Tampere University

Dmitrii Monakhov

ANALYSIS OF USER EXPLORATION PATTERNS DURING SCENE CUTS IN OMNIDIRECTIONAL VIDEOS

Department of Computing Sciences
Master thesis
April 2020

ABSTRACT

Dmitrii Monakhov: Analysis of User Exploration Patterns during Scene Cuts in Omnidirectional Videos
Master thesis
Tampere University
Supervisors: Professor Moncef Gabbouj, Bell Labs Distinguished Member of Technical Staff Igor Curcio
Degree Programme in Information Technology, MSc (Tech)
April 2020

Video content is usually represented as a sequence of scenes joined together. Two adjacent scenes can share the same semantic content, similar to filming the scene from different angles, or they can describe semantically different content. The methods that switch between two adjacent scenes are called scene transitions. In the case when two scenes are just concatenated with no additional effects, the transition is called a scene cut. The scene transition is a vital instrument for guiding user's attention in classical cinema, but the impact of scene transitions becomes more relevant in a 360° video environment. During the scene transition, the user loses his/her previous interest point, that may lead to a change in the exploration behavior and affect the content delivery system, diminishing the viewing experience of the user. In this work, we have studied how the user exploration behavior changes in terms of the exploration range and angular speed metrics, and we have investigated whether this change of exploration behavior is different when the scene cut is within the same semantic content compared to the semantically different scenes.
We conducted an experiment with 20 test subjects. The experiment consisted of two sets of videos. The first set included eight 20 second videos that were semantically different from each other. The videos were stitched together in a temporal domain using scene cuts. The second set was comprised of three videos with 4-5 scene cuts in each of them. The shots between the cuts were semantically consistent. We collected the positional data of head-mounted displays and converted it into the defined metrics, as per above explanation.
Our research showed an increase in exploratory behavior and also revealed that there was a delay between the scene transition and the start of the exploration. The results were attested using t-test procedures. We have also shown that the exploratory behavior is dominant in the inter-scene transitions compared to the intra-scene transitions.

Keywords: Omnidirectional video; 360 Degrees video; Exploration range; Scene transitions; Watching patterns; Scene cuts; Viewport dependent streaming.

.

PREFACE

It is always hard to estimate how much time one's work might take, and this thesis is no exception. Nevertheless, I can say for sure that I would not be able to finish this thesis without the help of my colleagues, mentors and friends.

I would like to thank Moncef Gabbouj for the supervision and guidance of the work. I am incredibly grateful for having a great opportunity to improve and utilize my skills.

Igor Curcio has helped me many times with the ways on how to proceed with the research, and I appreciate the experience I've gained while working with him.

I would like to also thank my colleagues: Deepa Naik, Henri Toukomaa and Sujeet Mate, who have helped me with different aspects of this work.

Finally, I would like to thank my parents, who were very patient and supportive through writing this thesis.

Tampere, April 2020

Dmitrii Monakhov

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS AND ABBREVIATIONS

DASH	Dynamic Adaptive Streaming over HTTP
ER	Exploration Range
FoV	Field of View
FPS	Frames Per Second
HEVC	High-Efficiency Video Coding
HMD	Head Mounted Display
Hz	Hertz
MPD	Media Presentation Description
QoE	Quality of Experience
RoI	Region of Interest
SRM	Similarity Ring Metric.
VR	Virtual Reality

1. INTRODUCTION

The recent achievements in virtual reality media have allowed capturing the content in the 360-degree Field of View (FoV). This type of content becomes more and more popular as Head Mounted Displays such as Occulus Rift and Samsung Gear VR become more accessible to the general public. Nowadays, one of the most popular methods of delivering the omnidirectional content to end users is by streaming it through the network. Modern network infrastructures allow efficient storage and distribution of the material and are much more convenient than the other methods, such as production and shipping of external storage devices. To achieve the same quality of viewing experience on level of watching a traditional movie in SD quality, the content should have 4K resolution and a frame rate of 30FPS or more, resulting in required bandwidth being equal to 100 MB/S [1]. Therefore, we can see that even for lower levels of quality, these restrictions require a large network bandwidth and a significant amount of storage space for keeping the files on a server.

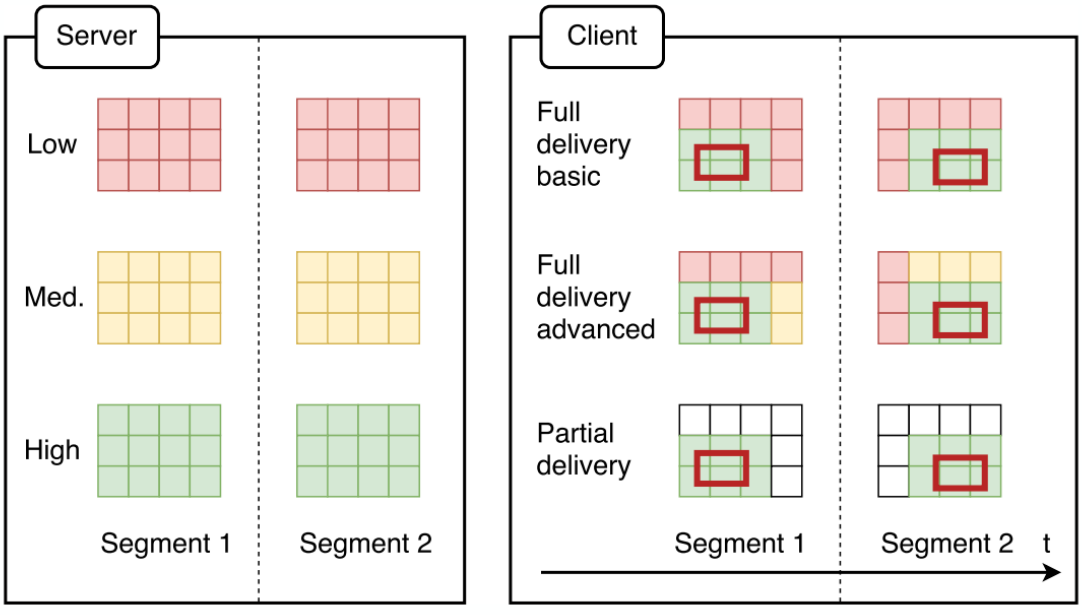

Figure 1. *An example of segment allocation and delivery during a watching session* [2]

The base operational point for 360-degree video streaming is transmitting the whole video at a constant high quality, regardless of the user's input (i.e., the direction the user

is looking at, called **viewport**). Transmitting video at constant quality is known as the **Viewport-Independent** streaming model [3].

However, the end user watching a 360-degree video on a Head-Mounted Display (HMD) sees only a part of the omnidirectional video panorama at a given time and, therefore, there is no need to stream the content outside the user's current viewport. The model that transfers the area inside the viewport in high quality and the surrounding space in lower quality (or not streaming it at all) is called the **Viewport-Dependent** streaming model, also known as viewport-adaptive model [3]. One implementation of that model is to split the viewing area into tiles [2], each having multiple quality levels. During the watching session, the playback client checks what tiles are intersected with the current viewport, and fetches the high-quality versions of those tiles. This method allows reducing the required network bandwidth considerably. This approach is illustrated in Figure 1.

However, this approach has some issues. One of the problems is related to the unpredictability of the user's viewing direction. If the head movement is slow enough, it is relatively easy to set up the switching between the low and high-quality tiles in the manner that the user will not notice it. On the other hand, if the user's motion is swift, he/she can go outside the area of the current high-quality viewport. The switch between low and high quality can be a time consuming operation, so the user may experience a lower quality video content before the switch.

If the user's head movement could be predicted, we would be able to mitigate this effect by prefetching the high quality tiles beforehand. One can assume that the user's head movement correlates with the change of the area of interest. The area of interest changes due to a set of factors, such as the user's personal preferences, the semantic content of the video and the staging/direction techniques. The user's movement due to his/her personal preferences is irregular and can hardly be predicted. This is not the case for staging and directing techniques. One of these techniques is scene transition. In this work, we would like to explore the problem of how the scene transition affects the user's exploration behavior, and whenever the difference in scene transition type impacts the behavior differently. This problem has been studied by us in [4].

While it is true that the **viewport-dependent** streaming model is more efficient in terms of bandwidth consumption, the implementations can still be wasteful due to the user's head movement. The reasoning for that lies in the structure of the content prepared for streaming. The video is fragmented into tiles in spatial domain, but it is also divided into multiple fixed-length segments in time domain, i.e. the video is divided into a collection

of short-length clips. The tiles selected at high quality must be buffered for the duration of the whole segment. If the user moves amid the duration of the segment, some parts of the high-quality area would not be seen and hence are wasted. This problem is quite important, and it has been studied by us in [5]. In this paper we have investigated how the head movement velocity, duration of the segment, and the type of trajectory (across the equator or angled) affects the bandwidth wastage. The experiment was done by simulating artificial trajectories and averaging the data wastage statistics across them. The results were compared with some real trajectories. We have shown that on average 15-20% is wasted due to head movement. We have also shown that the introduction of the trajectory prediction methods in the model can decrease the wastage by 4.3%.

Scene transitions are widely used in the traditional cinematography. The problem of scene transitions and user guidance in omnidirectional videos has been addressed in several scientific papers. Lasse T. Nielsen et al. explored the challenge of guiding the audience's attention in omnidirectional videos [6]. The authors presented a collection of methods that might help to guide the attention of the viewer. The authors clustered the guiding methods according to three aspects:

- Does the method explicitly direct viewers' attention, or it uses implicit cues?
- Is the cue coming from the scene itself and is it semantically a part of the content (diegetic cue), or is it external to the virtual environment (nondiegetic cue)?
- Does the method restrict the interaction with the scene in some way, or the user is free to interact with it?

For example, an explicit nondiegetic cue with no restriction on a HMD, tells what area the user should watch by directing to it with arrows.

The authors compared three guiding methods:

- Forced rotation to the area of interest: the user can freely explore the content, but the orientation of the virtual world is forcefully rotated towards the area of interest.
- Firefly: users can explore the content freely. A small flying firefly gives hints on the areas of interest.
- No additional guidance.

The authors of the mentioned work proceeded in the following manner: one clip was shown to 45 participants individually, with different types of cues as described above. After the viewing session, the authors quizzed the participants about their viewing experience using a questionnaire.

The results have shown a statistically insignificant difference between the methods, and the authors claimed that future studies would benefit from including behavioral and physiological measures of presence.

Exciting but statistically insignificant results of the paper demonstrate the necessity of a stricter measure of the human reaction.

Film editing in traditional cinema is a widely researched topic. However, editing films in virtual environments is a new area that has just began to be explored. In his scientific paper, Serrano et al. analyses how different types of scene cuts affect the user's viewing behavior [7].

The authors reviewed the papers regarding the cognition studies in movie content. The study made by Magliano et al. classified the edits into three groups: edits that are discontinuous in space, time and action (E1), edits that are discontinuous in space and time (E2), and edits that are continuous in space, time and action (E3). The participants of the experiment were asked to segment the video they watched into some meaningful events at two different scales: the most significant meaningful events and the least significant ones. The study showed that the user's segmentation aligned with the storyboard of the video and that E1 edits had the most substantial effect on the perceived discontinuity [8, 9].

In the paper [7] the authors reproduced the research described above in the VR environment. In their work the authors have selected sixteen different videos within the range of thirteen seconds to two minutes. These videos were edited into 216 clips, each clip containing two shots split up by an edit. Each shot lasted about six seconds. The authors have also noted that the VR shots could be much more prolonged compared to shots in conventional cinematography. The duration of a single camera shot in classical cinematography has been steadily decreasing since 1960, and today the average duration of a camera shot is around 5 seconds long [10]. For VR, such duration is too short to explore the environment in its entirety, and fast switching between different scenes can make the user feel uncomfortable. Based on the analysis of the VR videos by the authors, the average length of the shot for omnidirectional video is longer, totaling 20 seconds. It was also noted that the cuts of type E3 are much less frequent for VR videos. Based on the collected statistics, the continuity edits (E3) correspond to 2% of the total number of the edits. Therefore, the authors have excluded those types of edits, focusing only on the type E1 and E2 cuts. The authors confirmed that the users maintained the perception of continuity across the edit boundaries in VR narrative content.

The authors also analyzed how different scene cut parameters affect the viewer's behavior. There are three sets of conditions: 2 types of edit (discontinuous in time, space and action, discontinuous in time and space only), three alignments before and after the cut (0, 40 and 80 degrees region of interest (ROI) disparity before and after the cut) and nine region of interest configurations, that gives 54 different conditions in total. The authors introduced the following metrics:

- framesToROI – indicates how many frames it takes for a viewer to fixate on a new region of interest.
- percFixInside – indicates the percentage of total fixations inside a region of interest. To make this metric comparable between different configurations of ROIs, the authors computed this metric relative to the average percentage of fixations inside the ROI. This average was calculated for each ROI configuration separately and before the edit.
- ScanpathError – RMSE between of each scanpath and the corresponding baseline scanpath.
- Nfix – the ratio between the number of fixations and the total number of gaze samples after the edit.
- State sequences – user's fixations are split for different states: focusing on the first region of interest, focusing on the second region of interest (if it exists), focusing on the background and idle state (saccadic eye movement with no fixations). The authors performed state distribution analysis to general patterns of state sequences.

The main results of the paper were the following:

1. Regarding the influence of previous VR experience, no significant effect was found considering the metrics mentioned above.
2. The initial ROI alignment had a sufficient effect on the metric's values. The authors claimed that higher misalignment across the edit boundary leads to the exploratory behavior of the viewers after the edit and initial ROI search.
3. The type of edit did not show any significant effect within the defined metrics.
4. The authors pointed out that there exists a typical behavioral pattern while watching the video. At first, the user displays a strong exploration peak at the beginning of the session. The same peak appears after the edit; the duration of the exploration peak is around 1-2 seconds. The exploration peak is followed by attention

peak, which also lasts approximately 1-2 seconds. This behavior is consistent within different ROI alignments.

The problem of virtual environment exploration was considered by Sitzmann et al. in their paper [11]. The authors gave quantitative and qualitative results on how people explore virtual reality content. The authors tested 169 users and collected their scan paths while observing static omnidirectional panoramas. The authors found the following results:

- They compared the saliency maps generated from viewings in the HMD and through a desktop application. They concluded that there is no significant difference between these maps in these two cases and, therefore, since the desktop experiments were much easier to control, it might be possible to use these for collecting adequate training sets for data-driven saliency prediction in the future VR systems.
- The authors have shown that users maintain strong equatorial bias and that this bias can be used to improve the accuracy of the existing saliency predictors.
- The entropy of the saliency map has been studied. The authors have shown that entropy affects the transition time between the salient areas. The smaller is entropy , the shorter is the transition time.
- The final saliency maps converge to the same saliency map after 30 seconds, regardless of the starting head orientation.
- The authors claimed that, according to the collected data, the users appear to behave in two different ways: attention and re-orientation. Eye fixations happen in the attention mode when users have “locked-in” on a salient part of the scene, while movements to new prominent regions occur in the re-orientation mode. On average, users fully explored each scene within 19 seconds.
- The prediction obtained from head orientation data only is comparable with the results achieved with both the head and gaze data. This result is especially interesting regarding the subject of the work in this thesis.

The authors suggested the method of the scene’s alignment before and after the cut in order to avoid an abrupt change in content. For this purpose, they proposed to compute saliency maps for both scenes and align the maps horizontally so that the cross-correlation coefficient between the maps is maximized.

In our work, we take into account the equatorial bias and head orientation data described in this paper. The authors reveal that there were two peaks: the peak of exploration and the peak of attention, both lasting 1-2 seconds. Such peaks appear

regardless of the ROI configuration and alignment. The authors pointed out '*we found no significant effect of the type of edit in our metrics; the graphs suggest a difference that our metrics are not capturing'*. Our work tries to explore metrics that would show the difference between the types of edits.

The effect of transitions in the traditional cinematography is a deeply explored topic. Nevertheless, there are very few papers regarding scene transitions in virtual reality. In the paper by Liang Men et al. [12], the authors examine how different types of transitions influence the user experience of presence in VR. They compared four different types of transitions: SimpleCut transition (directly connecting two scenes), SuperFast Transition (swift transfer between two points that are camera positions for the given scenes), Fade transition (using a fade effect during the transition), Vortex transition (using the vortex effect during the transition) with help of a questionnaire that explored how the scene transitions affected the viewer's sense of presence, and how it changed the feeling of perceived realness and also if the scene transitions would generate some sensation in the users. The main conclusion of the authors was that the SimpleCut Transition provides maximum consistency of presence. In our work, we decided to also focus our attention on the SimpleCut transitions.

Current work is based on the set of papers mentioned above. According to [11], head movements have comparable power of prediction to head + eye movements and required more straightforward setup. Therefore, we concentrated our work on head movements only. The metrics introduced by Sitzmann et al. had shown no difference between different types of scene cuts. Nevertheless, the authors pointed out that at a glance, the state distributions between those types of cuts are different. We introduced new metrics that try to discriminate against this difference.

2. THEORETICAL BACKGROUND

2.1 Omnidirectional video streaming

Omnidirectional video is a type of media captured in a field of view larger than general 2D video. The content is captured from multiple cameras (or with one camera using a special lenses, such as fish-eye lenses [13]). Each camera records the corresponding part of a 360 panorama, which are then mapped into one image. Afterwards, The image is projected into one of the commonly used types of projections, such as equirectangular projection, cubic projection or other [14]. The advantages and disadvantages of different projection types and the mechanisms for efficient streaming of the projected data have been a matter of several scientific works [15, 16]. One of the ways to improve the viewing experience in omnidirectional videos is the possibility to show the content stereoscopically. On the other hand, 3D stereoscopic content brings new problems related to the capturing process and user's response to the in-frame motion and artificial depth of field in the video [17]. Another issue arises due to the multiple camera imaging and the problem of arranging left and right 3D images. This problem has been considered and worked out in works such as [18, 19].

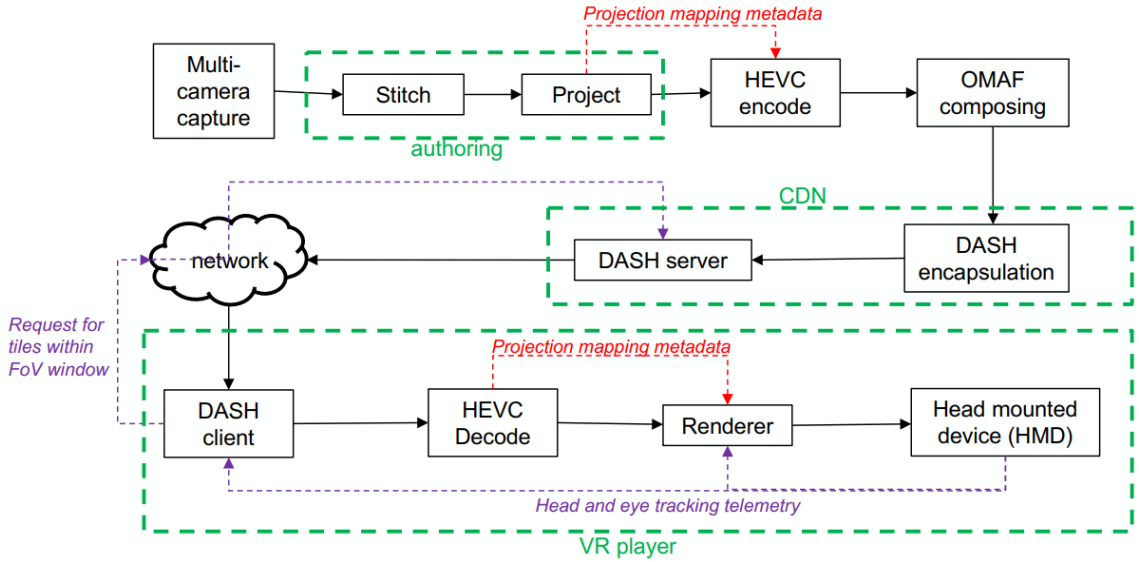

Figure 2. *Example of standard streaming framework for omnidirectional content [2]*

It should be pointed out that the omnidirectional video is extremely resource-consuming due to a large field of view of the frame. This leads to the huge size of the final data file and the necessity of the efficient compression procedure. Therefore, to achieve efficient

compression of the content, HEVC encoding procedures have been used. The advantages and comparison of the method to the older standard can be seen in [20].

After encoding, the data is encapsulated in accordance to the guidelines of the omnidirectional media format (MPEG-OMAF) [21] . To further decrease the bandwidth of the content, the video makes use of the independent tile-decoding procedure, as described in section 1.

The system should be able to switch the content quality with respect to the currently available bandwidth in order to take into account the volatile network bandwidth. MPEG-DASH based streaming systems achieve bandwidth adaptability of the stream by splitting the video stream into smaller segments [22]. Each segment has several representations of varying quality levels, and the system can choose the representation to download depending on the current bandwidth availability. Therefore, the content is divided into smaller segments of fixed duration and different quality levels that are stored separately. A Media Presentation Description (MPD) file is generated, allowing the player to request different representations of the segment. Segments and MPD files are stored on the server. The client requests the content based on the current user orientation and the available bandwidth, which is then rendered onto the screen. The whole procedure is shown in Figure 2.

The **viewport-dependent** streaming model dramatically reduces the required bitrate, achieving bandwidth savings up to 40% [16]. These savings can be further improved with the help of additional practices, such as using different tiling schemes, predicting tiles that will be viewed [23] or using unequal quality levels for each eye in case of stereoscopic video [24]. On the other hand, the **viewport-dependent** model introduces a new problem of **motion-to-high quality delay**. When the user moves his viewpoint into a low-quality area, the quality cannot change instantaneously to high since the high-quality tiles should be streamed to the user and decoded first. Downloading the currently watched segment in high quality would lead to some data wastage, since a part of the segment in low quality which was already viewed cannot be used again in high quality. So, to minimize the wastage, the switch from low to high quality usually happens between the segments, leading to the time delay between the low-high quality switch of at least one segment duration. It is possible to make the segment duration shorter, though doing so we decrease the compression advantages we get from the encoding algorithm. This switch delay can significantly reduce the user's QoE and it is vital to take into account while following the bandwidth restrictions.

2.2 Measurement and analysis

Measurement of the viewer's perception is a complex problem. To provide such a measurement, one should introduce some quantity that gives us the ability to measure and compare it. This measurement should reflect the influence of a film content on the user's mind, and the results of the measurements should be reproducible and comparable. In literature, such measurement is usually called a metric, however, one should note that it is not a metric in a strict mathematical sense. For VR content, it makes sense to measure some physical characteristics of the human body that may change according to the mental state of the subject.

In general, the reactions of individuals strongly depend on their cultural and social background. Quan et al. have shown that for the simple models such as linear regression, the prediction accuracy of viewer's future position drops significantly even with the modest increase of the prediction interval [25]. That is why it is quite challenging to build a strict mathematical model when it comes to human behavioral patterns.

Despite of that, it is possible to extract and prove some statements about human behavioral patterns through statistical analysis and hypothesis testing.

The general procedure for performing hypothesis testing can be defined in the following way [26].

First, we define our problem through the null and alternative hypotheses. The null hypothesis is a statement that describes the default position regarding the tested subject [27]. For example, we can assume that the introduction of scene transitions in the omnidirectional video does not affect the user's exploration behavior through any measurable metrics. The alternative hypothesis is a statement that contradicts the null hypothesis. Following our example, an alternative hypothesis can be the increase or decrease in some parameters related to the user, such as the head motion speed, after the observation of the scene transition.

After the hypothesis was defined, the next step is to collect observations that would give us information on our theory. This collection of observations is called a **sample**. The sample is obtained from the **population** – a complete set of observations that can be made [26]. In our example, the population represents a complete set of all possible head-motion trajectories for videos with scene transitions, and the sample would be the collection of head-motion trajectories that were gathered by us throughout the testing. The observation can be defined as an absolute value of the angular velocity of the head, before and after scene transitions. When collecting observations, it is important to randomize the sampling procedure. Unrandomized sampling can lead to strong bias within

the test and can make the results much less robust. For example, different age groups could have different reaction times, making the analysis for one age group unsuitable for the other ones. Therefore, it is essential to note the restrictions of one's research and comprehend how the data was sampled. In some cases it can be convenient to split the population and test the hypothesis as two separate samples to see whether the different populations respond differently.

The next step is to choose the test statistic, that checks whether or not our hypothesis holds. The selection of the statistic depends on many parameters, such as the type of the measurements, the number of measurements in the sample, the type of the statistic we are testing for (for example, it can be the sample mean or its variance) and population's distribution parameters. In some cases, we use several test statistics, one to check the parameters of our sample data, and then with the knowledge of our parameters, we choose the other test statistic that would finally verify our hypothesis.

The hypothesis testing is based on the idea that the observation of unlikely events should be infrequent [27]. If this event is observed during a real experiment, the null hypothesis is wrong. It is up to the researcher to set the threshold that marks the event to be rare. In common practice, the event is regarded to be unusual when the probability of it is less than 5%. Of course, this threshold can be changed when we want to be more confident about the hypothesis. Test statistic measures the probability of the given event with regards to the null hypothesis. Based on this probability, we either reject the null hypothesis and accept the alternative hypothesis or state that we do not have enough data to reject the null hypothesis.

In the next section, we will look at the several test statistics, which will help us to perform hypothesis testing on the subject's viewing patterns.

In our case we would like to see if the exploration behavior of the subject differs before and after the scene transition, so we will focus on the test statistics that work on paired data. In paired tests, we test a hypothesis between two distributions, and each observation from one distribution has a corresponding observation in the other one. For scene transitions, the paired observation is the fixation some characteristics for each subject before and after the transition. The test is provided on the distribution of differences instead of comparing parameters between distributions.

Paired Student t-test is one of the most commonly used tests for comparing means. Usually, the hypothesis for the test is represented in the following form:

- H0: the mean difference between observations is equal to zero.

- H1: the mean difference between observations is not equal to zero.

The test statistic for the test has the following form:

$$t = \frac{\bar{d}}{\sqrt{s^2/n}}$$

Where $\bar{d}$ is the mean difference between observations, s^2 is the standard deviation of the sample, and n is the number of observations in the sample. Test statistic follows the t-test distribution and we can determine the rareness of the event from it.

Our data should follow several assumptions for performing paired Student t-test [28] :

- The data were randomly sampled from the population and gives a representation of that population.
- d_i follows a normal distribution.
- d_i are independent of each other.

In general, the t-test is robust to the second condition as long as the distribution of d_i is close to normal, and the sample size is big enough [29]. There have been several works that have studied the effectiveness of the t-test under non-normality conditions, which have shown that the increase of the sample size above 80 can mitigate the effects of skewness and flatness of the distribution [30].

One can assess the normality of the distribution using the Q-Q plot method [31]. The method plots quantiles of the given distribution against quantiles of the alternative distribution. In order to check normality, we can fit a continuous standard distribution curve using the maximum likelihood estimation and compare quantile of the original distribution and the fitted distribution. This method doesn't provide the quantitative measure for assessing the normality, but it gives a visual estimate which in a lot of cases is enough to determine if the normality condition is satisfied. For more quantitative result one can perform Lilliefors test which asseses whether the maximum discrepancy between the data and the fitted normal distribution is significant enough to conclude the non-normality of the data [32].

In case when the researcher expects or identifies an extreme non-normality of the data, specifically related to heavy tailness of the distribution, other more robust measures can be used such as truncated/winzorized mean [33] or median-based methods such as Wilcoxon signed-rank test [26, 34].

The result of the statistical analysis typically is given by a p-value, which provides the probability of collecting the given sample with the computed statistic under the null hypothesis. The p-value is then compared to the selected threshold. If the p-value is less than the chosen threshold, the null hypothesis is rejected.

In this work, we consider two types of scene transitions, and it is important to analyze whenever the user's behavior is different in these two separate cases. To solve this problem, one can use one of the methods from the family of two-sample location tests. These test compare Here we will present a short review of two-sample methods.

Let us assume that the samples of sizes n_1 and n_2 were collected from two normal distributions with known variances $\sigma_1$and σ_2 and means μ_1 and μ_2. In this case, variable

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

where $\overline{X_1}$ and $\overline{X_2}$ are sample means, follows a standard normal distribution [26]. The null hypothesis is then represented in the following form:

H0: $\mu_1 - \mu_2 = d_0$

In case when we want to check the equality of the means of the distributions, d_0 is equal to zero. To test the hypothesis, we compute the z-score for our two samples

$$z = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

and its p-value. The p-value is then compared to our selected threshold.

In the common case, the variance of the distributions is not known. Assuming that the variances of the given distributions are equal, we can use a two-sample Pooled t-test. We define t-test statistic t in the following way:

$$t = \frac{(\overline{x_1} - \overline{x_2})}{s_p\sqrt{1/n_1 + 1/n_2}}, \quad s_p = \sqrt{\frac{(n_1 - 1)s_{X_1} + (n_2 - 1)s_{X_2}}{n_1 + n_2 - 2}}$$

Where $s_{X_1}$and s_{X_2}represent the variance of the samples. We can find the p-value of the computed t-test statistic from the t-test distribution with $n_1 + n_2 - 2$ degrees of freedom and compare it with the chosen threshold.

In other situations, we cannot assume that variances of the distributions are equal. In this case, one can use Welch's t-test [35]. The equation below describes the statistic for the test:

$$t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{{s_1^2}/{n_1} + {s_2^2}/{n_2}}}$$

The statistic follows the t-distribution with approximate degrees of freedom equal to

$$\upsilon = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}$$

Just like the Paired Student t-test, the statistics become more robust to the violation of the normality assumption with the increase of sample size [36].

Statistical methods can also be used to pick atypical values from the data. Those methods fall under the category of outlier detection. They can be used to remove elements that might have been incorrectly recorded or to distinguish some unusual behavior of the data. For example, outlier detection methods can be used to identify unusual head motions during the test. The methods vary in complexity, depending on our prior knowledge about the data. In the case when we know the distribution our data adheres to, the outlier points can be defined through the statistical moments and quantiles of the distribution. For example, one can pick outliers by checking if the value falls inside a specified interval inside the distribution. The interval cab defined through the inter-quantile range of the distribution or by expressing it as an area around the mean/median of the distribution [33]. In the case when the distribution is unknown, more sophisticated methods are used, usually based on clustering, density-based, or other [37].

The statistical approach provides powerful and robust methods to assess the hypothesis, but it is crucial to understand the limitations of each test and use the right tool in the right situation.

2.3 Film transitions

Film transitions describe a set of methods of connecting one scene of the movie to another. Conventional cinematography follows a set of rules that were invented in the early 20th century and provide comfortable film viewing. One of the main approaches used in modern cinematography is *continuity editing* [38]. By editing movies using this set of rules, the viewers perceive the content as a continuous flow, even though it can change temporally and spatially across the scene. For example, a technique called **insert shot**

can be used to change the scale of the frame without breaking the continuation perception for the viewer [39]. This shot takes into account how the size of the objects should change between cuts. The general rule states that the camera position can change only by one-two orders in scale. In this way the change becomes more natural for the viewer and is better at capturing his/her attention [39]. Different scales of the scene can be seen in Figure 3.

One of the most significant differences between VR and traditional cinema editing is the fact that generally traditional cinema directors employ the montage techniques to enhance the emotional response of the viewers and redirect the viewer's focus to the object of the scene. In case of the VR editing and cinematography, the creator should also consider how the montage affects user's perception inside the 3D space. Techniques like color adjustment, depth of field blurring or tone mapping in the stereoscopic content can lead to unpleasant viewing experience if done independently for the left and right images of the stereoscopic view. Depth perception and video stabilization also play an important role in the editing of the VR movie [40].

In traditional cinema, the duration of the shot in recent years is has been shortened in comparison with to how movies were edited in the previous century. On the other hand, the average shot length in VR can be 20 seconds or longer [7]. The rationale for this increase may be fact that the area that the user can observe in virtual reality is much larger than the screen used in the traditional cinema, thus it is reasonable to give the user more time to explore the environment.

The classification of editing tricks and scene cutting techniques is widely developed in conventional cinematography. Scene transitions are used to either create a continuous action filling or to break the continuity and transfer the user to the next episode. Scene transitions can be split into two main groups: transitions that conserve the continuity of perception between the scenes, and transitions that deliberately break that continuity. Continuity between the scenes can be defined within three axes: space (camera position), time, and action. According to the recent studies in psychological science, human perception consists of a chain of discrete events [41]. An event is a sequence of episodes that are consistent in terms of space, time, action, from which a human can convincingly predict the next episode from the previous set of episodes. If the next episode is unpredictable, then the previous set of episodes is stored in the long-term memory as a single event [9].

Figure 3. *A scene from the movie 'Little Caesar' represents different types of shots*

Modern editing techniques in cinema imitate this peculiarity of human perception. In general, scene transitions can be divided into three classes:

- Edits that are discontinuous in time, space, and action (action discontinuities).
- Edits that are discontinuous in time, space but continuous in action (spatial/temporal discontinuities).
- Edits that are continuous in time, space, and action (continuity edits).

One of the differences between VR setup and conventional cinematography is that the edits of the third type are transferred to the user since the user has control over the viewing direction. Therefore, our study focuses on the first two types of the scene cuts.

The way the transition effect is done depends on the director's vision of the scene. In some cases, the transitions can have some additional symbolic meaning, such as the passing of time, the similarity between the characters or other linked narratives. In general, the scenes that are continuous in action (continuity edits and spatial/temporal discontinuities) are connected with simple scene cuts, so that the change is less perceivable by the user. In case of action discontinuities, we usually would like to emphasize the start of the new scene, and therefore, additional effects such as Fade, Dissolve, or Swipe can

be used. In our work, we focused on the scene cuts since it is the most common method for connecting two scenes. For works concerning the effects of different scene transitions in VR, see [12, 42].

3. RESEARCH METHODOLOGY AND MATERIALS

3.1 Experimental setup and data collection procedure

Data collection procedure was performed in the following manner:

- Find and prepare visual stimuli for the tests.
- Enlist test subjects, complete pre-test fatigue questionnaire.
- Run test session, save observed trajectories in a convenient format, conduct post-test fatigue questionnaire.

The visual stimuli were divided into two parts: visual stimulus for **inter-scene** transitions (changes between two completely different scenes) and visual stimulus for **intra-scene** transitions (between the same scene, but different camera positions). For the first part, we selected eight clips presented in a Table 1.

Table 1. *Content used for the Inter-scene transitions experiment*

Title (Youtube)	Offset (sec.)	Href
360° Wife Carrying I 100 Moods From Finland	5	https://www.youtube.com/watch?v=E2dOKkg0ozY
360° Barn Dance I 100 Moods From Finland	5	https://www.youtube.com/watch?v=skUHYb7FqK4
GoPro Fusion Spatial Audio Demo - Skiing	25	https://www.youtube.com/watch?v=nOoWRhp9ZiA
String quartet in Turku - 360 - Spatial Audio	180	https://www.youtube.com/watch?v=gkh4yW3WVn0
VR 360 Wildfire Roller Coaster Onride POV Silver Dollar City Branson MO	45	https://www.youtube.com/watch?v=jBHTKOtGDZU
Survive a Bear Attack in VR	25	https://www.youtube.com/watch?v=g7btxyIbQQ0
See aircraft carrier jet launches, air operations for the first time in 3D VR	25	https://www.youtube.com/watch?v=H4Q0RLoeyuY
Pole vault	0	ftp://ftp.ient.rwth-aachen.de/testsequences/testset360

The clips were edited to be 20 seconds long, with starting offset defined in the Table 1 for capturing the most interesting section of the given clip. To make the quality between clips consistent, they were edited to the following specifications: Video resolution –

3840x1920 and the framerate – 30 FPS. After the clips modification, they were concatenated one after another with no additional effects added, thus setting the instantaneous transitions between the clips (scene cuts).

For the second experiment, we wanted to study the transitions that happened within the same semantic content, i.e., **intra-scene** transitions. Three different clips were selected from YouTube for that experiment. The clip details can be seen in the Table 2.

Table 2. *Content used for the Intra-scene transitions experiment*

Name (Youtube)	Duration (sec.)	Short name used in thesis	Href (Youtube)
Aunkai Bujutsu, a martial art based on body awareness and the basics	62	Martial	/watch?v=cUkrXfFzFcQ
Armored Combat League NYC VR 360	66	Armor	/watch?v=Zsl2bX7UT7g
Lions 360° | National Geographic	76	Lions	/watch?v=sPyAQQklc1s

The clips consisted of several scenes filmed in the same area. The transitions for some of the clips were not instantaneous. Therefore, the transition effects were edited out. Some scenes from the 'Lions' clip were deemed to be too gruesome for viewing, so we also removed them from the final version. The video resolution and frame rate were set as in the previous clip collection. The period between the scene cuts was roughly equal to 10 seconds, and each clip contained four scene transitions.

The Mean Spatial and Temporal Perceptual Information [43] characteristics of the clips can be seen in the Figure 4. Spatial perceptual information is defined as the maximum standard deviation over pixels for Sobel-filtered frames. Spatial Information shows how much information (detail) frames in the clip contain on average individually. Temporal perceptual information is computed as the maximum of the difference between the intensities of two adjacent frames. Temporal information shows the rate of change between the subsequent frames or how dynamic the content of the clip is. The calculations were done using a Python script, based on the description from the ITU-T recommendations [43, 44].

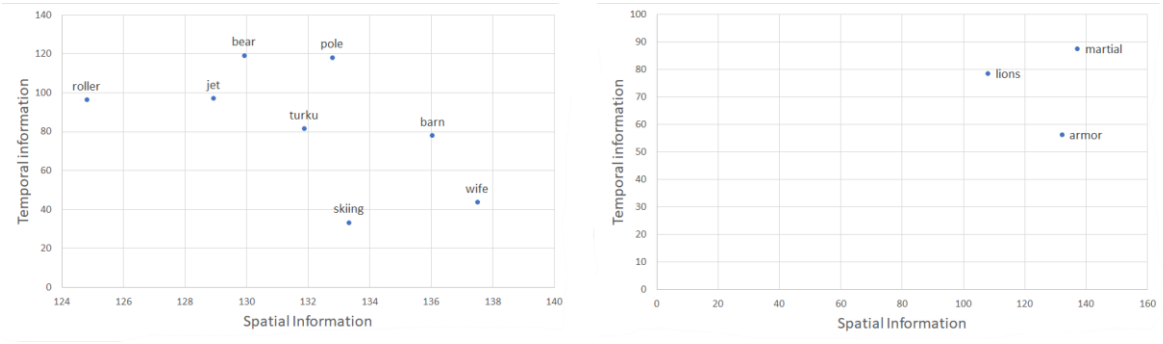

Figure 4. *Spatial and Temporal Perceptual information of the clips. Left – characteristics of the clips for inter-scene transitions, right - characteristics of the clips for intra-scene transitions. The spatial information values are similar across clips, while the temporal information differs from clip to clip.*

Subsequently, the clips were converted to the HEVC DASH-streamable format with the use of the MP4Box and FFMPEG tools. We used the **viewport-independent** streaming model, i.e., the quality of the video did not depend on the position of the viewport. Motion-to-photon delay in the **viewport-dependent** streaming model is influenced by encoding parameters such as the number of tiles and the DASH segment duration, and this delay can interfere with the way how the user explore the content. In our work, we wanted to analyse the natural reactions to scene cuts with minimal additional interference. Therefore, an independent model was chosen for these experiments.

Our test system was analogous to a typical streaming system, as described in the prior art (see Figure 5). The system consisted of two components: the server and the client. The server stored streamable content with the MPD file, as well as the viewed trajectories. We implemented a simple manager for clip selection and start/pause controls. The client requested and rendered the content based on the viewer's current orientation. We used the combination of Oculus Gear VR HMD and Samsung Galaxy 8 phone for our head-mounted display, in which the smartphone also worked as a client. The content was transferred to the phone via a wireless network, reducing the number of cables needed. The viewing orientation was collected based on the smartphone's gyro sensor data. Data points were sampled at a frequency of 50 **Hz**. The orientation data was saved in the yaw-pitch-roll-timestamp format. Yaw values were bounded in the range [-180, 180]; pitch and roll values were bounded in the range [-90,90]. The timestamp values were represented as a time interval between the start of the clip and the moment the data point was sampled, calculated in milliseconds.

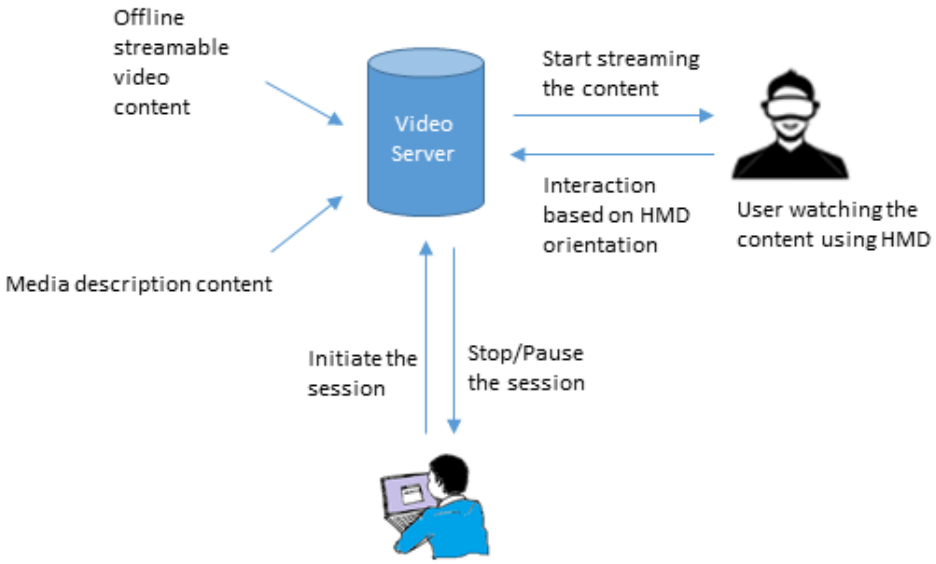

Figure 5. *The test setup used in the experiments*

After the clips were prepared, we selected 20 people to participate in our test session,17 males, 3 females. Before the watching session, we tuned the inter-pupillary distance of the head-mounted display to accommodate it for each test subject. We also conducted a pre-fatigue questionnaire to check whether any possible side effects could impede the user's viewing trajectory. The survey was based on [44]. We verified the levels of pain in shoulders, difficulty in concentration, nausea, sleepiness, dizziness, stabbing pain and fatigue in eyes, stiffness of the neck, pain in the back, forehead or temples, difficulties with focusing, dry or watering eyes, feeling that the eyes are looking in different directions, double vision. The user was informed that he was allowed to close his eyes during the watching session to re-establish a clear vision, if needed.

The users were not restricted in how they should explore the content. During the watching session, the user resided in a rotating chair, so the change of the viewing orientation could occur due to either head or chair rotation. The users could take a small break between clips if needed. After the viewing session was finished, we conducted a post-test questionnaire to see how this test affected the users. The trajectory data was saved into a database and could be extracted into the Excel format for further analysis.

3.2 Research methodology

The collected data was imported to Python as a dictionary for further analysis. The structure of the dictionary is shown in Figure 6. The study was done with the help of the Python's scientific computing libraries, NumPy and SciPy.

The analysis was done only on Yaw values for the following reasons. According to [11], the user retains strong equatorial bias while watching omnidirectional videos. Additionally, the chosen videos also distribute the salient regions of the clip along the equator, so the most significant impact in the trajectory should happen along the yaw axis. With these details taken into consideration, we decided only to use yaw values. Pitch values assumed to be close to the equator.

The user's exploration behavior can differ significantly from person to person. Ideally, we wanted our users to explore similar content in a scene, so the factors that defined their behavior would be comparable. For example, if the clip has a fast-moving object in one part of the scene and some static landscape in the other, the user's exploration behavior would be different depending on what part of the scene was observed. To remove users who viewed content atypically on average, we used the Similarity Ring Metric [45].

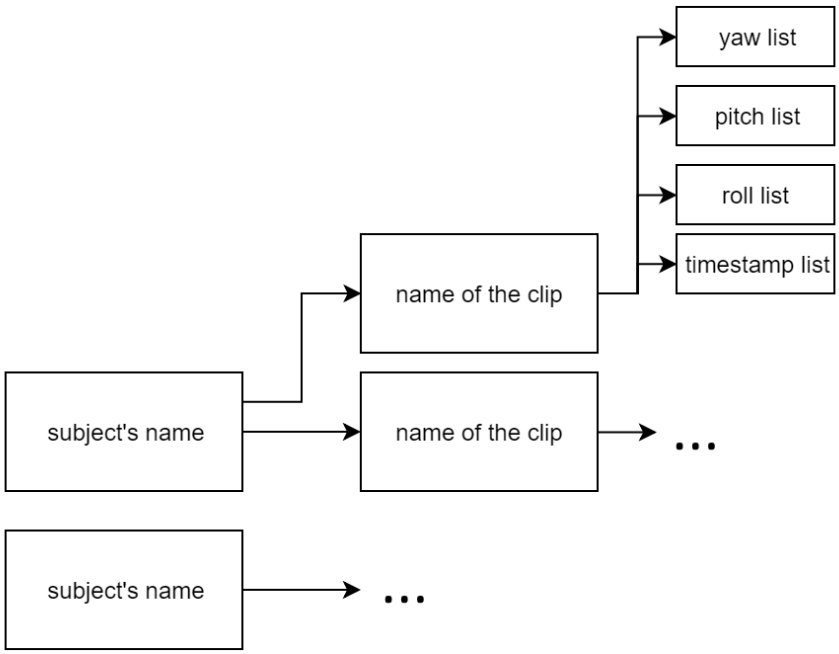

Figure 6. *Structure of collected data*

First, we computed the typical trajectory by taking a median of yaw values from all the users within the given timestamp. Then, we bounded a 'ring' around each point with a diameter equal to the horizontal FOV of the head-mounted device, which in our case was equal to 100°. If the user was inside the ring for the given timestamp, at least 50% of the

content he explored intersected with the content explored by watching it within the typical trajectory. Then we computed the percentage of time the user was inside the ring by consecutively checking each point from the given user's path. By setting the threshold on that percentage, we could either accept or reject the user's trajectory as an outlier. By adjusting the ring diameter and 'time within the ring' threshold, we could make the outlier rejection procedure stricter or more relaxed. In our work, we set the limit to be equal to 50% of the time being inside the ring. The visual representation of the procedure can be seen in the Figure 7.

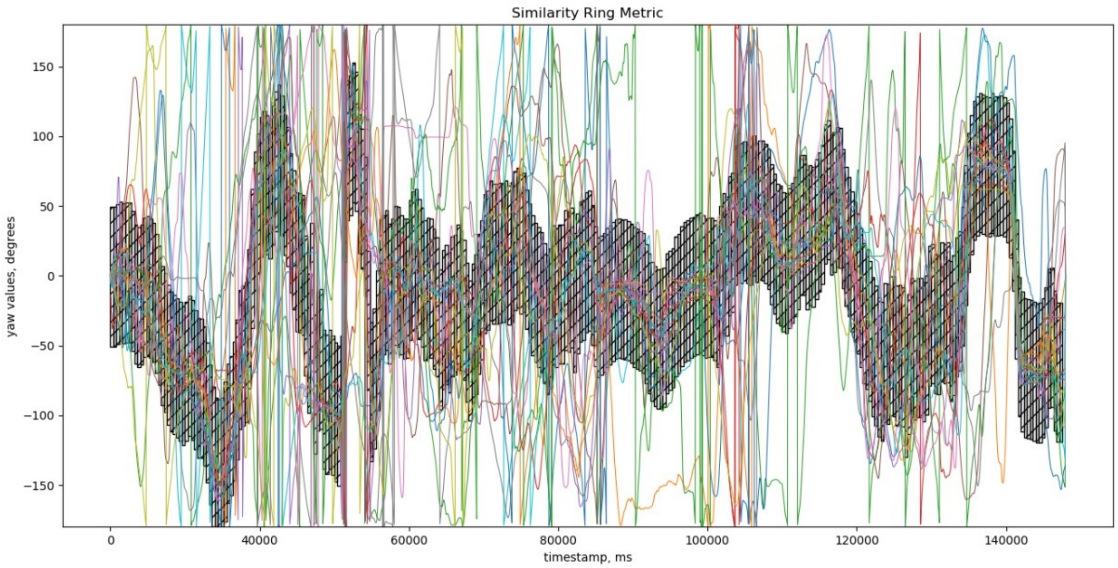

Figure 7. *Similarity Ring Measure for inter-scene transition experiment*

In our work we wanted to define the metrics that would show the change in viewing behavior without associating it within the specific viewed content. So, our analysis was focused on the changes in the **Exploration Range** ***(ER)*** and the magnitude of angular velocity during the content exploration.

The exploration Range defines the maximum angular distance the user traversed during a specified time interval. To compute ER, we divided data into small intervals of specified duration, found the maximum and minimum yaw values in the given interval, and calculated the difference between them.

$$ER_{t_k} = \mathrm{Max}\left(Yaw_{t_k}, Yaw_{t_k+dt}, \dots Yaw_{t_k+\Delta t}\right) - \mathrm{Min}\left(Yaw_{t_k}, Yaw_{t_k+dt}, \dots Yaw_{t_k+\Delta t}\right),$$

Where dt – sampling period, and Δt – the duration of the selected interval. In our experiments, we chose the interval with a duration of 250ms, which for our data was equivalent to taking five consecutive samples.

The next characteristic we observed was the magnitude of angular speed. The angular speed measures the speed within which the users explore the given content. Our setup

did not allow direct measurements of angular speed. Instead, we approximated the metric by computing derivatives of the angular values using finite differences. Due to the noise in the measurements of the user's position and the fact that differentiation emphasizes high frequencies of the original signal and therefore increases the noise in the derivative signal, it was essential to perform a smoothing procedure before computing the angular speed. In our work, we used the Savitzky-Golay filter [46], that performs least-squares polynomial approximation within an interval around each data point. The window size of the filter and the order of fitted polynomial was chosen to be 16 and 3, respectively. To compute the magnitude, we took the absolute values of the signal derivative. Below, the metrics used in the experiment will be compared between each other.

For the Exploration Range, it is necessary to introduce the duration of the interval within which we collect values for a single exploration range sample. This makes the metric less generalized since different interval sizes would lead to different results. On the other hand, the metric is more accurate for the viewport displacement computing in a case when the user makes fast head movements in a small vision cone. Another advantage is that this metric more accurately represents the user's region of interest displacement in case of larger intervals. This makes it perfect for comparing the change in viewers' orientation concerning DASH segment length and size of the video tile. The angular-speed metric is independent of the duration interval, making results more universal.

For the chosen metrics, the change of the yaw values in time is more important than the absolute value of them. In our case, the yaw data was bounded in the interval [-180, 180]. It should be mentioned that points 180 and -180 correspond to the same single point (cut point), and a small change in the position around point 180 or point -180 may lead to a significant jump in metric values. This effect arises because the yaw values are represented on a circle, points 180 and -180 of which are stitched together. To avoid the jump when the user crossed those points, we projected the points from the circle to the regular number line using the following procedure. We iterated over yaw values and computed the difference between current and previous values. If this difference was more significant than a predefined threshold (the threshold was set to 320 degrees), we assumed that the user went over the stitch point. In this case, we shifted the point so that the resulting curve becomes smooth. This procedure is illustrated in the Figure 8.

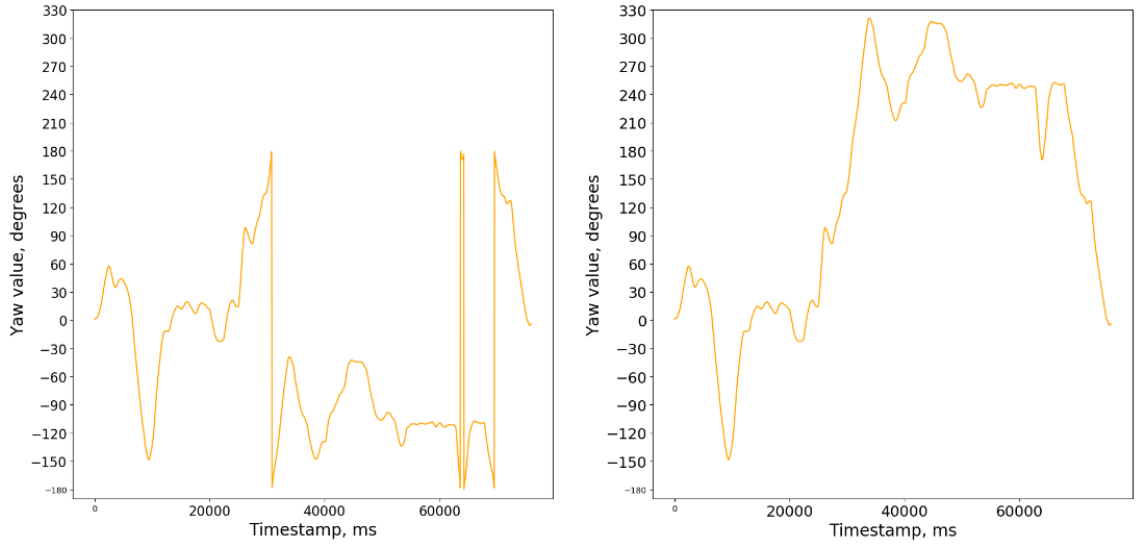

Figure 8. *Yaw unrolling procedure. Left – original yaw values, right – unrolled version*

The next characteristic we wanted to study was the reaction time of the user during a scene cut. Close inspection observing the above-described metrics showed that the change in the behavior did not happen instantaneously but with a small delay after the start of the transition. This delay between the end of transition and the start of the abnormal watch behavior is called **reaction time**. We derived the following procedure for determining the reaction time delay:

First, we computed the histogram of exploration ranges for each user. From the inspection of those histograms, we concluded that exponential distribution could be a good fit for the given exploration range data (example in Figure 9).

In our study, we have assumed that the behaviour that the user exhibits after the scene cut is atypical from the normal behavior in terms of the metrics defined. We can consider of the problem of detecting the start of post-scene-cut exploration as an outlier detection problem, i.e., we want to find the timestamp after the scene transition where the user's exploration range would differ significantly from the typical values. Therefore, we needed to define some threshold for the exploration range values, at which we assume the user's behavior to be abnormal. One of the ways to establish such limit is by using the Tukey criterion [47]. This criterion defines the outlier threshold in terms of inter-quartile distance of the distribution. The threshold can be expressed as

$$threshold = Q_3 + \lambda(Q_3 - Q_1)$$

Where Q_1 and Q_3 are the first and third quantile of the observed distribution, and λ is a scalar, which defines the strictness of the outlier test. Usually, λ is set to be 1.5 or 3; in our case, we have chosen it to be equal to 1.5.

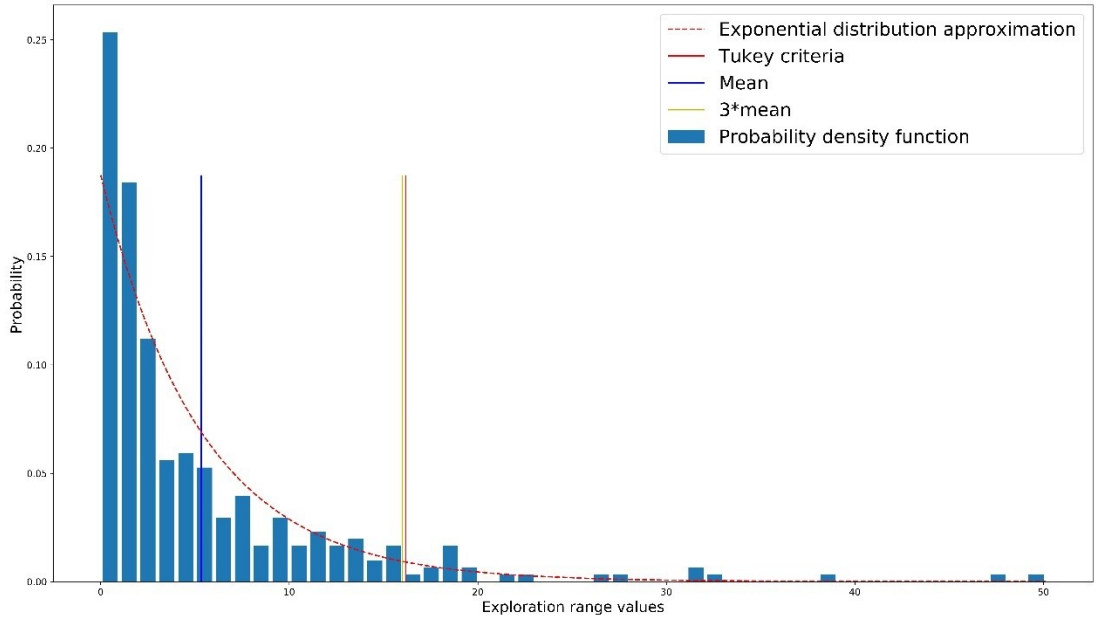

Figure 9. *Example of exponential distribution fit*

Assuming that the user's exploration range distribution can be approximated with the exponential distribution, we can simplify the criteria by expressing threshold using the mean of the distribution:

$$Q_1 = \log\frac{4}{3} \cdot mean, \;\; Q_3 = \log 4 \cdot mean$$

$$threshold = Q_3 + 1.5 \cdot (Q_3 - Q_1) =$$

$$\log 4 \cdot mean + 1.5 \cdot (\log 4 \cdot mean - (\log 4 - \log 3) \cdot mean) \;\; =$$

$$(\log 4 + 1.5 \cdot \log 3) \cdot mean \;\; \approx 3 \cdot mean$$

Therefore, we can assume that the exploration range values that are **3** times larger than the mean exploration range of the user are the sign of the abnormal exploration behavior of the user.

The reaction time was computed using the following procedure:

- For each subject in a fixed clip
    1. Compute the average exploration range of the user
    2. For each scene cut point in the clip
        - Collect the next *K* points of the exploration range data
        - Find the first value *V* that is three times larger than the average exploration range of the given user
        - If such value does not exist, take a local maximum of those points.

- The reaction time for the given scene cut and the user is equal to a duration of a time interval between the start of a new scene and the point *V*.

3. Average the user's reaction time over scene cuts of the clip using median or mean.

- Average the reaction time over users to compute the reaction time of the clip.

The value *K* was selected to be 16, that is equivalent to the next 4 seconds after the scene cut in the case when the exploration range is computed over intervals of 250ms.

The visual representation of the algorithm can be seen in Figure 10.

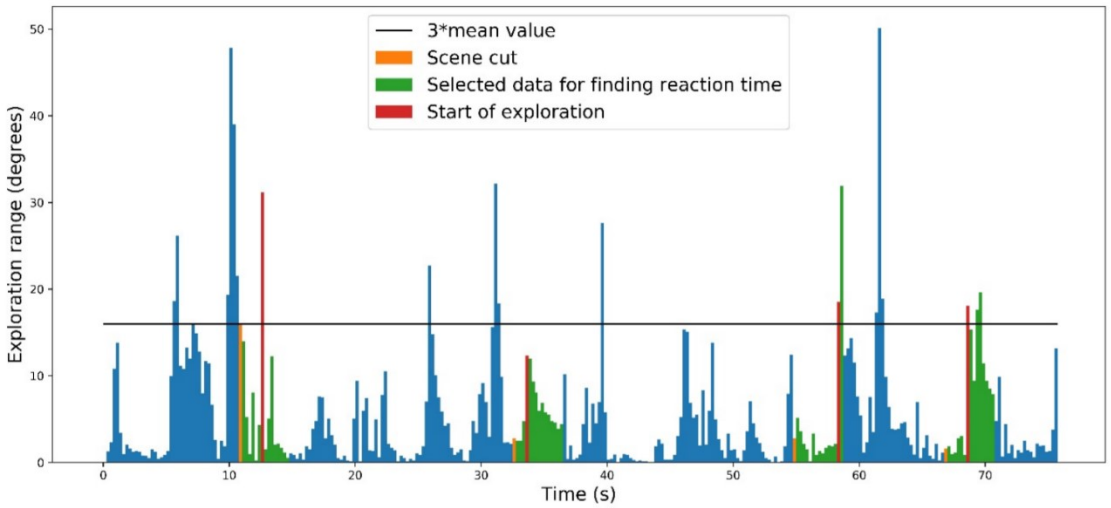

Figure 10. *Reaction Time computation for one of the users*

4. RESULTS

4.1 Analysis of inter-scene transitions

The **inter-scene** transition analysis was performed on the data, which had a variety of different clips concatenated together. First, we completed the SRM outlier procedure, as described in section 3.2. The individual scores of each test subject can be observed in the Figure 11.

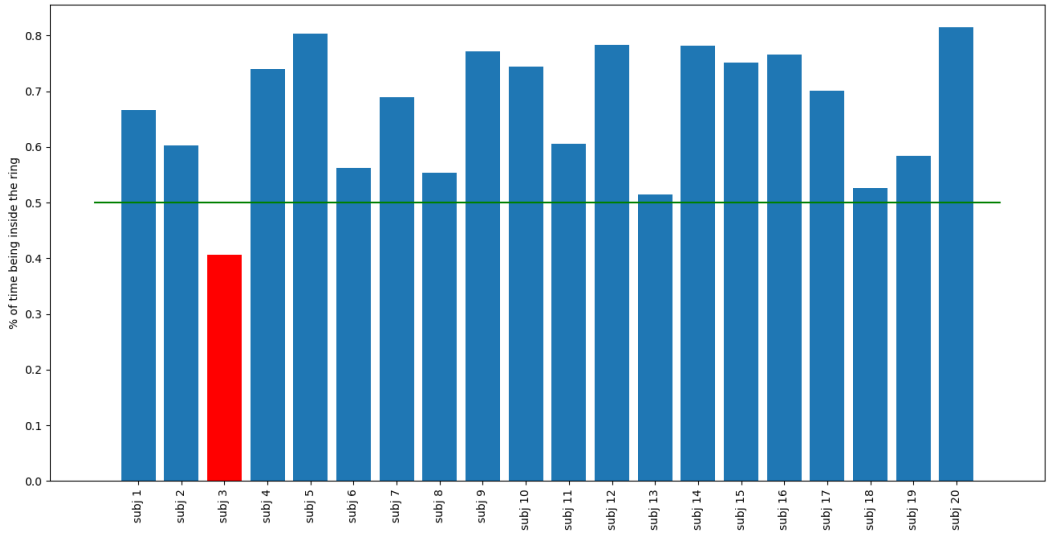

Figure 11. *SRM scores for test subjects in the inter-scene transition experiment. One subject was excluded.*

One test subject retained the score of 40.6%, that was less than the set threshold, and therefore the user was excluded from the subsequent analysis.

Next, we computed Exploration Range statistics for each user. Average values of the Exploration Range can be seen in the Figure 11.

From Figure 12, one can see an increase in the metric after the scene cut (scene cuts are marked as orange bars). On average, the rise in Exploration Range was equal to 56% relative to the average Exploration Range of the clip. It should be mentioned that this increase varies and depends on the content semantics. For example, the second scene cut was masked by a peak, due to the rapid movement of an area of interest in the frame before the end of the second clip in the sequence. The movement would result in the viewer's head re-orientation and, therefore, creating a local maximum.

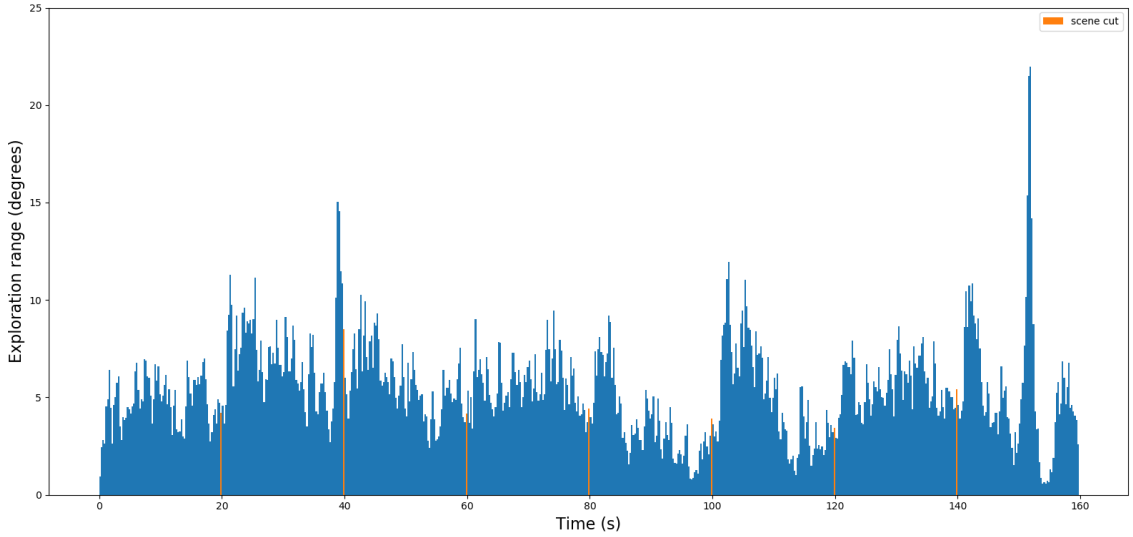

Figure 12. *Average ER values for the inter-scene transition experiment*

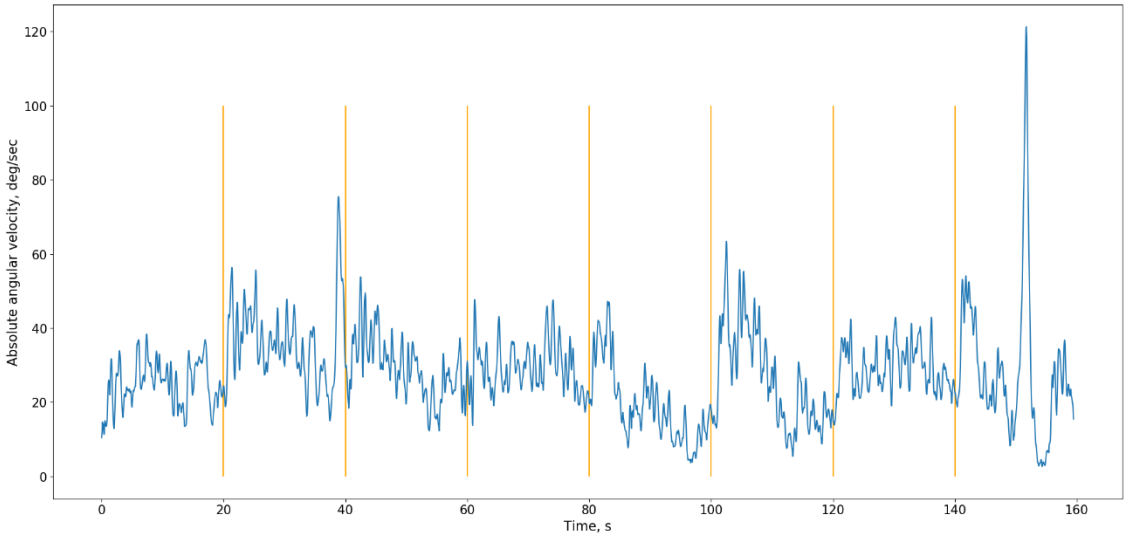

Figure 13. *Average angular velocity values for the inter-scene transition experiment*

Next, we analyzed the changes in angular velocity after the scene cuts for the given sequence. The angular velocity represents the speed within which the user rotates the head during the viewing session. From Figure 13 we observed some similar patterns, as we see in the Exploration Range plot. The average increase in the angular speed during the scene cut is equal to 5.5%.

We also computed the average reaction time after the scene cut, as described in section 3.2. The reaction time was calculated for all the test subjects individually and then averaged, resulting in the mean reaction time equal to 1800ms. During this period, there is an opportunity to tune the parameters of the streaming setup in such way that it would mitigate the unfavourable effects of the user's movement after the scene cut.

We performed hypothesis testing to see if the hypothesis of increased exploration would have a strong support under our data. To check the hypothesis, we chose the Student t-test statistic based on the mean Exploration Range before and after the scene cut. Before performing the test, we checked whether our sample distribution followed the normality assumption. The data was obtained in the following manner: we calculated Exploration Range values during the time interval of two seconds just before the scene cut and averaged them. This computed value represented the Exploration Range before the scene cut. The appropriate dataset after the scene cut is averaged over a two-second interval that starts one second after the scene cut (to take into account the reaction time of the subjects). In our pair test, we used the differences between the two datasets mentioned above (before and after the scene cut) to obtain an independently sampled dataset.

The normality assumption was checked using of QQ plot. From Figure 14, one can see that our sample closely follows the normality assumption.

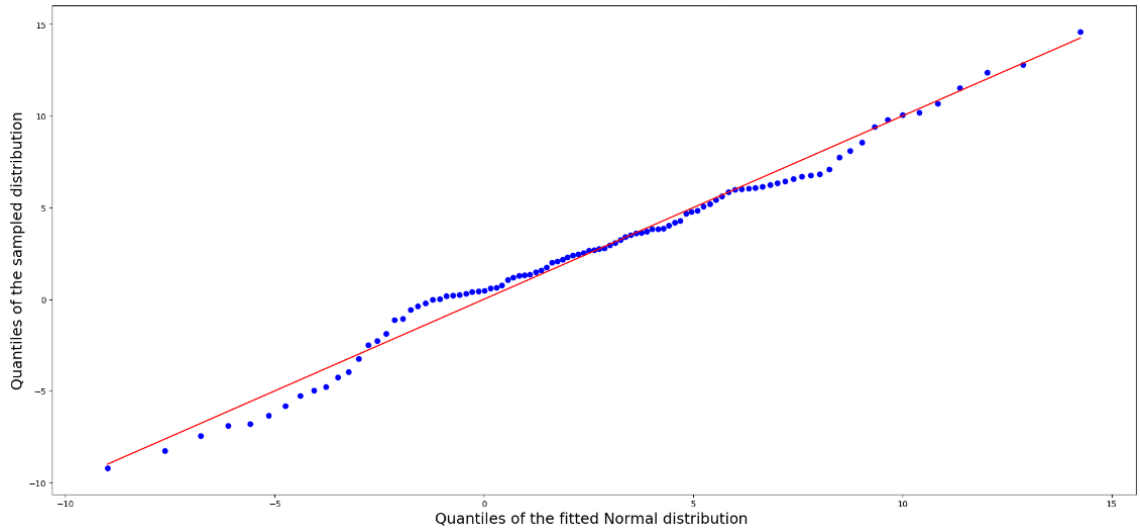

Figure 14. *Quantile-quantile plot of difference distribution for inter-scene transition experiment*

P-value obtained from student t-test was equal to $6.2 \cdot 10^{-7}$ %. This means that the probability getting this set of data under the condition that nothing changes after the scene

cut is almost equal to zero, so we can state with a high degree of certainty that there is a difference in user's exploration behavior before and after the scene cut.

4.2 Analysis of intra-scene transitions

In the case of **intra-scene** transitions, three clips were prepared. Every clip included approximately four scene cuts. Similar to the previous experiment, we performed the SRM outlier procedure. For the 'Armor' and 'Lion' clips, no test subjects were excluded from the analysis. The average SRM score for the clips show that in general trajectories were quite similar to each other, as seen in Table 3. Two test subjects were excluded for the 'Martial' clip since their trajectory differed drastically from the other ones. The SRM values can be seen in Figure 15.

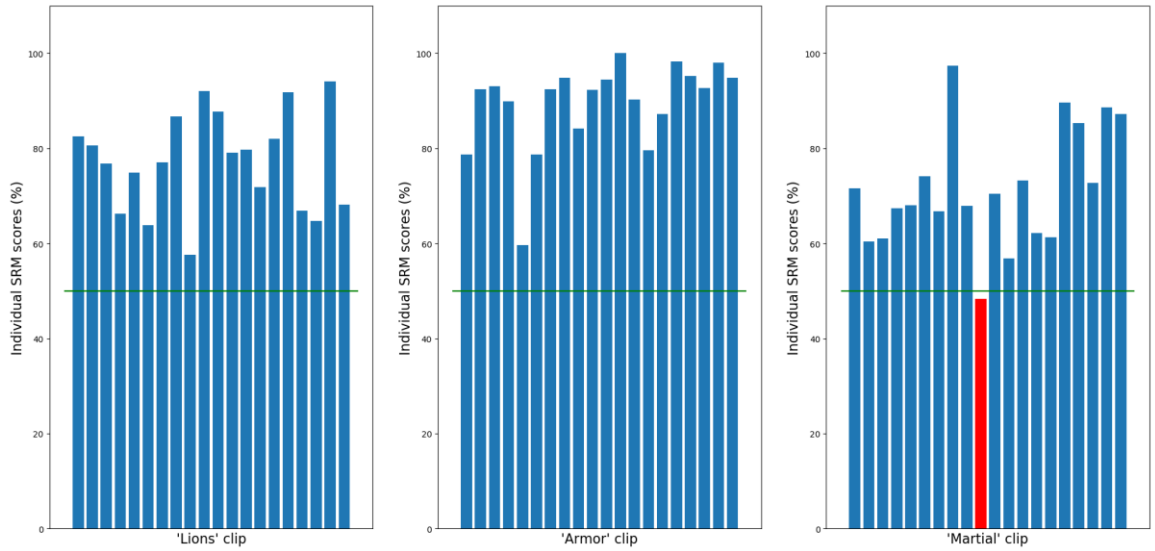

Figure 15. *Similarity Ring metric for 'Lions', 'Armor' and 'Martial' clip. The red color represents the excluded subjects*

Table 3. *SRM scores for clips in intra-scene transition experiment*

Clip	SRM score, %
Armor	0.92
Lions	0.88
Martial	0.80

The analysis of the Exploration Range for **intra-scene** transitions has shown similar user's exploration behavior during scene cuts, as in the **inter-scene** transition experiment. The increase in Exploration Range varied from 28% to 76%. The Figures 16-18 below show the average Exploration Range for 'Armor', 'Martial', and 'Lions' clips respectively.

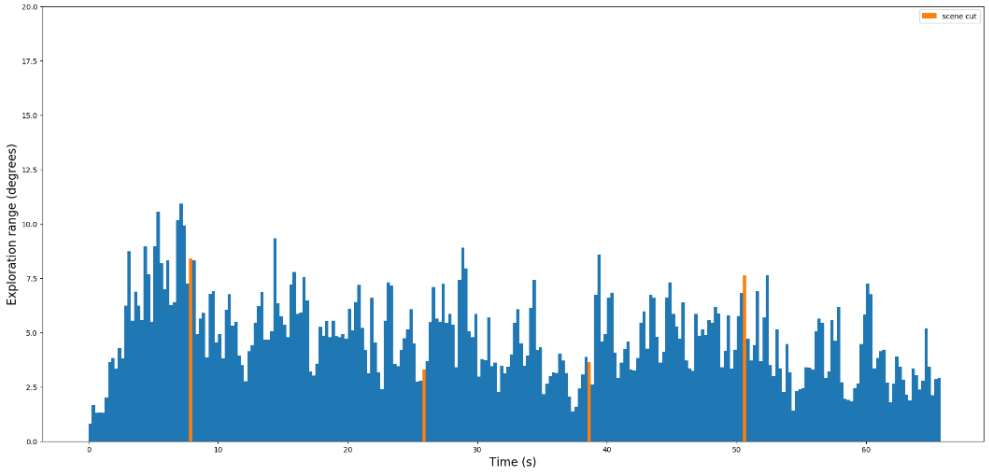

Figure 16. *Average ER values for 'Armor' clip*

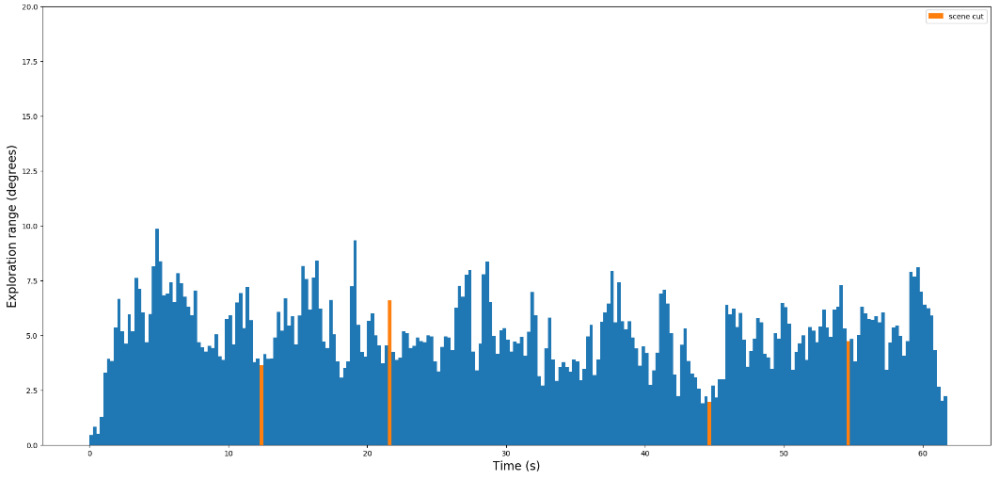

Figure 17. *Average ER values for 'Martial' clip*

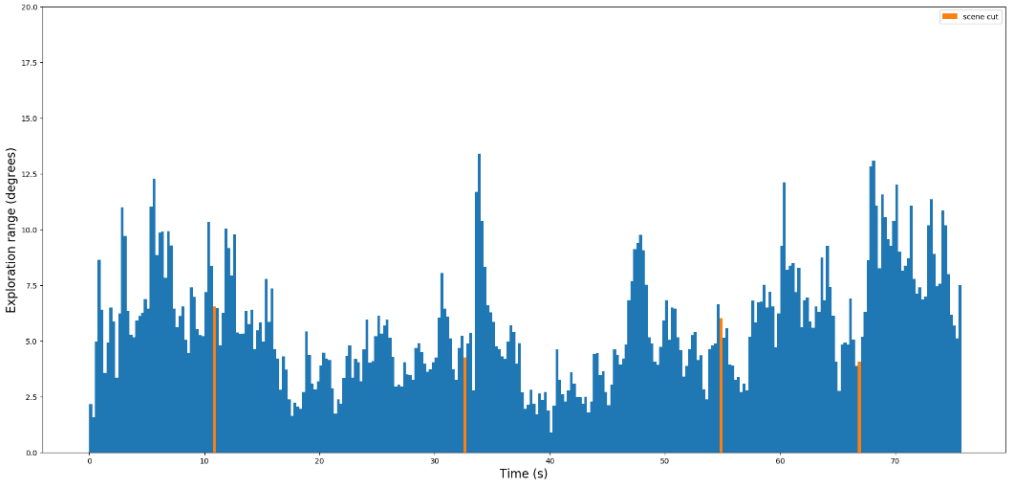

Figure 18. *Average ER values for 'Lions' clip*

The angular velocity plots also show similar patterns as in the **inter-scene** transition experiment. The plots can be seen in Figures 19-21.

The average increase in Exploration speed after the scene cut is equal to 4.4% relative to exploration speed before the cut. We have found out that the drastic content change results in the significant metric change no matter whether the content switches with the **intra-scene** transition or changes using the **inter-scene** transition. For example, in the 'Lions' clip, the scene cuts happen in the same semantic content, but the increase in Exploration Range was comparable to the results in the **inter-scene** transition experiment due to a pronounced distinction in camera location for that clip.

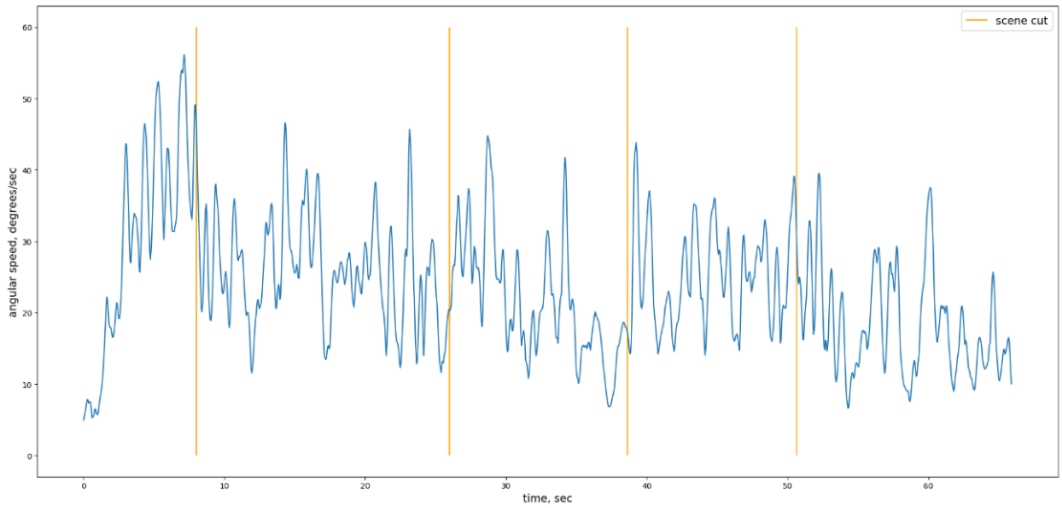

Figure 19. *Average Absolute Angular Velocity values for 'Armor' clip*

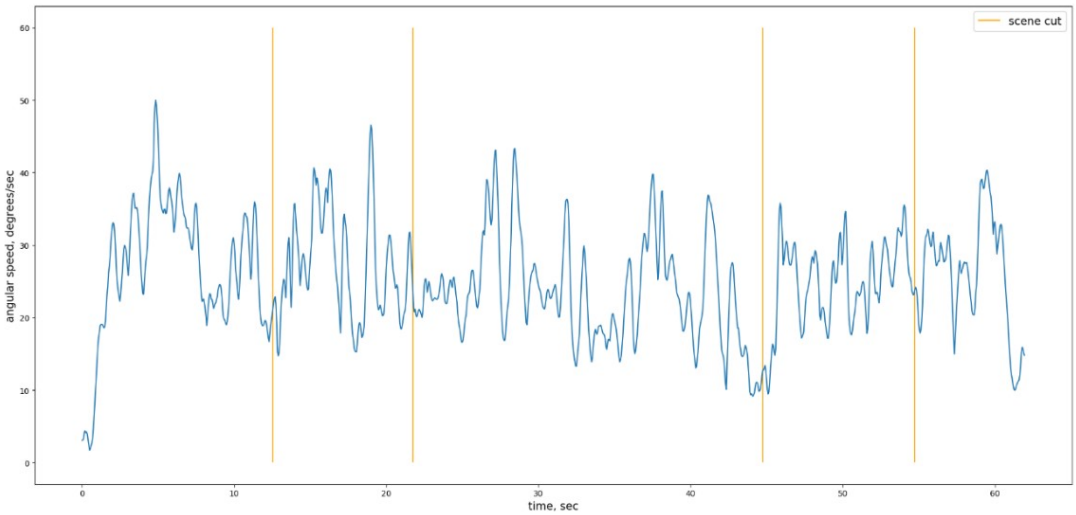

Figure 20. *Average Absolute Angular Velocity values for 'Martial' clip*

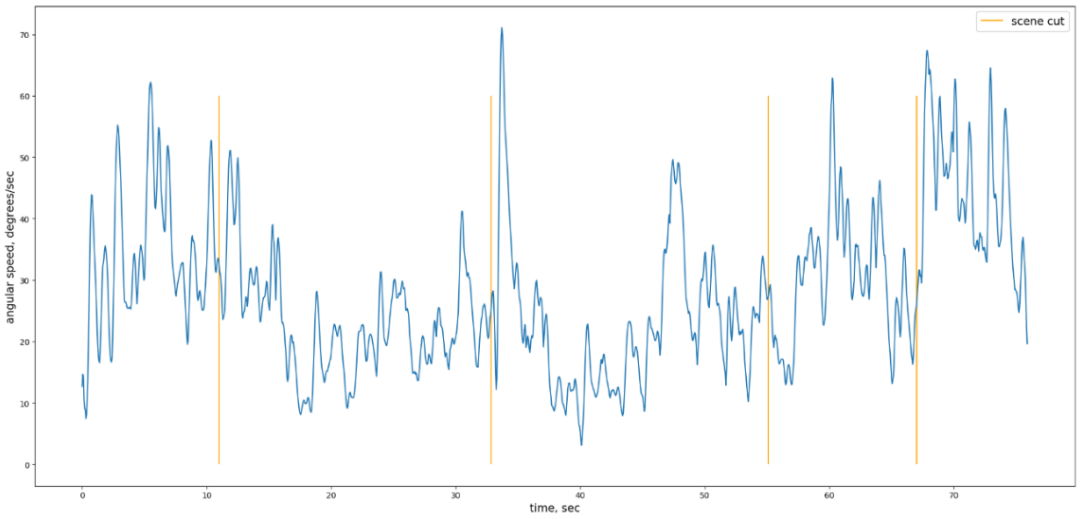

Figure 21. *Average Absolute Angular Velocity values for 'Lions' clip*

The mean reaction time has been computed for each clip, and it appears that the reaction time decreases as the content gets more diverse between scene cuts. The average reaction times for each clip can be seen in the Table 4.

Table 4. *Reaction time for clips in the inter-scene transition experiment*

Video	Mean Reaction Time
Armor	1500
Lions	1400
Martial	1800

We have provided the same statistical tests for the **inter-scene** transition experiment. One can see that the tails of the sampled distribution are heavier than fitted the normal distribution. In this case, the normality assumption might not take place, and the Student t-test may lead to the wrong result. Therefore, we have also computed the Wilcoxon signed-rank test statistic, which is more robust to the violation of the normality assumption. The QQ plot can be seen in the Figure 22.

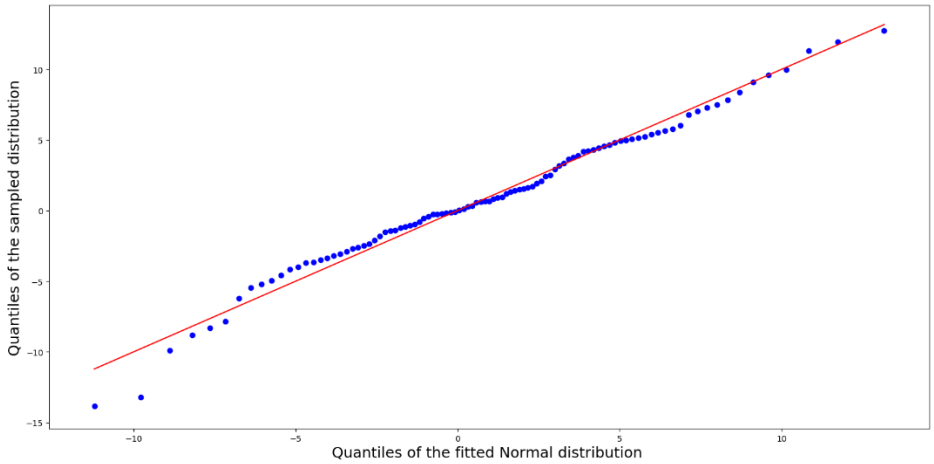

Figure 22. *Quantile-quantile plot of difference distribution for intra-scene transition experiment*

The p-values for Student t-test and Wilcoxon signed-rank test statistics are equal to 0.5% and 0.1%, respectively. The test shows that the **intra-scene** transition forces the users to change their exploration behavior. Though it should be noted although both intra- and inter- scene transitions lead to a change in the user's behavior; the statistical significance of the **inter-scene** transition effect is six orders greater than one of the **intra-scene** transition.

4.3 Comparison between inter- and intra- scene transitions

We have acquired the following results:

- We have shown that in case with the **intra-scene** transition, the Exploration Range and Angular velocity of a user after the scene cut record a small change due to the similarity of the scene content before and after scene cut.
- The content switch between dissimilar scenes would result in the change of Exploration Range and angular velocity, regardless of intra- or inter- scene transition.
- There is a complex relationship between the reaction time, type of transition, and the content itself. Nevertheless, the reaction time for the **intra-scene** transition is on average higher than the reaction time for the inter-scene transition.

The comparison for average Exploration range values in clips for the inter-scene transition and for the intra scene transition can be seen in the Table 5. The average relative increase in Exploration Range is equal to 77%.

Table 5. *Comparison between average exploration range during scene cuts and during normal viewing*

Name	Avg. exploration range in degrees (without scene cuts)	Avg. exploration range in degrees (with scene cuts)
Video cut (Inter-scene)	4.3	6.7
Armor (Intra-scene)	3.4	5.1
Lions (Intra-scene)	3.7	6.9
Martial (Intra-scene)	3.6	5.0

We have checked the statistical significance of the increase in the Exploration range between inter- and **intra-scene** transitions. Since the variance of the population is unknown, and there is a doubt about the equality of variances for inter- and intra-scene transitions, we used the Welch t-test that takes into account the issues mentioned above. The sample distributions for changes in Exploration Range can be seen in the Figure 23.

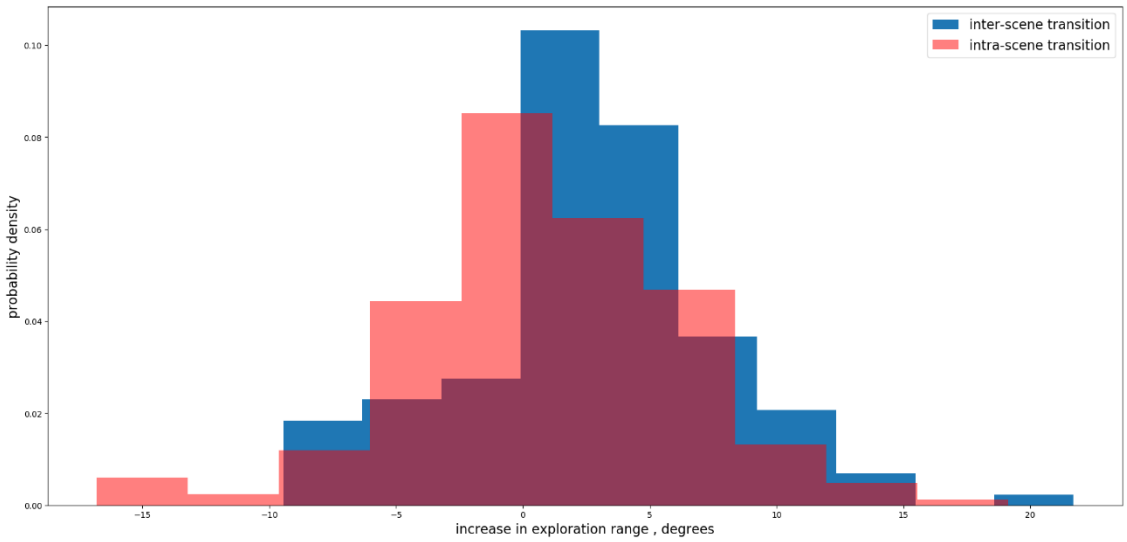

Figure 23. *Histograms of differences for inter- and intra-scene transitions*

We have checked the equality of the mean values for the distributions of differences for the inter- and intra-scene transitions being used in the statistical tests mentioned above. The calculated p-value is equal to 2.7%, showing that the expected values of the distributions are significantly different. Since the t-test value is positive and equal to 3.023, the increase of Exploration Range for **inter-scene** scene transition on average was more significant than the rise in another type of the cut.

4.4 Discussion

Our experiment has shown that the user's exploration behavior changes with the introduction of the scene cut. This is confirmed by the statistical inference based on Exploration Range data analysis. We reveal that the type of scene transition affects the user's head movements differently, and on average the **intra-scene** transition leads to faster head movements than the **inter-scene** transition. It was also discovered that there was a delay between the change in the exploration behavior and the scene transition, averaging about 1.5 seconds.

It should be pointed out that some issues would be interesting to explore in future researches. First, we analyzed the movement in the yaw direction only. While the most of the changes happen in the yaw dimension, the investigation of change in the pitch and roll dimension may lead to some new results. Second, there was some experimental imprecision related to the **intra-scene** transition experiment. For the selected clips, the average duration between scene transitions is equal to 13.3 seconds, while in the **inter-scene** transition experiment, this duration was equal to 20 seconds. One may argue that if the duration between the transitions is too short, the participant might not explore the omnidirectional scene throughly enough; therefore, the unusual behavior induced by the previous scene transition may interfere with the current scene transition. This may be a reason for the increase in the p-value for the **intra-scene** transition test compared to the p-value in the **inter-scene** transition test.

Our findings reveal that for three clips in the **intra-scene** transition experiment, the Exploration Range increase and reaction times differ from clip to clip. This means that the content of the clip itself affects the viewer's behavior. The influence of the clip's content on the viewer's reaction is a sophisticated problem, which requires much more in-depth analysis than the one given in our work.

Another area the research can be developed is to see how the transition method would affect the reaction time of the user. Reaction time can be used to prepare for the increased exploration after the scene cut (for example, it is possible to switch to shorter DASH segments for the duration of the increased exploration, so that the motion-to-high-quality delay would be decreased, improving the overall quality of experience).

5. CONCLUSIONS

In our work we designed and conducted an experiment that compares the viewing behavior before and after the scene transition in the omnidirectional video. We have prepared two sets of omnidirectional videos: the first set of clips was edited together to check the effect of scene transitions between different semantical locations (**inter-scene** transitions), the second set of clips contained three clips with multiple scene transitions in each of them, where the transition happened within the same location (**intra-scene** transition). The clips were then shown to twenty people. During the watch session, the subject's viewing patterns were recorded in terms of the yaw, pitch and roll angles of the HMD, sampled at the constant rate and stored for the further analysis.

We introduced metrics for analyzing the subject's behavior, mainly Exploration Range, Similarity Ring Metric, and the magnitude of angular speed. Exploration Range and magnitude of angular speed describe how swift the user's movement is at the given moment of time. On the other hand, Similarity Ring Metric describes how similar viewer's exploration patterns were for the given clip. Viewers with the exploration patterns highly dissimilar from the general viewing pattern for the clip were removed from the sample to reduce the amount of factors that could affect the exploration behavior for the scene transition. We developed a Python program for extracting the above-mentioned metrics from the collected traces. Based on those metrics, we developed hypothesis testing procedures that determined a set of patterns during the scene cut. The testing procedures were implemented using the SciPy library for the Python programming language.

The following results were obtained:

- The videos were watched with a high degree of similarity between different users, based on the Similarity Ring Metric.
- The introduction of the scene cut into the video leads to increased exploration behavior for both inter- and intra-scene transitions. On average, the movement after the scene transition is 77% higher than during normal exploration.
- The response to the **inter-scene** transition differs from the one to the intra-scene transition, that is the **inter-scene** transition leads to higher exploration on average.
- There exists a delay between the transition and the start of the exploration behavior that we characterize as reaction time. This delay depends on the content and is constrained by 1-2 seconds. The reaction time for the **intra-scene** transition is, on average, higher than the reaction time for an **inter-scene** transition.

Some of these results can be further explored in future works, for example, reaction time delay can work as a helpful mechanism for mitigating the bandwidth spikes during the increased exploration after the transition. Another direction for further research would be to check whether different transition effect can decrease the rise in exploration metrics. We hope that the analysis presented in this work would be helpful for the further development of streaming systems for omnidirectional videos.

6. REFERENCES

[1] Mangiante S, Klas G, Navon A, GuanHua Z, Ran J, Silva MD. VR is on the Edge: How to Deliver 360° Videos in Mobile Networks. Proceedings of the Workshop on Virtual Reality and Augmented Reality Network; Los Angeles, CA, USA. New York, NY, USA: Association for Computing Machinery; 2017.

[2] Schatz R, Zabrovskiy A, Timmerer C. Tile-based Streaming of 8K Omnidirectional Video: Subjective and Objective QoE. Evaluation. 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX). ; 2019.

[3] Curcio IDD, Naik D, Toukomaa H, Zare A. Subjective Quality of Spatially Asymmetric Omnidirectional Stereoscopic Video for Streaming Adaptation. Smart; Multimedia; Cham: Springer International Publishing; 2018.

[4] D. Monakhov, D. Naik, I. D. D. Curcio, H. Toukomaa. Analysis of User Exploration Patterns during Scene Cuts in Omnidirectional Videos. SMPTE 2018. 2018:1-20.

[5] Monakhov D, Curcio IDD, Mate S. On Data Wastage in Viewport-Dependent Streaming. ; 2019.

[6] Nielsen LT, M oller MB, Hartmeyer SD, Ljung TCM, Nilsson NC, Nordahl R, et al. Missing the Point: An Exploration of How to Guide Users' Attention during Cinematic Virtual Reality. Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology; Munich, Germany. New York, NY, USA: Association for Computing Machinery; 2016.

[7] Serrano A, Sitzmann V, Ruiz-Borau J, Wetzstein G, Gutierrez D, Masia B. Movie editing and cognitive event segmentation in virtual reality video. ACM Transactions on Graphics (TOG). 2017 Jul 20,;36(4):1-12.

[8] Magliano JP, Zacks JM. The Impact of Continuity Editing in Narrative Film on Event Segmentation. Cognitive Science. 2011;35(8):1489-1517.

[9] Zacks JM, Speer NK, Swallow KM, Maley CJ. The Brain's Cutting-Room Floor: Segmentation of Narrative Cinema. Frontiers in human neuroscience. 2010;4:168.

[10] Delong J, Brunick K, Cutting J. Film through the Human Visual System: Finding Patterns and Limits. In: ; 2012.

[11] Sitzmann V, Serrano A, Pavel A, Agrawala M, Gutierrez D, Masia B, et al. Saliency in VR: How Do People Explore Virtual Environments? IEEE Transactions on Visualization and Computer Graphics. 2018 Apr;24(4):1633-1642.

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

[13] Nakano M, Li S, Chiba N. Calibration of fish-eye camera for acquisition of spherical image. Systems & Computers in Japan. 2007;38(6):10-20.

[14] M. Jamali, F. Golaghazadeh, S. Coulombe, A. Vakili, C. Vazquez. Comparison of 3D 360-Degree Video Compression Performance Using Different Projections. 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE). 2019:1-6.

[15] Ghaznavi-Youvalari R, Zare A, Aminlou A, Hannuksela MM, Gabbouj M. Shared Coded Picture Technique for Tile-based Viewport-adaptive Streaming of Omnidirectional Video. IEEE Transactions on Circuits and Systems for Video Technology. 2018.

[16] Zare A, Aminlou A, Hannuksela MM, Gabbouj M. HEVC-Compliant Tile-Based Streaming of Panoramic Video for Virtual Reality Applications. Proceedings of the 24th ACM International Conference on Multimedia; Amsterdam, The Netherlands. New York, NY, USA: Association for Computing Machinery; 2016.

[17] Terzić K, Hansard M. Methods for reducing visual discomfort in stereoscopic 3D: A review. Signal Processing: Image Communication. 2016;47:402-416.

[18] Liu Q, Su X, Zhang L, Huang H. Panoramic video stitching of dual cameras based on spatio-temporal seam optimization. Multimedia Tools Appl. 2018.

[19] S. Li. Binocular Spherical Stereo. IEEE Transactions on Intelligent Transportation Systems. 2008;9(4):589-600.

[20] F. Bossen, B. Bross, K. Suhring, D. Flynn. HEVC Complexity and Implementation Analysis. IEEE Transactions on Circuits and Systems for Video Technology. 2012;22(12):1685-1696.

[21] Omnidirectional Media Format [Internet].; 2017 [cited 01.02.2020]. Available from: https://mpeg.chiariglione.org/standards/mpeg-i/omnidirectional-media-format.

[22] T. C. Thang, Q. Ho, J. W. Kang, A. T. Pham. Adaptive streaming of audiovisual content using MPEG DASH. IEEE Transactions on Consumer Electronics. 2012;58(1):78-85.

[23] de la Fuente, Y. S., Bhullar GS, Skupin R, Hellge C, Schierl T. Delay Impact on MPEG OMAF's Tile-Based Viewport-Dependent 360° Video Streaming. IEEE Journal on Emerging and Selected Topics in Circuits and Systems. 2019;9(1):18-28.

[24] Curcio I, Toukomaa H, Naik D. Bandwidth Reduction of Omnidirectional Viewport-Dependent Video Streaming via Subjective Quality Assessment, Proceedings of the ACM Int. Workshop on Multimedia Alternate Realities. Mountain View, CA; 27 October; ; 2017.

[25] Qian F, Ji L, Han B, Gopalakrishnan V. Optimizing 360 Video Delivery over Cellular Networks. Proceedings of the 5th Workshop on All Things Cellular: Operations, Applications and Challenges; New York City, New York. New York, NY, USA: Association for Computing Machinery; 2016.

[26] Walpole RE, Walpole RE, Myers RH, Myers SL, Keying Y. Probability and statistics for engineers and scientists. 7th ed ed. Upper Saddle River, NJ: Prentice Hall; 2002.

[27] Bruce P, Bruce A. Practical Statistics for Data Scientists: 50 Essential Concepts. O'Reilly Media; 2017.

[28] Newcombe RG. An Introduction to Medical Statistics, 2nd edn (1995) Author: J. Martin Bland Publisher: Oxford University Press, Oxford ISBN: 0-19-262428-8. European journal of orthodontics. 1996;18(3):308.

[29] Lumley T, Diehr P, Emerson S, Chen L. The importance of the normality assumption in large public health data sets. Annu Rev Public Health. 2002;23:151-169.

[30] Ratcliffe JF. The Effect on the t Distribution of Non-Normality in the Sampled Population. Journal of the Royal Statistical Society.Series C (Applied Statistics). 1968;17(1):42-48.

[31] e-Handbook of Statistical Methods [Internet].; 2020 [updated 01.02.;]. Available from: https://www.itl.nist.gov/div898/handbook/eda/section3/qqplot.htm.

[32] Lilliefors HW. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. Journal of the American Statistical Association. 1967;62(318):399-402.

[33] Wilcox RR. Introduction to Robust Estimation and Hypothesis Testing. Amsterdam: Academic Press; 2012.

[34] Wilcoxon F. Individual Comparisons by Ranking Methods. Biometrics Bulletin. 1945;1(6):80-83.

[35] Lu Z, Yuan K. Encyclopedia of Research Design; pages 1621-1623. In: Thousand Oaks; Thousand Oaks, California: SAGE Publications, Inc; 2020.

[36] Fagerland MW. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? . 2012;12(1).

[37] Han J, Kamber M, Pei J. 12 - Outlier Detection. Data Mining (Third Edition). 2012:543-584.

[38] Swenberg T, Eriksson PE. Effects of Continuity or Discontinuity in Actual Film Editing. Empirical Studies of the Arts. 2018;36(2):222-246.

[39] Ascher S, Pincus E. The filmmaker's handbook: a comprehensive guide for the digital age. , 2008 ed. New York: Plume; 2007.

[40] Liu C, Huang T, Chang M, Lee K, Liang C, Chuang Y. 3D Cinematography Principles and Their Applications to Stereoscopic Media Processing. Proceedings of the 19th ACM International Conference on Multimedia; Scottsdale, Arizona, USA. New York, NY, USA: Association for Computing Machinery; 2011.

[41] Kurby CA, Zacks JM. Segmentation in the perception and memory of events. Trends in Cognitive Sciences. 2008;12(2):72-79.

[42] A. MacQuarrie, A. Steed. The Effect of Transition Type in Multi-View 360° Media. IEEE Transactions on Visualization and Computer Graphics. 2018;24(4):1564-1573.

[43] ITU-T P.910 Subjective video quality assessment methods for multimedia applications [Internet].; 2008 [cited 01.02.2020]. Available from: https://www.itu.int/rec/T-REC-P.910.

[44] ITU-T P.916 Information and guidelines for assessing and minimizing visual discomfort and visual fatigue from 3D video [Internet].; 2016 [cited 01.02.2020]. Available from: https://www.itu.int/rec/T-REC-P.916.

[45] I.D.D. Curcio. Similarity Ring Metric (SRM) for Subjective Evaluation of 360-Degree Video ISO/IEC JTC1/SC29/WG11 MPEG #119 meeting, document m41303, Turin, Italy. 2017.

[46] Savitzky A, Golay MJE. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. Analytical Chemistry. 1964;36(8):1627-1639.

[47] Beyer H. Tukey, John W.: Exploratory Data Analysis. Addison-Wesley Publishing Company Reading, Mass. — 1977, XVI, 688 S. Biomedical Journal. 1981;23(4):413-414.