods in [41] to analyze between scan-paths. Among these methods one being when two samples have a Euclidean distance that is less than a certain threshold then an overlap is said to occur. Several geographic movement data visualizations were tested for eye movement data in [2]. One of which was the distance function of path similarity analysis.

[1] https://v-sense.scss.tcd.ie/?p=2477

Visualizing viewer behavior in a 360° video has been the subject of a number of recent papers. Bender [4] investigated attentional synchrony or "how does the gaze of multiple viewers exhibit a high degree of clustering in space and time?". To do this the gaze data of 21 viewers was tracked across two narrative cinematic VR films. Complied heatmaps were then used to measure the viewers' attention; this had the effect of being unable to isolate particular user data or provide a complex statistical analysis.

In [28] Löwe et al. developed a visualization to illustrate attentional synchrony across multiple viewers of a 360° video. They proposed a view similarity measure to illustrate this information. Three other visualizations as part of their proposed visual analytic work-flow are: a limited view from the participants' perspective, a 3D sphere-mapped version of the video to provide spatial context, and an unwrapped view of the entire frame to provide global context.

An analytics tool was developed for 360° VR in [3] that allows someone without coding skills to select areas in the scenes that were key to the story. The motivation behind the paper is quite close to the reasoning in this one. The main difference in the methodology being the tracker used here is developed for a post-production tool for greater ease as to how it could be included in the post-production environment and our aim to analyze the effectiveness of the filmmakers' artistic intent of their films from their direct input.

Visual attention. As our work can be seen as pilot survey for visual attention modeling and saliency prediction for 360° video, we briefly describe related works in this area. Good overviews in visual attention research can be found in [8, 11, 22]. Visual attention for 360 contents, however, is a relatively new research area with only a few publications in the last decade. For instance, the authors of [6] presented a computational model of dynamic visual attention on the sphere which combines static features (i.e. intensity, chromatic, and spherical orientation) and motion features.

More recently, a testbed suitable for subjective evaluations of 360° video was introduced in [45]. The authors of [12] introduced a dataset of head movements of users watching 360° video. The dataset includes data collected from 59 users watching five 360° videos on an HMD. In [44] a simple approach to treat raw experimental head direction trajectories in omnidirectional content to obtain visual attention maps was proposed. The authors of [16] collected viewport data of 32 participants for 21 360° images and proposed a new method, fused saliency maps, to transform the gathered data into saliency maps. Later, the authors of [36] proposed a saliency estimation approach for 360° video.

3 SCAN-PATH METRICS

Measuring the similarity between the director's cut and the scan-paths of users wearing an HMD is not only of interest for the director who wants feedback if a viewer can actually follow his story or not; the integration of a director's cut is also of interest for streaming solutions as proposed by MPEG [43, 47]. In order to measure the similarity between the director's cut and the collected scan-paths, we chose the following metrics for the evaluation in Section 5:

(1) Angle between the vector of the director's cut and the vector of the user scan-path.

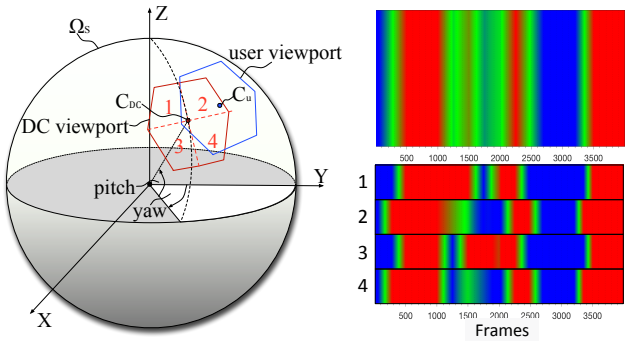


Figure 2: Left: Sphere Ω_S with DC viewport and user viewport. C_{DC} and C_u are the viewport centers of the DC and the user. 1 to 4 are the individual quarters of the DC viewport. Top right: DCS map with overall viewport overlap (red: full overlap, red: 50% overlap, blue: no overlap), bottom right: DCS map with overlap of viewport quarters.

(2) Euclidean distance between the vector of the director's cut and the vector of the user scan-path.
(3) Percentage of overlap between the DC viewport and the user viewport.
(4) Percentage of overlap per quarter between the DC viewport and the user viewport.
(5) Percentage of amount of frames with a viewport overlap of at least 50%.

While the first two metrics are standard metrics, the measurement of viewport overlaps is not commonly used to measure scan-path similarities. However, the percentage of viewport overlaps gives a direct indication if the viewer is looking into the right direction. Measuring the percentage of overlap per viewport quarter also gives additional temporal and directional information of the viewing directions over time. Moreover, for streaming applications where a so-called director's view is provided as an extra bitstream, the degree of overlapping viewports between the director's view and the user's viewport has a high impact on the rate-distortion and thus the quality at the user's side.

Figure 2 shows schematically a spherical surface Ω_S with the DC viewport, an overlapping user viewport, and the center locations of the viewports. Furthermore, the DC viewport is divided into four symmetrical quarters Q_1 to Q_4. It can be assumed that the user can follow the story if the director's cut (the center location of the DC viewport) is within the user viewport, i.e. if the overlap between the DC viewport and the user viewport is at least 50%.

We also introduce two new simple but efficient visualization methods to display viewport overlaps, the so-called *Director's Cut Similarity maps* (*DCS maps*), which are color-coded representations of the percentages of viewport overlaps on a frame-by-frame basis (see right graphs in Figure 2). While the upper DCS map visualizes the overall viewport overlaps between the director's cut and the scan-path, the lower DCS map visualizes the viewport overlaps per quarter. They should give additional information to the filmmakers, namely from where or into which direction the users are drifting. As an example, if the viewport overlap is 100% (all 4 quarters are red) and then quarters 1 and 3 are becoming green and blue while

quarters 2 and 4 are still red, the filmmaker can easily notice that the users are drifting to the left and thus check for reasons which might cause this behavior.

This visualization has been developed together professional VR filmmakers who need intuitive visual feedback about how good their directional cues work directly during post-production and not after the film is released. Current metrics and their visualization do not fulfill this task as they either do not provide the information over time (e.g. heat maps provided by YouTube, etc.) or they are too complex and not intuitive [5].

4 METHODOLOGY

In order to compare a director's cut with scan-paths recorded during a subjective experiment, we used a dataset of eight monocular 360° videos for testing (see Figure 3). The dataset has a wide range of content types including documentary, advertisement, tourism and education. Each 360° video is in the equirectangular format with various resolutions and frame rates. Table 1 describes the characteristics of the 360° videos used in this work.

4.1 Collection of director's cuts

To collect the director's cuts for the given set of 360° videos, we let five filmmakers set position markers in the equirectangular format of their own videos. The setting of the position markers was done with The Foundry's professional compositing software *Nuke*[2] using the *Tracker* node; *Nuke* together with the plugin suite *CaraVR*[3] are widely used in 360° video post-production and thus quite suitable for integration into current work-flows. The filmmakers were instructed to set tracking points manually at the intended viewing locations with keyframes every 2nd second. Tracking points between the keyframes were calculated through linear interpolation. Fig. 4 illustrates the setup in Nuke with the tracker node applied to the video in order to create the director's cut.

Finally, the *Nuke* project files were uploaded to the website Tracksperanto[4] in order to export the tracks into a suitable format and store them for later computations.

Together with the director's cut, the filmmakers were asked to provide additional information about plot points and directional cues used to attract attention of the viewers. In particular, the filmmakers were asked to provide the level of importance for the story ("plot point", "essential plot point", "not relevant") and the intended viewing behavior ("maintain attention", "free exploration", "not relevant") within certain frame ranges. Besides this, the following directional cues were requested:

- Sound ("character/object", "other sound cues")
- Environment ("brightness/contrast/color", "visual effects elements", "other environment cues")
- Motion/action ("camera motion", "character/object motion", "other motion cues")

4.2 Collection of user data

4.2.1 Test subjects. Subjective experiments were conducted with 20 participants (16 males and four females). The participants were aged between 22 to 46 with an average of 30.8 years. 50% of the participants had a medium familiarity with visual attention studies; 35% and 15% of the participants had no and high familiarity with visual attention studies respectively. Furthermore, eight participants wore glasses, and all of the participants were screened and reported normal or corrected-to-normal visual acuity.

4.2.2 Test-bed. We developed a test-bed to collect the viewport tracking data for a given set of 360° video from the participants. The test-bed was implemented using two APIs, namely, `three.js` [10] and `WebVR` [1]. The former enabled us to create and display GPU-accelerated 3D graphics in a web browser. The latter enabled the creation of fully immersive VR experiences in a web browser, allowing us to display a set of 360° videos without the use of any other specific software. The participants viewed each 360° video on the Oculus Rift CV1 while the test-bed continuously recorded their head orientation. In parallel with the video, the audio data was sent to the integrated headphone of the HMD.

4.2.3 Test procedure. Subjective tests were performed as *task-free* viewing sessions, *i.e.,* each participant was asked to look naturally at each presented 360° video while seated in a freely rotatable chair. Each session, which lasted approximately 30 minutes, was split into a training and a test session. During the training session, one minute of the *Help* [18] 360° video was played to ensure a sense of familiarity with the viewing setup. Then, during the test session, the test videos were randomly displayed while the individual viewport trajectories were recorded for each participant.

After each presented video, we inserted a short questionnaire period where the test subjects were asked to answer some questions, while a gray screen was displayed. Three of the questions, which we evaluate in Section 5 are

- Q_1: Did you feel any discomfort?
- Q_2: Did you feel immersed in the environment/ engaged with the video?
- Q_3: Did you feel any disorientation?

The full list of questions and answers for further evaluation is provided with the dataset. Before playing the next 360° video, we reset the HMD sensor to return to the initial position.

5 EVALUATION

Subjective data was collected from participants for the dataset as described in the previous section. These allow us to explore similarities between the intended viewing directions of the filmmakers and the actual viewing direction of the users. In particular, we are interested to analyze these similarities at certain plot points, provided by the filmmakers, which might be essential to follow the story. Furthermore, if directional cues had been provided by the filmmakers, the similarity measures can give an indication about the efficiency of the used directional cues. However, well pre-defined stimuli of directional cues are actually necessary to eventually assess the efficiency objectively. Finally, the examination and partial evaluation of the questionnaires might give additional information when evaluating the statistics of the similarity measures.

[2] https://www.foundry.com/products/nuke
[3] https://www.foundry.com/products/cara-vr
[4] http://tracksperanto.guerilla-di.org

(a) *360Partnership* (b) *Cineworld* (c) *DB* (d) *Jaunt*

(e) *Smart* (f) *Luther* (g) *Vaude* (h) *War*

Figure 3: Sample frames from the eight 360° videos used for the experiment.

Table 1: Description of the dataset. The *Help* video is the training video.

Video	Content Description	Resolution	FPS	Duration
Help [18]	Science fiction film: alien destroys buildings and objects; slow moving camera.	3840×2160	30	1m
360Partnership	Documentary: urban Indian quarters and schools; camera mostly static with long shots.	3840×1080	30	6m17s
Cineworld	Commercial: dark interior with forced viewer attention by use of graphic arrows on screen; moving camera.	2560×1280	30	1m
DB	Commercial: bright lit interior and exterior scenes; slow paced moving camera.	4096×1024	30	3m58s
Jaunt	Commercial: scene of a parties interior. Actor addresses camera. Slow moving camera.	2304×1152	60	2m52s
Smart	Commercial: camera point of view inside moving car; fast movement outside of car.	2880×1440	60	2m7s
Luther	Tourism: various German interior and exterior sites; high amount of cuts; camera mostly static.	4096×2048	30	4m25s
Vaude	Commercial: scenic mountain exteriors and factory floor interior; slow moving camera.	4096×2048	30	2m25s
War	Education: exterior trenches in World War 1; mostly static camera.	4096×1152	25	3m25s

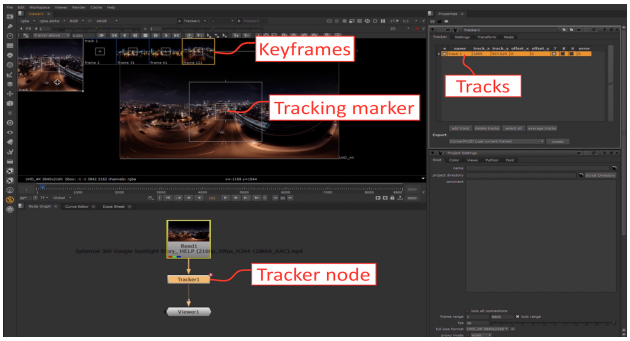


Figure 4: Screenshot of Nuke with tracker node to create the Director's Cut at certain keyframes.

5.1 Subjective questionnaire

The answers to questions Q_1 to Q_3 give information about the condition of the test subjects when watching each of the videos under test. Thus, we first evaluate these questions using a 3-point scale: "no", "maybe", and "yes" using the Kruskal-Wallis non-parametric test. We found no statistical differences (p>0.05) among the answers for question Q_1 (Q_1: $\tilde{\chi}^2$=13.71, df=7, p=0.06). However, significant differences (p<0.05) were found among the answers of the Q_2 and Q_3 (Q_2: $\tilde{\chi}^2$=18.65, df=7, p=0.01; Q_3: $\tilde{\chi}^2$=15.27, df=7, p=0.03). The number of the answers ("no", "maybe", "yes") of the 20 test subjects and eight test videos for the questions $\{Q_1, Q_2, Q_3\}$ is reported in Table 2.

Table 2: Mean values of answers ("no", "maybe", "yes") to the point-scale answered questions for all test subjects.

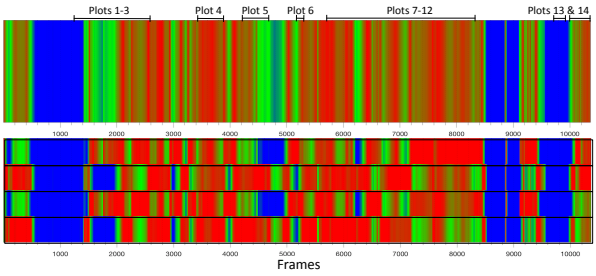
Video	Q_1^v	Q_2^v	Q_3^v
360Partnership	(**16**, 2, 2)	(2, 4, **14**)	(**19**, 1, 0)
Cineworld	(**12**, 2, 6)	(**13**, 2, 5)	(9, 1, **10**)
DB	(**18**, 1, 1)	(6, 1, **13**)	(**15**, 3, 2)
Jaunt	(**17**, 0, 3)	(4, 2, **14**)	(**16**, 2, 2)
Smart	(**9**, 5, 6)	(5, 1, **14**)	(**15**, 1, 4)
Luther	(**16**, 1, 3)	(2, 5, **13**)	(**17**, 1, 2)
Vaude	(**15**, 1, 4)	(2, 7,**11**)	(**14**, 2, 4)
War	(**13**, 4, 3)	(3, 5, **12**)	(**12**, 4, 4)

As can be seen in the table, the *Smart* and *Cineworld* videos seem to cause a relatively high degree of discomfort with six participants answering yes for question Q_1, which is likely caused by the moving camera. Furthermore, the level of immersion for *Cineworld* is quite low with 13 test subjects answered "no" at question Q_2. Interestingly, half of the test subjects (ten participants) felt disorientated in the *Cineworld* video (see Q_3), although graphical arrows were used as directional cues in order to guide the viewer.

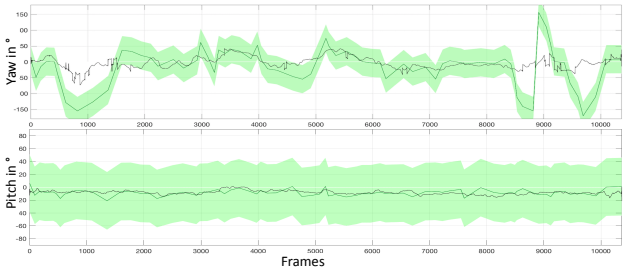
5.2 Similarity measures

For measuring the similarity between the director's cut and the users' scan-paths, we applied the metrics as introduced in Section 3 and show exemplary the results of two videos, *Jaunt* and *Smart*, in

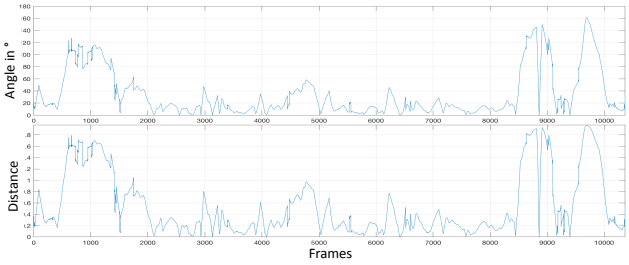
(a) Screen shots of plot points 4 and 12 with ROI



(b) DCS maps with plot points

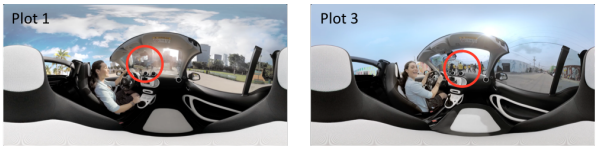


(c) Director's cut (dark green) with viewport area (light green) and user's scan-path (black)

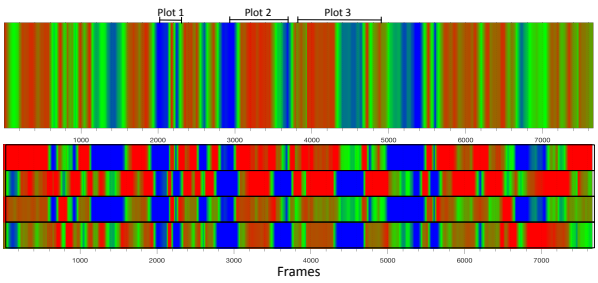


(d) Angle and distance between director's cut and user's scan-path

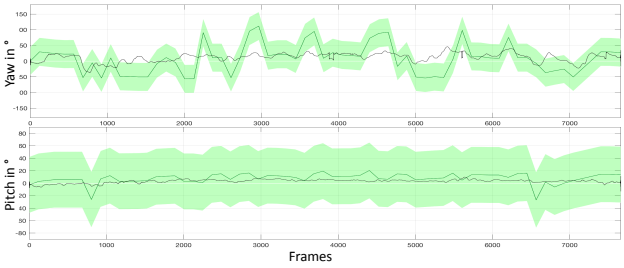
Figure 5: Visualization of metrics for the *Jaunt* video



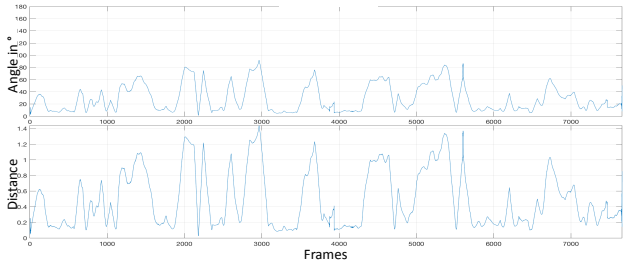
(a) Screen shots of plot points 1 and 3 with ROI



(b) DCS maps with plot points



(c) Director's cut (dark green) with viewport area (light green) and user's scan-path (black)



(d) Angle and distance between director's cut and user's scan-path

Figure 6: Visualization of metrics for the *Smart* video

Figures 5 and 6. The graphs show mean values of all 20 test subjects. The individual graphs of all test subjects and all videos are available with the dataset.

Figures 5a and 6a show two frames from certain plot points of the videos with the ROI highlighted. The plot points, which were provided by the filmmakers, are displayed on top of the DCS maps in Figures 6b and 6b, respectively. A first look at both figures shows a relatively high similarity between the director's cuts and the users' scan-paths, which can be seen in the large red and green areas in the DCS maps and the overlaps of the scan-paths with the viewport area for yaw and pitch in Figures 5c and 6c. Furthermore, as expected, the scan-paths are equator biased, i.e. users tend to look into the direction of the equator rather than the pole caps of the omnidirectional video.

When taking a closer look at the *Jaunt* video in Figure 5, one can notice three larger dissimilarities between the scan-paths between frames 550 and 1,200, frames 8,500 and 9,100, and between frames 9,550 and 10,000. The first two areas are no plot points (only a small part of the first area belongs to plot point 1). No specific directional cues have been applied to attract users' attention. Furthermore, the filmmaker's intent was to let the users explore the environment. Thus, it is unlikely that the director's cut and the users' scan-paths have a high degree of similarity. However, the third area contains plot point 13 where a logo was composed into the scene as directional cue. Here, the logo was composed twice, at yaw 0° and 180°. It is obvious that the test subjects did not follow the fast turn of the head as intended by the filmmaker as can be seen in Figure 5c.

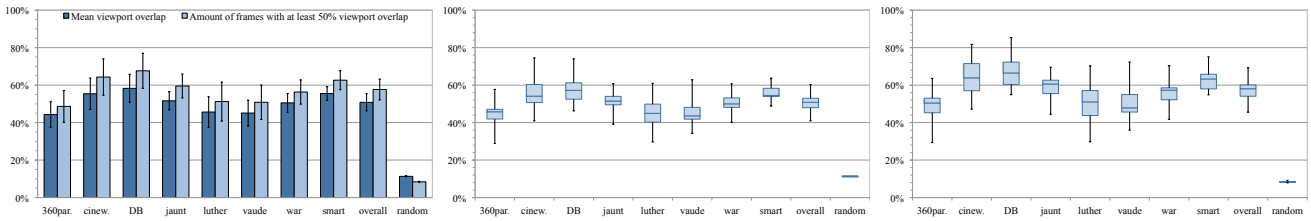


Figure 7: Evaluation of viewport overlaps between the scan-paths for all videos and participants. Left: mean values and standard deviation, middle: boxplot of average viewport overlap, and right: boxplot of frames with at least 50% viewport overlap.

However, as the test subjects mainly look at yaw 0°, they are still able to see the logo.

The *Smart* video has only three major plot points (see Figure 6). While the first plot point has a relatively low similarity between the director's cut and the users' scan-paths, the other two plot points show a relatively high similarity. At this point, we need to mention that the camera was moving quite fast with the speed of the car as it was mounted inside the car. The motion cue of the camera will likely let the users look into the direction of the motion (i.e. yaw 0°) in order to not get motion sick. This video had the highest value of discomfort, probably caused by the motion of the car. Thus, while the director used additional directional cues, such as acting characters outside the car, the effect of the camera's motion was a more dominant cue.

The statistical evaluation of viewport overlaps between the director's cut and the users' scan-paths for all videos across all frames are presented in the diagrams in Figure 7. As reference, we computed a random director's cut and 20 random scan-paths with uniform distribution taking the equator bias into account. The left diagram shows the mean values and standard deviation of the average viewport overlaps and amount of frames with at least 50% viewport overlap for all 20 participants, respectively. The other two diagrams show the boxplots with the same viewport overlap measures.

Obviously, all videos under test have a higher similarity between the director's cut and the users' scan-paths than the randomized scan-paths. The overall mean viewport overlap is 50.79% and the mean amount of frames with a viewport overlap > 50% is 57.64% while the random reference is 11.35% and 8.42%. At this point, we want to mention that the viewport of the Oculus Rift CV1 only covers between 11% and 12% of the sphere.

As can be seen in Figure 7, the *DB* video has the highest amount of viewport overlaps, followed by *Cineworld* and *Smart*. On the other side, *360partnership* has the smallest similarity between director's cut and users' scan-paths. Here, the reason is quite straight forward; the video is a documentary which was produced in a way as if it is a traditional documentary with the only difference that a 360° camera was used. The video has neither essential plot points nor are special directional cues included in order to attract users' attention. Only the action of characters within the environment lets users look into the direction of the director's cut at certain points in time.

For five of the eight videos, the filmmakers had provided additional information like plot points and directional cues used to attract users' attention. Figure 8 shows the statistics for the individual plot points for all five videos. As reference, we added the statistics of the total plot point frames as well as all frames of each video.

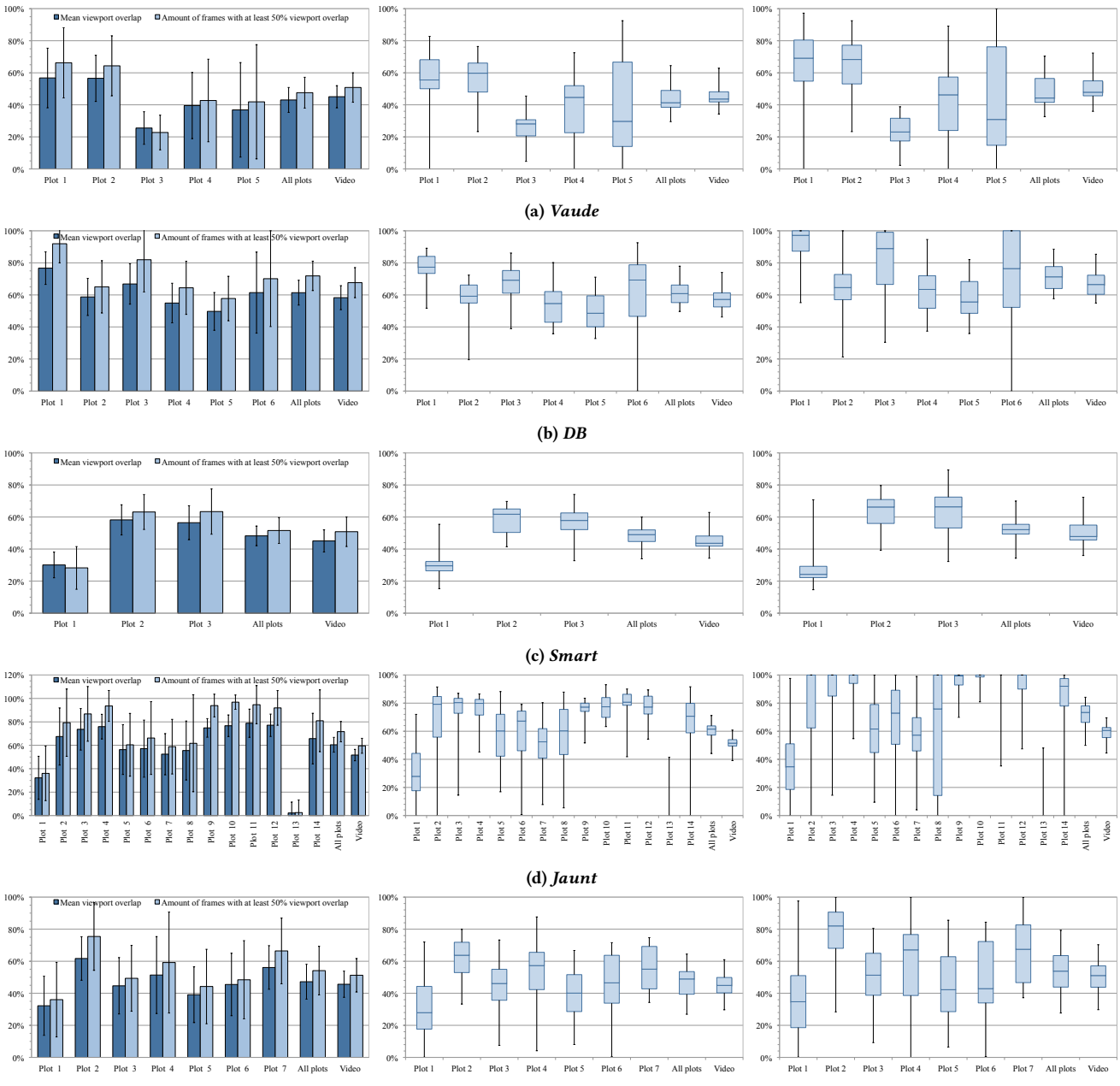
Vaude. The overall similarity between the director's cut and the users' scan-paths for all plot point frames in the *Vaude* video is, within the statistical error tolerance, lower than across the entire video, . This means that the director was not able to attract higher attention at plot points. This is mainly caused by conflicting directional cues. While the main directional cue is the voice of the character, the landscape, i.e. the mountain scenery, is an environmental cue which causes the viewers to freely look around while listening to the main character. In particular, for plot point 3, the voice of the bikers is too weak to maintain users' attention.

DB. The *DB* video shows a slightly higher similarity for all plot point frames compared to all frames of the video. Especially, plot point 1 has a significantly higher similarity, which is caused by an overlay, i.e. a visual effects element, to attract users' attention. Similar overlays have also been used for plot points 3, 5 and 6.

Smart. For the *Smart* video, the similarity for all plot point frames and the entire video is nearly identical. While plot points 2 and 3 have a higher similarity, plot point 1 has a much lower similarity compared to the entire video. Moreover, users seem to look more often directly into the direction of the director's cut than slightly next to it as the percentage of frames with a viewport overlap > 50% is slightly smaller than the average viewport overlap. The low similarity is, as previously mentioned, likely caused by the fast camera motion which is again a conflicting directional cue to the environmental cue, namely the actors on the street.

Jaunt. The *Jaunt* video has plenty of plot points as it is a commercial with many story-centric graphical effects. The director composed many graphical overlays into the environment which the user should recognize and added special sound cues which help to maintain user attention. Plot point 1 has no essential directional cues which explains the low similarity. The very low similarity of plot point 13 was previously explained with Figure 5.

Luther. For the *Luther* video, the similarity for all plot point frames and the entire video is also nearly identical. This video is actually different to most of the other videos as it has a relatively large amount of cuts while at each cut a new scenery is introduced. This usually results in a new exploration of the users in the scene, i.e. strong directional cues are necessary to maintain attention to certain parts of the scene. In plot points 2 and 7, sound cues like thunder (plot point 2) and a voice over saying "look how it goes up" are used and seem to increase the visual attention.



(a) *Vaude*

(b) *DB*

(c) *Smart*

(d) *Jaunt*

(e) *Luther*

Figure 8: Evaluation of viewport overlaps between the scan-paths for all plot points and participants. Left column: mean values and standard deviation, middle column: boxplot of average viewport overlap, and right column: boxplot of frames with at least 50% viewport overlap.

6 CONCLUSION

In this paper, we provided a new combined dataset for visual attention analysis and introduced a novel simple but efficient metric and visualization method for similarity measures between a director's cut and users' scan-paths. The metric has a high relevance for compression and streaming applications towards adding an additional bitstream utilizing the director's view, because the degree of overlapping user viewports with the director's view has a high impact on the rate-distortion and thus the quality at the user's side. Furthermore, the visualization of viewport overlaps using color-coded DCS maps makes it easy for content creators to analyze if users are able to actually follow the story. Once visual attention in 360° video (i.e. where users look in the presence of directional cues) can

accurately be predicted, e.g. with deep learning approaches and sufficiently high amount of available data for learning. Such visualization could be integrated in common post-production software applications in order to give feedback to content creators directly during the editing.

We then collected eight intended scan-paths, the so-called director's cuts, including the 360° videos, plot points and directional cues from five professional VR filmmakers and performed a subjective experiment with 20 test subjects who watched the videos while their head orientation was recorded.

Then, the subjective test data was compared against the director's cuts using the proposed metrics. The results show that the mean viewport overlap and the mean amount of frames with a viewport overlap > 50% for all videos are 50.79% and 57.64%, respectively. These numbers indicate that the utilization of an additional bitstream containing the director's view, which was introduced by MPEG, will likely increase the quality at the consumer side as a large portion of pixels of the user's viewport can directly be rendered from the director's view. However, this still needs to be evaluated in a real streaming application, which is beyond the scope of this paper. Another aspect in this context is the question if the director's cut, i.e. the intended scan-path and not the director's view, should be integrated into the metadata of the omnidirectional media format (OMAF). This, however, requires that the filmmakers have efficient and easy-to-use tools available to create such director's cuts. In this work, we presented a method to create a director's cut with the professional and commonly used post-production software *Nuke*, which allows a simple integration into current post-production work-flows.

The results also give good indications to filmmakers, namely, how effective certain directional cues might work to attract viewers' attention. Storytelling in cinematic VR is still in an early stage and all stakeholders are at the beginning of a learning curve. Although pre-defined and separated stimuli would be necessary to objectively assess certain directional cues, the collected data and evaluation give good insights into the interaction between directional cues. For instance, the fast camera motion in the *Smart* video was a strong directional cue which even caused discomfort for many test subjects. As camera motion is a strong directional cue and viewers want to avoid to become motion sick, environmental cues may compete with motion cues and thus become less effective. On the other side, arbitrary environmental cues (like the mountain scenery in *Vaude*) might be in conflict with sound cues, which the director actually used to guide the viewer.

Finally, the entire dataset, i.e. the director's cuts including the directional cues and plot points as well as the scan-paths of the test subjects, is publicly available with this paper and can be accessed with further details at https://v-sense.scss.tcd.ie/?p=2477. To our knowledge, it is the first time that director's cuts of professional VR filmmakers together with subjective data has been provided to the scientific community. This combined dataset can be seen as a first important step to contribute to new streaming concepts like the one introduced by MPEG. Moreover, it can further be used to develop and test saliency prediction approaches, which can then be integrated into post-production applications, and new streaming solutions by creating and utilizing the additional 'director's view', which can easily be rendered from the director's cut, i.e. it is an important contribution for the multimedia community.

Although the collected data is much richer than presented in this paper, e.g. we asked between six and seven video related questions after each video screening and seven general questions after the entire experiment, it is beyond the scope of this paper to evaluate all of it. This will be part of future studies. Especially the directional cues and the answers to all of the questions will be evaluated together with further experiments for storytelling in VR.

ACKNOWLEDGEMENT

We would like to thank the VR filmmakers Angus Cameron, Soenke Kirchhof, Josef Kluger, Declan Dowling and Jack Morrow for fruitful discussions, their great support by providing the Director's Cuts for their VR films and for their feedback to develop the DCS maps.

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/2776.

REFERENCES

[1] 2017. WebVR: Bringing Virtual Reality to the Web. https://webvr.info/. (Feb 2017).

[2] Gennady Andrienko, Natalia Andrienko, Michael Burch, and Daniel Weiskopf. 2012. Visual analytics methodology for eye movement studies. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2889–2898.

[3] Paulo Bala, Mara Dionisio, Valentina Nisi, and Nuno Nunes. 2016. IVRUX: A tool for analyzing immersive narratives in virtual reality. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Frank Nack and Andrew S. Gordon (Eds.). Springer International Publishing, Cham, 3–11. https://doi.org/10.1007/978-3-319-48279-8_1 arXiv:9780201398298

[4] Stuart Marshall Bender. 2018. Headset Attentional Synchrony: Tracking the Gaze of Viewers Watching Narrative Virtual Reality. *Media Practice and Education* (May 2018), 1–20.

[5] Tanja Blascheck, Kuno Kurzhals, Michael Raschke, Michael Burch, Daniel Weiskopf, and Thomas Ertl. 2014. State-of-the-Art of Visualization for Eye Tracking Data. In *Eurographics Conference on Visualization (EuroVis)*.

[6] Iva Bogdanova, Alexandre Bur, Heinz Hugli, and Pierre-Andre Farine. 2010. Dynamic visual attention on the sphere. *Computer Vision and Image Understanding* 114, 1 (2010), 100 – 110. https://doi.org/10.1016/j.cviu.2009.09.003

[7] Ruud Bolle, Yiannis Aloimonos, and Cornelia Fermüller. 1997. Toward motion picture grammars. In *Computer Vision — ACCV'98*, Roland Chin and Ting-Chuen Pong (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 283–290.

[8] Ali Borji and Laurent Itti. 2013. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 185–207. https://doi.org/10.1109/TPAMI.2012.89

[9] Ulrich Bunke. 1992. Relative index theory. *Journal of functional analysis* 105, 1 (1992), 63–76.

[10] Ricardo Cabello et al. 2017. JavaScript 3D library. https://threejs.org/. https://github.com/mrdoob/three.js/. (Feb 2017).

[11] Marisa Carrasco. 2011. Visual attention: The past 25 years. *Vision Research* 51, 13 (2011), 1484–1525. https://doi.org/10.1016/j.visres.2011.04.012

[12] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 2017. 360-degree video head movement dataset. In *Proceedings of the 8th ACM Multimedia Systems Conference (MMSys)*. ACM, Taipei, Taiwan, 199–204. https://doi.org/10.1145/3083187.3083215

[13] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D. Gilchrist. 2010. ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods* 42, 3 (01 Aug 2010), 692–700. https://doi.org/10.3758/BRM.42.3.692

[14] Simone Croci, Sebastian Knorr, Goldmann. Lutz, and Aljosa Smolic. 2017. A Framework for Quality Control in Cinematic VR Based on Voronoi Patches and Saliency. In *International Conference on 3D Immersion (IC3D)* (2017-12-11). IEEE, Brussels, Belgium, 1–8.

[15] Fabien Danieau, Antoine Guillo, and Renaud Dore. 2017. Attention guidance for immersive video content in head-mounted displays. In *2017 IEEE Virtual Reality (VR)*. IEEE, Los Angeles, CA, USA, 205–206. https://doi.org/10.1109/VR.2017.7892248

[16] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic. 2017. Look around you: Saliency maps for omnidirectional images in VR applications. In *Proceedings of the 9th*

International Conference on Quality of Multimedia Experience (QoMEX). IEEE, Erfurt, Germany, 1–6. https://doi.org/10.1109/QoMEX.2017.7965634

[17] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. 2012. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods* 44, 4 (01 Dec 2012), 1079–1100. https://doi.org/10.3758/s13428-012-0212-2

[18] Justin Lin (Director). 2015. Help (2015). http://www.imdb.com/title/tt4794550/. (2015).

[19] Rebecca M Foerster and Werner X Schneider. 2013. Functionally sequenced scanpath similarity method (FuncSim): Comparing and evaluating scanpath similarity based on a task's inherent sequence of functional (action) units. *Journal of Eye Movement Research* 6, 5 (2013).

[20] Rorik Henrikson, Bruno Araujo, Fanny Chevalier, Karan Singh, and Ravin Balakrishnan. 2016. Multi-device storyboards for cinematic narratives in VR. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, Tokyo, Japan, 787–796.

[21] Sebastien Hillaire, Anatole Lecuyer, Remi Cozot, and Gery Casiez. 2008. Depth-of-Field Blur Effects for First-Person Navigation in Virtual Environments. *IEEE Computer Graphics and Applications* 28, 6 (Nov 2008), 47–55. https://doi.org/10.1109/MCG.2008.113

[22] Laurent Itti and Ali Borji. 2014. Computational models: Bottom-up and top-down aspects. *The Oxford Handbook of Attention* (2014), 1–20.

[23] Tina Kjær, Christoffer B Lillelund, Mie Moth-Poulsen, Niels C Nilsson, Rolf Nordahl, and Stefania Serafin. 2017. 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*. ACM, Gothenburg, Sweden, 4.

[24] Thomas C. Kübler, Enkelejda Kasneci, and Wolfgang Rosenstiel. 2014. SubsMatch: Scanpath Similarity in Dynamic Scenes Based on Subsequence Frequencies. In *Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA '14)*. ACM, New York, NY, USA, 319–322. https://doi.org/10.1145/2578153.2578206

[25] Christopher A Kurby and Jeffrey M Zacks. 2008. Segmentation in the perception and memory of events. *Trends in cognitive sciences* 12, 2 (2008), 72–79.

[26] Vladimir I Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 8 (Feb 1966), 707–710.

[27] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. 2017. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver, Colorado, USA, 2535–2545.

[28] Thomas Löwe, Michael Stengel, Emmy-Charlotte Förster, Steve Grogorick, and Marcus Magnor. 2015. Visualization and analysis of head movement and gaze data for immersive video in head-mounted displays. In *Proceedings of the Workshop on Eye Tracking and Visualization (ETVIS)*, Vol. 1.

[29] John William Mateer. 2017. Directing for Cinematic Virtual Reality: how traditional film director's craft applies to immersive environments and notions of presence. *Journal of Media Practice (author-produced version)* 18, 1 (5 2017), 14–25. https://doi.org/10.1080/14682753.2017.1305838

[30] Rafael Monroy, Sebastian Lutz, Tejo Chalasani, and Aljosa Smolic. 2018. SalNet360: Saliency maps for omni-directional images with CNN. *Signal Processing: Image Communication* (2018). https://doi.org/10.1016/j.image.2018.05.005

[31] Afshin Taghavi Nasrabadi, Anahita Mahzari, Joseph D. Beshay, and Ravi Prakash. 2017. Adaptive 360-Degree Video Streaming Using Scalable Video Coding. In *Proceedings of the 2017 ACM on Multimedia Conference (MM '17)*. ACM, New York, NY, USA, 1689–1697. https://doi.org/10.1145/3123266.3123414

[32] Lasse T Nielsen, Matias B Møller, Sune D Hartmeyer, Troels C M Ljung, Niels C Nilsson, Rolf Nordahl, and Stefania Serafin. 2016. Missing The Point: An Exploration of How to Guide Users' Attention During Cinematic Virtual Reality, In VRST '16 Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technolog. *VRST '16 Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*, 229–232. https://doi.org/10.1145/2993369.2993405

[33] University of Nantes and Technicolor. 2017. Salient360!: Visual attention modeling for 360° images grand challenge. (2017). http://www.icme2017.org/grand-challenges/

[34] Cagri Ozcinar, Ana De Abreu, Sebastian Knorr, and Aljosa Smolic. 2017. Estimation of optimal encoding ladders for tiled 360° VR video in adaptive streaming systems. In *The 19th IEEE International Symposium on Multimedia (ISM 2017)* (2017-12-11). IEEE, Taichung, Taiwan, 45–52.

[35] Cagri Ozcinar, Ana De Abreu, and Aljosa Smolic. 2017. Viewport-aware adaptive 360° video streaming using tiles for virtual reality. In *2017 IEEE International Conference on Image Processing (ICIP)* (2017-09-30). IEEE, Beijing, China, 2174–2178. https://doi.org/10.1109/ICIP.2017.8296667

[36] Cagri Ozcinar and Aljosa Smolic. 2018. Visual Attention in Omnidirectional Video for Virtual Reality Applications. In *10th International Conference on Quality of Multimedia Experience (QoMEX)* (2018-05-29).

[37] Amy Pavel, Björn Hartmann, and Maneesh Agrawala. 2017. Shot Orientation Controls for Interactive Cinematography with 360 Video. In *UIST '17 Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. Quebec City, QC, Canada, 289–297. https://doi.org/10.1145/3126594.3126636

[38] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. A dataset of head and eye movements for omni-directional images. In *Proceedings of the ACM International Conference on Multimedia Systems (MMSys)*. ACM, Taipei, Taiwan, 205–210.

[39] Ana Serrano, Vincent Sitzmann, Jaime Ruiz-Borau, Gordon Wetzstein, Diego Gutierrez, and Belen Masia. 2017. Movie editing and cognitive event segmentation in virtual reality video. *ACM Transactions on Graphics* 36, 4 (Jul 2017), 47:1–47:12. https://doi.org/10.1145/3072959.3073668

[40] Alia Sheikh, Andy Brown, Zillah Watson, and Michael Evans. 2016. Directing attention in 360-degree video. In *IET Conference Proceedings*. Institution of Engineering and Technology, Amsterdam, Netherlands, 29 (9 .)–29 (9 .)(1). http://digital-library.theiet.org/content/conferences/10.1049/ibc.2016.0029

[41] Stephen V. Shepherd, Shawn A. Steckenfinger, Uri Hasson, and Asif A. Ghazanfar. 2010. Human-Monkey Gaze Correlations Reveal Convergent and Divergent Patterns of Movie Viewing. *Current Biology* 20 (2010), 649–656.

[42] Shamus Smith, Tim Marsh, David Duke, and Peter Wright. 1998. Drowning in immersion. *Proceedings of UK-VRSIG* 98 (1998), 1–9.

[43] Emmanuel Thomas and Alexandre Gabriel. 2017. *Director's view streaming with conventional encoded 360 scenes (viewport dependent scheme) - OMAF*. Technical Report MPEG2017/ M40585. JTC1/SC29/WG11, ISO/IEC, Hobart, AU.

[44] Evgeniy Upenik and Touradj Ebrahimi. 2017. A simple method to obtain visual attention data in head mounted virtual reality. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, Hong Kong, 73–78.

[45] Evgeniy Upenik, Martin Rerabek, and Touradj Ebrahimi. 2016. A testbed for subjective evaluation of omnidirectional visual content. In *Proceedings of the Picture Coding Symposium (PCS)*.

[46] Mirjam Vosmeer and Ben Schouten. 2017. Project Orpheus A Research Study into 360° Cinematic VR. In *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '17)*. ACM, New York, NY, USA, 85–90. https://doi.org/10.1145/3077548.3077559

[47] Eric Yip and Mary-Luc Champel. 2017. *List of proposed features and technologies for OMAF amendment*. Technical Report MPEG2017/ W17236. JTC1/SC29/WG11, ISO/IEC, Macau, China.