A Taxonomy and Dataset for 360° Videos

Afshin Taghavi Nasrabadi, Aliehsan Samiei, Anahita Mahzari, Ryan P. McMahan, Ravi Prakash
The University of Texas at Dallas, U.S.A.
{afshin,aliehsan.samiei,anahita.mahzari,rymcmaha, ravip}@utdallas.edu

Mylène C.Q. Farias, Marcelo M. Carvalho
Department of Electrical Engineering
University of Brasilia (UnB)
{mylene,mmcarvalho}@ene.unb.br

ABSTRACT

In this paper, we propose a taxonomy for 360° videos that categorizes videos based on moving objects and camera motion. We gathered and produced 28 videos based on the taxonomy, and recorded viewport traces from 60 participants watching the videos. In addition to the viewport traces, we provide the viewers' feedback on their experience watching the videos, and we also analyze viewport patterns on each category.

KEYWORDS

360° video, Dataset, Viewport, Virtual Reality

1 INTRODUCTION

Omnidirectional or 360° video is one of the many Virtual Reality (VR) technologies with a growing popularity. 360° video applications range from entertainment to education. These videos are usually watched through Head Mounted Displays (HMD) that enable viewers to explore a scene and look in any direction from a specific point in the scene. However, this new medium poses new challenges for content producers and service providers. For example, 360° videos should have a high spatial resolution (4K or above) to provide an acceptable level of Quality of Experience (QoE) for viewers. Therefore, processing and streaming this type of content is very demanding.

Several solutions have been proposed to stream and render 360° videos based on real-time users' viewport [3, 12]. They take advantage of the fact that users, at any point in time, view a limited portion of the video. To provide a high quality video inside a users' viewport, these methods need to know the users' viewport beforehand. This is typically done using viewport prediction methods. Since the accuracy and duration of existing viewport prediction methods are limited, viewport cannot be accurately predicted for time intervals longer than one second [13]. This limits the usefulness of viewport prediction under fluctuating network conditions as the video client has to buffer a long duration of the video to cope with network variations. Any mismatch between the predicted and actual user viewport can be detrimental to QoE. Another interesting challenge is storytelling, which, so far, has not been well-defined for this type of media. Unlike traditional videos, viewers are not limited to watch only a specific part of the scene determined by producers. Also, the effect of camera motion and scene-cuts are very different, which requires that producers know how to guide user attention. Currently, there are ongoing efforts to study the effect of different editing techniques on the QoE of 360° videos [16].

Solving the above-mentioned challenges requires study and analysis of user behaviors while watching 360° videos. Publicly available viewport datasets facilitate these studies for several reasons. First, study of viewport traces enables content producers to understand which aspects of a 360° video are important for users and how their attention can be guided. Second, datasets can help to develop and test viewport prediction and saliency detection methods. Moreover, these traces can be used by other researchers for the purpose of running experiments related to 360° video streaming, as well as salience and visual attention modeling. Since users' viewport patterns are highly influenced by the video content, it is important to have various types of 360° videos in a dataset. Moreover, a taxonomy of videos could be developed and videos could be appropriately classified on the basis of a set of attributes. If there is a high correlation between viewing patterns for videos in the same category, and significant differences between videos in different categories, then taxonomy information could be leveraged for viewport prediction. Today, there are several published datasets for users' viewport traces [5][8][2][11][18]. However, they do not provide a comprehensive taxonomy of videos.

We propose a taxonomy for 360° videos that classifies videos based on camera motion and number of the moving targets in a video scene. We gathered and produced 28 scene-cut free videos based on the proposed taxonomy. We designed a subjective experiment in which 60 viewers watched a subset of the videos. In addition to providing the viewport traces for each viewing session, our dataset[1] includes the viewers' feedback about their experience after watching each video. Viewers described what they focused on and rated their perception of presence level of discomfort. These responses can be very helpful to study viewport traces. We also present preliminary analysis of user data that could be helpful in designing viewport prediction algorithms.

2 RELATED WORK

Several datasets provide head movement traces of users watching 360° videos. Corbillon *et al.* [2] gathered a dataset of viewport traces for five videos with 59 participants. Lo *et al.* provided a dataset captured with 50 subjects watching 10 videos [11]. Saliency

[1] https://gitlab.com/UTDdslab/360dataset

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

and motion map of videos are also available in the dataset. The authors designed a viewport prediction method based on their dataset in [6]. Wu *et al.* [18] provided a dataset with 18 videos watched by 48 participants. They classified the videos based on their genre, such as sports, documentary, etc. This classification is very general and does not characterize the intrinsic properties of a scene. Their test procedure included two types of experiments. In the first experiment, subjects just watched the videos. In the second experiment, after each video, subjects were asked specific questions about the video content. This type of experiment forces viewers to pay attention to the content of videos. As a consequence, viewport samples are more scattered in the first experiment than in the second one. Duanmu *et al.* [5] provided a dataset of viewport traces in which videos were watched on a computer monitor, with the users navigating the video using mouse. The dataset includes 12 videos and 50 users. The authors compared the similarities and differences of viewing patterns between HMD and computer-based viewing sessions.

Fremerey *et al.* [8] provided a dataset of head movements from 48 subjects watching twenty different videos. Participants also filled a Simulator Sickness (SS) Questionnaire after each set of five sequences. Although the overall discomfort was not very high, female participants experienced a higher discomfort level, which increased over the course of different videos. David *et al.* provided a dataset of head and eye movements [4] for nineteen videos watched by 57 subjects. The dataset includes head+eye and head-only saliency maps and scan-paths. Interestingly, their results show that there are some differences between head-only and head+eye saliency maps, which are not highly correlated. According to the authors, this is caused by a loss of information in head-only maps. Generally, users' behavior depends on the content of the video. However, users' viewport is biased towards the center of the video. According to Fremerey *et al.* [8], most of the time users watch the areas closer to the center of the videos, with 90% of the time within ±30° deviation from equator, and 50% of the time within the same interval from the center in horizontal direction. But, in horizontal direction, the viewports are more distributed, and nearly 50% of users watched 330° of horizontal view in all video sequences.

In [1], thirty 360° videos were shown to 32 participants. Videos were classified into 5 categories based on motion, but no distinction was made between the motion of object(s) in the scene versus camera motion.Their analysis shows that viewport patterns change for different categories. Users viewport distribution along yaw axis tend to be more uniformly distributed for videos without moving objects. In another study [14], a dataset is created based on 6 videos with duration of 10 seconds, which is much shorter than previous studies. 17 participants have watched the videos, and each video was randomly repeated. They found that most of the fixation points are around moving objects. Moreover, videos with high motion complexity have fewer fixation points. Xu et al. [19] have mined their own dataset for the purpose of viewport prediction. 58 subjects watched 76 panoramic videos. Analysis of their dataset shows that there is center bias, and there is similarity in the magnitude and direction of viewport changes when they are co-located.

3 TAXONOMY

Our goal is to design a taxonomy of 360° videos that puts videos with similar user viewing patterns in the same category. Users' head movement can be triggered by different features of the content. Several studies on viewport dynamics suggest that user attention is guided by moving targets in the scene [14]. Therefore, existence of moving objects plays an important role in a taxonomy. Additionally, we believe camera motion can affect viewer attention. In regular videos, camera motion dictates what users see in a scene. Although in 360° videos users are free to look in any direction, camera motion can alter user behavior and transform the motion of moving objects. So, we also would like to study the effect of camera motion on user viewport.

Therefore, we propose a two dimensional taxonomy for 360° videos based on the type of camera motion and number of moving objects. We classify videos into five different categories based on camera motion: 1- Fixed, 2- Horizontal, 3- Vertical, 4- Rotational, 5- Mixture of previous camera motions. Regarding moving objects in a scene, we study the effect of the number of moving targets in a scene on viewport changes. So we have three categories: 1- No moving object, 2- Single moving object, 3- Multiple moving objects. By comparing videos from these categories, we can study the effect of moving objects on viewport pattern. This taxonomy results in a total of fifteen categories shown in Table 1 with each category corresponding to a <camera movement, number of moving targets> combination.

3.1 Videos used in the study

We consider two videos for each category, so we can examine any similarity in viewport patterns for two videos from the same category. Each video has duration of one minute. During the one minute cut, the videos have no scene-cuts. So there is no discontinuity during the video. In Table 1, each video has a numerical ID from 1 to 30, referred to as *videoID*. Each cell contains two videos from the same category. For each video, resolution and frame rate are specified. Most of the videos were chosen from YouTube, but, we also produced several videos for categories such as rotational camera movements (videos 4,19,20,22,23,24). For YouTube videos, the links to videos and start time are inside brackets. The full URL is the concatenation of www.youtu.be/ and the address in the table. If the source video is longer, we use only one minute of the video. For the recorded videos, we used Samsung Gear360 camera. Spatial and temporal complexity of videos is depicted in Figure 1 [10]. All videos can be found in our dataset. In our study, as we focus on the visual stimuli only, we removed audio from the videos.

We did not include any video for the category of vertical camera motion with one moving object. There is no publicly available video containing only vertical camera motion and one moving object of at least one minute duration without scene cuts. We tried to produce two videos for this category using a camera mounted on a drone. But the outcome was too shaky and could cause discomfort to viewers. So we decided not to include them in our study.

4 EXPERIMENTAL METHODOLOGY

To study users' behavior under the proposed taxonomy, we designed a subjective experiment where participants watch a set of

Table 1: Taxonomy and video links

		Number of Moving Targets		
		None	**Single**	**Multiple**
Camera Motion	**Fixed**	1) 3840x1920 25fps [ESRz3-btZIA (0:40)] 2) 3840x2160 29fps [30cSb3wTc7U (0:00)]	3) 3840x2048 29fps [ULixPLH-WA4 (0:07)] 4) 3840x1920 29fps	5) 3840x2048 30fps [7IWp875pCxQ (0:18)] 6) 3840x2048 29fps [ze_w7Lh97Co (0:05)]
	Horizontal	7) 3840x2160 25fps [9XR2CZi3V5k (0:01)] 8) 3840x2048 29fps [6TlW1ClEBLY (0:45)]	9) 2560x1440 29fps [tVsw0DvAWdE (0:15)] 10) 3840x1920 29fps [cNlQrTkXkOQ (0:15)]	11) 3840x2160 24fps [jMyDqZe0z7M (0:00)] 12) 3840x2160 30fps [2Lq86MKesG4 (0:12)]
	Vertical	13) 3840x1920 29fps [DgxmQvWEGBU (0:04)] 14) 3840x1920 29fps [elhdcvKhgbA (0:14)]	15) —————— 16) ——————	17) 3840x2048 30fps [jau-Ric7kls (1:11)] 18) 3840x2160 29fps [905_oiaJN_0 (0:15)]
	Rotational	19) 3840x1920 29fps 20) 3840x1920 29fps	21) 3840x2160 29fps [ZRFIdczxxkY (0:04)] 22) 3840x1920 29fps	23) 3840x1920 29fps 24) 3840x1920 29fps
	Mixed	25) 3840x2048 25fps [HiRS_6BCyG8 (0:00)] 26) 3840x2160 29fps [L_tqK4eqelA (5:30)]	27) 3840x1920 30fps [AX4hWfyHr5g (0:00)] 28) 3840x1920 29fps [VGY4ksezNkY (2:11)]	29) 3840x2160 25fps [p9h3ZqJa1iA (0:00)] 30) 3840x1906 29fps [H6SsB3JYqQg (1:00)]

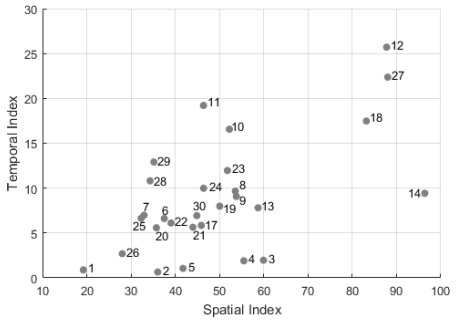

Figure 1: Spatial and Temporal complexity of videos in the dataset

360° videos, using a HMD, and answer a set of questions after each video. The experiment has three main parts: 1) *Training*, where participants answer an entry questionnaire and watch an introductory video[2]; 2) *Main session*, where participants watch one video from each of the taxonomy categories and answer a questionnaire after each video; 3) *Exit survey*, where participants answer a final questionnaire about their overall experience. Our experiment has been approved by the University Institutional Review Board.

The entry questionnaire is a background survey, which asks the participant's gender, age, and level of experience with using VR technology, including how many times the person has watched 360° videos. Then, the subject watches an introductory video to become familiar with the experiment and adjust the HMD. During the experiment, participants sit on a swivel chair and are free to look in any direction. We used Oculus Go HMD for this study, which is an all-in-one mobile, cable-free HMD. So, participants can rotate their head without cable interference. Since the videos do not have audio, users wear a headphone that plays white noise to eliminate auditory distractions.

In the main session, each participant watches the videos in a shuffled order of categories to compensate for temporal effects. The shuffled list has one (of the two) video in each category. When the playback is finished for each video, a gray screen is shown. Then, the subject takes off the HMD and asked the following questions:

Q1) Please describe what you saw while watching the video.
Q2) What did you focus on while watching the video?

[2]https://youtu.be/mlOiXMvMaZo starting at 0:30

Then, participants rate their presence level. We used the following four questions from the self-location portion of the Spatial Presence Experience Scale [9]:

Q3) I felt like I was actually there in the environment of the video.
Q4) It seemed as though I actually took part in the action of the video.
Q5) It was as though my true location had shifted into the environment of the video.
Q6) I felt as though I was physically present in the environment of the video.

Each question can be answered on a 5-point scale, from 1 (*do not agree at all*) to 5 (*fully agree*).

At the end of this questionnaire, we examine users' discomfort using a discomfort score [7]. The participant chooses their discomfort score in the range from 0 to 10, where 10 is the highest discomfort level and 0 is the lowest. The experiment is terminated if a subject chooses the maximum score at any moment. At the end of the experiment, an exit survey is conducted that asks for participants to choose their three favorite videos. All questionnaires are answered on a desktop computer.

4.1 Recording head movements

We developed a video player in Unity using Pixvana SPINPlay SDK[3] that plays back videos and records users' head orientation samples at the rate of HMD's screen refresh rate which is 60Hz. The player on the HMD is connected to a server on a PC. The server controls video playback on the HMD and collects recorded traces from HMD.

Before presenting the format of the recorded viewport dataset, we explain the coordinate system and the video projection format that were used. All videos are in equirectangular projection. Figure 2 depicts how an equirectangular video is shown to a viewer, along with the coordinate system. During playback, the video is mapped to a sphere. The coordinate system is defined such that the Z axis always points out to the center of equirectangular video. If we assume that a viewer looks in the direction of Z, then the Y axis points up and the X axis points right. In the beginning of video playback, the sphere is rotated along the Y axis to bring the center of the video to the front of the viewer. Note that the image on the surface of the sphere is mirrored, because the image is viewed form the inside of the sphere.

[3]https://pixvana.com/spin-sdk/

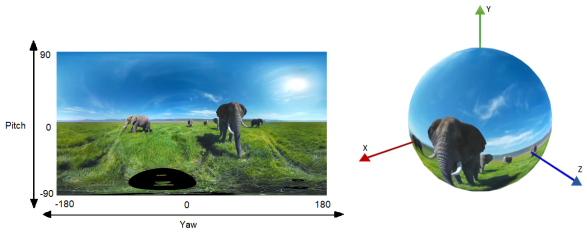

Figure 2: Left: Equirectangular frame. Right: The frame mapped to sphere and the coordinate system.

For each frame rendered on the HMD screen, we record the timestamp of the sample and the head rotation quaternion relative to the Z axis. We use a quaternion coordinate system because it is able to represent a rotation with a higher accuracy, if compared to Euler angles. Also, it does not have the gimbal lock problem [17]. We get rotation samples from `UnityEngine.XR.InputTracking` class in Unity. `GetLocalRotation(XRNode.CenterEye)` function provides the quaternion rotation of VR HMD. We record the samples and current timestamp inside `Update()` function loop that runs per rendered frame. In addition to the quaternion coordinate system, we also include the Cartesian coordinate system that points out to the center of user's viewport. Therefore, the format of the recorded samples is the following:

$$\{timestamp, Q_x, Q_y, Q_z, Q_w, V_x, V_y, V_z\}$$

where $\{Q_x, Q_y, Q_z, Q_w\}$ denote the components of the quaternion of viewport rotation and $\{V_x, V_y, V_z\}$ specifies the vector from the center of sphere to the center of viewport. We store all samples in a CSV file.

4.2 Dataset Structure

Our dataset has four main data folders: traces, videos, questionnaires, and viewportOverlays videos. The Traces folder contains the viewport traces of all participants. Each participant is assigned a 6-character number: the *SubjectID*. The Traces folder contains one sub-folder for each participant, which are named according to the *SubjectID*. It contains a CSV file, i.e., a trace file for each video watched by this participant. The CSV filename format is *SubjectID_VideoID.csv*. The Videos folder contains all videos used in the experiment, and the corresponding filename is $VideoID.mp4$. The *Questionnaires* folder contains the participants responses to all questions: *BackgroundQuestionnaire.csv* contains responses to the background survey, while *PerVideoQuestionnaire.csv* contains the answers to the questionnaire filled after each video. The header of questionnaire files contains the questions.

The *ViewportOverlays* folder contains all videos with an overlay layer of viewers viewport centers. We represent subject's viewport center using a 10° Gaussian kernel, and created a heatmap for each frame by applying the kernel for each viewport sample. This heatmap was merged to the original video frame, and finally the video was recreated from these frames. These videos provide useful visual representation of how videos were watched by viewers. We refer to these videos as *viewport-overlaid* videos. Figure 4 shows example frames of these videos.

5 RESULTS

A total of 60 persons participated in our study, from which 28.3% of participants were female. Each subject watched 14 videos plus the introductory video, with each video being watched by 30 viewers. Table 2 shows the age distribution of experiment participants, along with their experience with VR technologies.

Table 2: Subjects distribution and VR experience

Gender	Age	Mobile VR Exp.	Room Scale VR Exp.	360 Exp.
17 Female 43 Male	18-21 : 19 22-25 : 16 26-29 : 15 30-33 : 7 >40 : 3	Never: 17 1-5 times: 31 6-10 times: 4 11-20 times: 3 >20 times: 5	Never: 37 1-5 times: 15 6-10 times: 5 >10 times: 3	Never: 21 1-5 videos: 28 6-10 videos: 6 >10 videos: 5

5.1 Viewport Pattern

Analysis of the distributions of yaw and pitch angles over the whole duration of videos shows that pitch angle is biased along the equatorial section of the video. However, for videos captured from higher altitudes, the samples are more distributed, e.g., videos 17, 18, and 27. For the yaw angle, the distribution depends on the content and location of regions of interest. For example, videos 23 and 24, which contain rotational camera motion with multiple moving objects, have an almost uniform angle distribution. The dataset contains the histograms of yaw and pitch angles (see 'histogram' folder).

To analyze users' viewport pattern, we use the clustering algorithm proposed by Rossi *et al.* [15] which clusters viewports based on their overlap in spherical domain. We use an angle threshold of $\pi/5$[4]. The clustering algorithm divides a video into 3 second chunks, and viewports which are less than $\pi/5$ apart for 60% of the duration of a chunk are placed in the same cluster. A smaller number of clusters means that most viewers focused on specific parts of the video, while a higher number means that there were no common central focal points. Figure 3 shows the average number of viewport clusters for each category, with the error bars indicating the standard deviation associated to the averages. Notice that the number of clusters decrease as moving targets are added to the scene, with the exception being the videos with vertical camera movement. Figure 4 shows how a large number of viewers tracked the moving targets from frame a to frame b, which is one second later. Notice from this figure that a group of observers seem to be following moving targets (in this case, the man in the frames).

For the category corresponding to vertical camera movement and multiple moving targets, viewports are more dispersed compared to no moving target videos (see green bar in the third group in Figure 3). One possible reason for this behavior is that for most of the duration of these videos, the camera is located at high distance from the ground level and viewers have a landscape view. The viewport-overlaid version of these videos show that viewers were more interested in the landscape view, and did not focus on any specific area of the video.

Figure 5 shows the clustering results separated per video (and ordered by category) and Figure 6 depicts the corresponding number

[4]To run the clustering algorithm on our dataset see the 'scripts' folder in the dataset

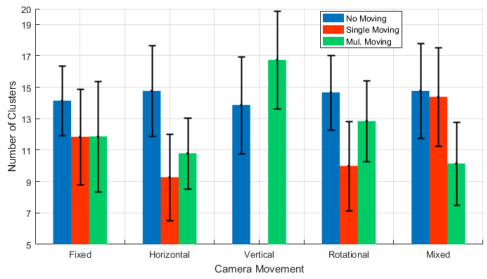

Figure 3: Barplot of the average number of clusters per category, along with the error bars.

Figure 4: Two frames from video 3 that show viewers tracking a moving target (walking man).

of viewers in the largest cluster for each of these videos. Notice that in some categorical pairs, such as (9,10), (13,14), and (27,28), the number of clusters for the videos is very different. For example, video 9 shows a view from a racing car that chases another car, while video 10 shows a woman walking in the woods. Both these videos were classified as having a horizontal movement and a single moving target. But, although for both videos the viewport-overlaid versions show a high concentration of viewports on the moving target, the speed of camera for video 9 is much higher and there are few objects (other than the two cars) in the video. Most likely, for single moving target videos, when the camera movement direction is aligned with the moving target, viewers are influenced to look at the target. Therefore, the number of clusters is smaller for this type of videos.

Videos 13 and 14 have a view from inside a glass elevator, with a vertical camera motion and no moving targets. One of the main differences between these two videos is that in video 14 (at time 0:22) the elevator stops and the door opens, attracting a lot of the viewers' attention and, therefore, reducing the number of clusters. Figure 7 shows how the number of clusters changes over time for videos 13 and 14, where a drop in the number of clusters can be seen in the curve for video 14 after 22s. Looking more closely, Figure 8 shows the viewport-overlaid views of the frames at instant 14s and 26s of video 14. Videos 17, 18, and 27 are also shot from a high altitude, and based on viewers feedback and viewport-overlaid videos, also had a landscape view that viewers found more interesting.

Comparing categories with one moving target, videos with camera motion can have fewer clusters if the camera follows the target. For example, in videos 8, 9, and 21, the camera moves according to the moving target, and the number of clusters is less compared to fixed camera category. In video 21, moving target is always at the center of video, and this video has more viewers in one cluster compared to similar videos without camera motion, e.g., videos 3 and 4. Video 22 has camera rotation but it does not follow the moving target, and viewers were more dispersed compared to video 21. However, for mixed camera motion, the pattern is not the same, as we mentioned that video scenery and camera elevation can affect the viewers.

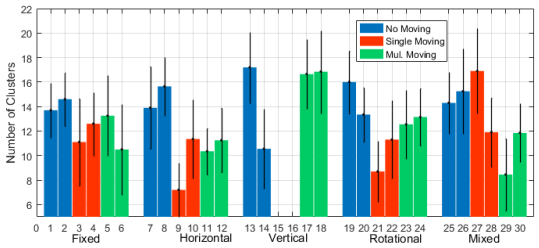

Figure 5: Barplot of the average number of clusters per video, along with the corresponding error bars.

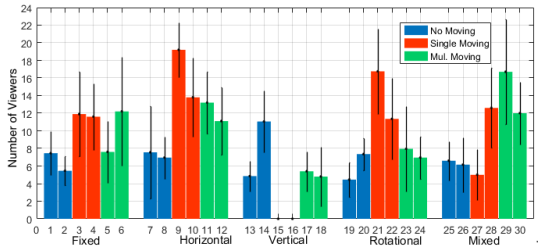

Figure 6: Barplot of the average number of viewers per video in the most populated cluster, along with the error bars.

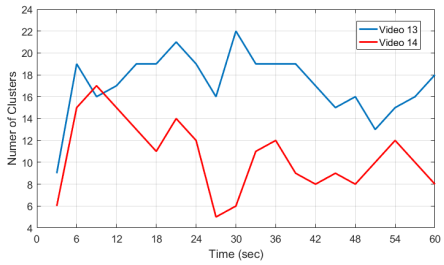

Figure 7: Number clusters during video playback for videos 13 and 14

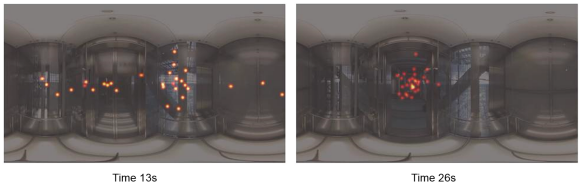

Figure 8: Two frames from viewport-overlaid version of video 14

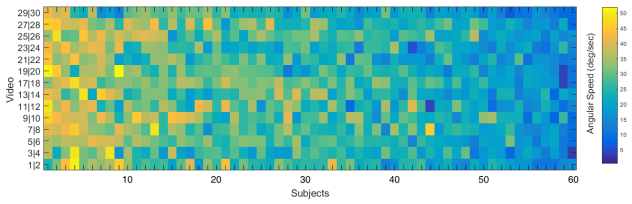

Figure 9: Subjects' head-movement speed heatmap

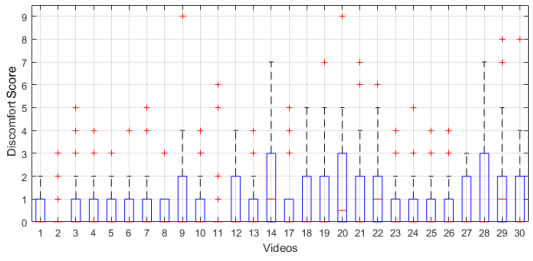

Figure 10: Boxplot of discomfort score per video

We also observed that viewers have different head movement and angular speed patterns. Some viewers tend to be still and rotate their head occasionally compared to others. Figure 9 shows the heatmap of average angular speed of viewers for each video. We measured the head movement speed at time intervals of 1 second, which is the great-circle distance between two head orientation samples divided by the total time. In this figure, each row corresponds to a category in the taxonomy and each column to a viewer. Viewers are sorted according to their average angular speed over all videos. It can be observed that viewers with a higher average speed watched all videos with a higher speed, which are represented by bright yellow colors in the map in Figure 9. On the other hand, barring a few exceptions, slower viewers had smaller average speeds, which are represented by darker blue colors in the map, for all videos. This suggests that users could be possibly classified on the basis of their head movement speed. So, head movement speed of a specific user in previous viewing sessions could be used to predict his/her future viewport.

5.2 Questionnaires

After watching each video, viewers chose their discomfort level. Figure 10 shows the box-plot graph of the users' responses. Although the range of scores is from 0 to 10, for most videos, the median score value was close to 0. More specifically, videos with fixed camera movements received low discomfort scores, while videos with camera motion received slightly higher discomfort scores. Also, for the presence level questions, Q3 to Q6, the average score is around 3, and the detailed responses are available in the dataset.

6 CONCLUSION

In this paper, we presented a taxonomy and dataset for 360° videos. We analyzed viewport traces according to the number of viewer clusters in each video. Generally videos with moving targets have fewer clusters, and preliminary investigation of viewport overlays on videos suggest that users tend to look at moving targets. However, there are some exceptions based on camera location and video scenery. For example, we observed that videos captured from higher altitudes have more dispersed viewport distribution irrespective of the number of moving objects. Some viewers tend to explore the scene more aggressively while others tend to be more passive, regardless of the nature of the video. In future work, we are going to study the relationship between viewports and moving objects in more details.

REFERENCES

[1] Mathias Almquist, Viktor Almquist, Vengatanathan Krishnamoorthi, Niklas Carlsson, and Derek Eager. 2018. The Prefetch Aggressiveness Tradeof in 360 Video Streaming. In *Proceedings of ACM Multimedia Systems Conference. Amsterdam, Netherlands.*

[2] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 2017. 360-degree video head movement dataset. In *Proceedings of the 8th ACM on Multimedia Systems Conference.* ACM, 199–204.

[3] Xavier Corbillon, Gwendal Simon, Alisa Devlic, and Jacob Chakareski. 2017. Viewport-adaptive navigable 360-degree video delivery. In *Communications (ICC), 2017 IEEE International Conference on.* IEEE, 1–7.

[4] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. 2018. A dataset of head and eye movements for 360Âř videos. In *Proceedings of the 9th ACM Multimedia Systems Conference.* ACM, 432–437.

[5] Fanyi Duanmu, Yixiang Mao, Shuai Liu, Sumanth Srinivasan, and Yao Wang. 2018. A Subjective Study of Viewer Navigation Behaviors When Watching 360-Degree Videos on Computers. In *2018 IEEE International Conference on Multimedia and Expo (ICME).* IEEE, 1–6.

[6] Ching-Ling Fan, Jean Lee, Wen-Chih Lo, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. Fixation prediction for 360 video streaming in head-mounted virtual reality. In *Proceedings of the 27th Workshop on Network and Operating Systems Support for Digital Audio and Video.* ACM, 67–72.

[7] Ajoy S Fernandes and Steven K Feiner. 2016. Combating VR sickness through subtle dynamic field-of-view modification. In *3D User Interfaces (3DUI), 2016 IEEE Symposium on.* IEEE, 201–210.

[8] Stephan Fremerey, Ashutosh Singla, Kay Meseberg, and Alexander Raake. 2018. AVtrack360: an open dataset and software recording people's head rotations watching 360Âř videos on an HMD. In *Proceedings of the 9th ACM Multimedia Systems Conference.* ACM, 403–408.

[9] Tilo Hartmann, Werner Wirth, Holger Schramm, Christoph Klimmt, Peter Vorderer, André Gysbers, Saskia Böcking, Niklas Ravaja, Jari Laarni, Timo Saari, and others. 2015. The spatial presence experience scale (SPES). *Journal of Media Psychology* (2015).

[10] ITU-T Recommendation P.910. 2008. Subjective video quality assessment methods for multimedia applications. (April 2008).

[11] Wen-Chih Lo, Ching-Ling Fan, Jean Lee, Chun-Ying Huang, Kuan-Ta Chen, and Cheng-Hsin Hsu. 2017. 360 video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference.* ACM, 211–216.

[12] 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.* ACM, 1689–1697.

[13] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. 2018. Your Attention is Unique: Detecting 360-Degree Video Saliency in Head-Mounted Display for Head Movement Prediction. In *2018 ACM Multimedia Conference on Multimedia Conference.* ACM, 1190–1198.

[14] Cagri Ozcinar and Aljosa Smolic. 2018. Visual attention in omnidirectional video for virtual reality applications. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX).* IEEE, 1–6.

[15] Silvia Rossi, Francesca De Simone, Pascal Frossard, and Laura Toni. 2018. Spherical clustering of users navigating 360° content. *arXiv preprint arXiv:1811.05185* (2018).

[16] 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 (TOG)* 36, 4 (2017), 47.

[17] Wikipedia. 2018. Gimbal lock. (2018). https://en.wikipedia.org/w/index.php?title=Gimbal_lock&oldid=849456729 [Online; accessed 9-February-2019].

[18] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang. 2017. A Dataset for Exploring User Behaviors in VR Spherical Video Streaming. In *Proceedings of the 8th ACM on Multimedia Systems Conference.* ACM, 193–198.

[19] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. 2018. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence* (2018).