Analysis of the Relationship between Content and Interaction in the Usability Design of 360° Videos

Nicholas Caporusso[1], Meng Ding[1], Matthew Clarke[1], Gordon Carlson[1], Vitoantonio Bevilacqua[2], Gianpaolo Francesco Trotta[2]

[1] Fort Hays State University, 600 Park Street, Hays, United States
[2] Polytechnic University of Bari, Via Edoardo Orabona, 4, Bari, Italy
{n_caporusso, mkclarke2, gscarlson}@fhsu.edu, m_ding6@mail.fhsu.edu, {vitoantonio.bevilacqua, gianpaolofrancesco.trotta}@poliba.it

Abstract. In the recent years, 360° images and video became a popular format for designing, producing, and consuming information. Immersive videos enable users to control the viewing angle at playback, and experience content in a unique fashion. Therefore, understanding how viewers interact with 360° video is crucial for improving their design, production, distribution, and consumption. In this paper, we introduce a model for categorizing 360° video based on the type of points of interest in the scene, and we present a study in which we analyzed heatmaps and changes in the viewing angle to identify the key features of regions of interest, the common interaction patterns within video categories, and the relationship between content design and engagement.

Keywords: Information visualization · Heatmap · Viewing angle · Immersive video

1 Introduction

Thanks to recent technology advances in video cameras, full 360° images and videos rapidly gained popularity as the best tools for experiencing locations (e.g., museums), environments (e.g., the rainforest), and activities (e.g., skiing) in an immersive fashion. Nowadays, affordable equipment supports shooting, producing, and playing high-quality 360° content with minimal technical effort. Moreover, most digital media platforms support uploading, playing, streaming, and even interacting with 360° videos using standard protocols. In the recent years, the production of 360° content and its consumption through popular platforms has significantly increased, and the integration of immersive video in online communities, mobile applications, and head-mounted devices is contributing to rendering 360 video a standard for storytelling. According to a recent report [1], 360° videos already are in a dominant position (99.37%) compared to VR content. In addition, according to recent reports, as the market of 360° cameras will grow over 1500% over the next six years, this opens new opportunities and scenarios for content production and utilization [2].

Several studies focused on technological aspects of 360° videos, such as, development of hardware rigs and mounts for optimal multi-camera alignment, research on

software for stitching images from multiple sources, and implementation of algorithms for optimizing video streams. Conversely, less attention has been dedicated to human factors and, specifically, to the design and fruition of 360° videos from a user interaction standpoint, though enhancing the quality of the experience is a key competitive factor [3]. As a result, despite the amount of material being produced both by amateurs and by professionals, there are no guidelines about how to create storyboards to fully unlock the potential of 360° content. Moreover, there are no best practices about how to organize Points of Interests (POIs) and visual cues in the scene so that users can change their viewing angle (VA) and navigate the story according to the intent of the director. Consequently, besides some experimental attempts, most of the experiential aspects of 360° content are still unclear and usability of 360° videos is yet to be investigated.

In this paper, we focus on the human aspects of 360° video, and we analyze the relationship between the content of images and user interaction (i.e., viewing angle). Specifically, the objective of our study is two-fold: by considering type, number, positioning, and movement patterns of POIs in the scene, we associate the video with a category that describes the syntactic complexity of the scene, that is, the amount and type of POIs in the video (i.e., no POIs, single fixed POI, single moving POI, multiple fixed POIs, or multiple moving POIs). This, in turn, together with the semantics of the video, might help predict user interaction dynamics and, thus, design both the story and the scene to include visual and auditory cues that guide users' viewing angle to help them experience the content.

Furthermore, we detail the results of a study in which we investigated the correlation between video design (i.e., syntactic complexity and content type) and interaction dynamics (i.e., how users navigate the scene). In our experiment, we tracked changes in viewing angles to generate heatmaps that reflect users' orientation, and we analyzed heatmap patterns within each category to validate the initial classification in terms of syntactic complexity. Also, we discuss qualitative information collected from participants using questionnaires aimed at understanding the relationship between syntactic complexity and overall performance of the video experience, evaluated on a two-dimensional emotion space. The consistency in our findings might suggest ways to approach the design of 360° videos based on a specific desired narrative outcome.

2 Related work

Indeed, the main feature introduced by 360° video is providing viewers with opportunities to control and change the viewing angle at playback. Therefore, as discussed by several studies, at any given time viewers focus on a specific portion of the video: a common property for several types of interactive videos is that there is a region of interest (ROI) currently viewed by the user [4] whereas the others are available, though not visible. This, in turn, has several implications in terms of pipeline of: (1) narrative design, (2) video recording, (3) data storage and transfer, (4) content visualization, (5) user experience, and (6) information perception and processing. Consequently, 360° content require a completely different approach to their design, devel-

opment, and distribution, with respect to traditional video. For instance, the availability of content over a full 360° angle, which offers a completely immersive viewing experience, requires data transmission of parts of the video that will not be displayed to the user, resulting in bandwidth consumption, increased loading time [5, 6], and potential quality reduction in the ROI. Also, the design and production of content should accurately consider the positioning of POI in the video, to avoid placing items that are crucial to the narrative in areas which potentially will be outside of the current ROI, at playback. To this end, the authors of [4] found that movement, sound and lighting cues from the fixation regions are the basic and effective methods for directing user's attention. Moreover, as immersive videos place the viewer at the center of the scene, the spatial organization of content becomes crucial to ensure a comfortable experience: [7, 8] reported that a distance of three meters between the user viewing point and items offered a good balance of being close enough to see clearly and creating a sense of immersion, whereas objects that are at a shorter distance may be perceived as unnaturally close and cause discomfort as they invade viewer's virtual personal space. Furthermore, being introduced in a realistic immersive environment may bias the audience and create the feeling that the scene is happening at that moment. As a result, in watching videos featuring people, viewers' engagement increases if they are acknowledged by the characters in the scene or if there is eye contact with them [7, 8].

Analyzing viewing patterns is crucial to understanding interaction with 360° video and, consequently, to improving the quality of content design, production and consumption. Recent studies focused on common ROIs, and they demonstrated that users' viewing directions are closely correlated for most of the videos [8]. In [9], the authors analyzed users' behavior when watching 360° videos in 5 categories based on content, that is, exploration, static focus, moving focus, rides, and miscellaneous. They utilized orientation and velocity parameters to compare angle distribution, Point of View (POV), and exploration phase, and they found common patterns in user interaction, that is, (1) angles distribution was highly dependent on video content, (2) viewpoints had minimal changes if there were clear characters in the scene, (3) at the beginning of a new video, viewers explored the scene to understand where the focus should be, and (4) the viewer tended to make large rotations towards the front than towards the back of a video.

Furthermore, other studies utilized heatmaps, which are a data visualization method that provides an intuitive way to identify regions of high and low concentration of a parameter, such as, ROI: areas of an image are color-coded based on the weighted number of ensemble members in that specific region [10], to show the point density interpolation within the area. Heatmaps, also known as intensity maps [11] are extensively utilized in several applications, such as, geographic data visualization for highlighting density of houses, crime reports, or roads or utility lines influencing a town or wildlife habitat [12, 13]. Heatmaps have been utilized in Human-Computer Interaction to study attention [14] and to produce maps of observation patterns, in combination with eye-tracking and gaze acquisition devices. Among their benefits, heatmaps render extremely easy to interpret the distribution over an area, because they use color coding to produce maps that quickly elicit relations and stimulate visually compari-

sons, facilitating differentiation between ROIs that received more attention and areas with less or no fixations. However, the main drawback of heatmaps is that they are not suitable for dynamic stimuli, such as, videos or systems whose interface changes as users interact with them [14]. Nevertheless, current video distribution platforms integrated heatmaps as a convenient visualization technique for analyzing the consumption of videos: as they aggregate information but conserve the essence of the forecast [10], heatmaps are offered to video publishers to visually support the human decision maker in understanding patterns of visualization and ways to improve the design of content for better audience engagement.

In this paper, we describe a study in which we analyzed the ROIs of 360° videos using static visualization methods (i.e., heatmaps) and the dynamic components of user interaction (i.e., time and rotation of the viewing angle) to identify the key features for content categorization and to achieve a classification of immersive content based on common user interaction patterns.

3 Study

We designed an observational study to evaluate the appropriateness and applicability of our model, and to compare it with categorizations presented in the most recent literature [9], which utilizes 5 classes, that is, exploration, static focus, moving focus, rides, and miscellaneous. Conversely, our model has four categories:

- contemplation, in which there are no explicit POIs, there is a fixed POV, and the video results in limited frame changes; as a result, the users would contemplate the scene in relation to implicit visual cues; gestalt
- observation: one fixed POI dominates the scene, which is watched from a fixed POV;
- tracking, in which the viewer has a fixed POV and the movement of one POI, which could consist in a group of elements, is expected to determine changes in the ROI;
- exploration, in which viewers have a fixed POV and are presented with a scene consisting on multiple moving POIs, or they are on a moving POV (e.g., in a car).

In our model, adjacent categories can overlap, so that an exploration video can contain contemplative elements, depending on the design of the content.

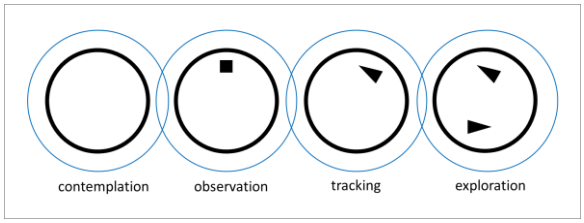

Fig. 1. Categorization model for 360° content depending on the type of elements in the scene. The square indicates a fixed primary POI, whereas triangles represent moving POIs.

A custom web-based 360° player was developed using Three.js [X], an open source library based on WebGL, which supports rendering video as texture material. Therefore, by mapping equirectangular frames on the surface of an empty sphere, the 3d engine enables the user to view the spherical projection from the inside, and to control the viewing angle using standard input peripherals or acceleration sensors (i.e., on a mobile device). We implemented a client-side script that continuously tracks the viewing angle and converts it into a pixel location in the equirectangular image. Width and height coordinates are stored in a flat file with their timestamp relative to the beginning of the video. For this experiment, we utilized a sampling rate of 10 Hz.

A set of 16 videos were selected among material available on the Internet, with the intent of maximizing differences between them in terms of type (e.g., contemplative, experiential), content (e.g., nature, music, story), number of points of interest (i.e., no POIs, one fixed POI, one moving POI, multiple fixed POIs, and multiple moving POIs), and amount and speed of action. Also, the selection considered the categories identified by other studies. Fragments from each video were extracted to standardize their length to one minute and to make sure that the scene is captured from one shooting angle, only. They were compressed to 1080p (1920 × 1080 pixels) variable-bitrate progressive MP4 formats at a data rate of 30 frames per second.

In this study, we did not acquire gaze using an eye tracking device (ETD), to replicate the same information that video players can record when users watch 360 videos, and to render our solution applicable and comparable to currently available distribution and analytics platforms. Also, we did not use a head-mounted display (HMD), because the majority of users interact with immersive videos using standard screens, due both to market penetration of this technology and to the limited availability of HMDs in the interaction.

A total of 29 subjects participated to the experiment in person. However, the study had a mortality of 6 who could not continue their involvement. Therefore, data from 23 participants were recorded and analyzed. The majority of subjects (65.22%) aged 25-35, 26.08% were 18-24, and 8.70% were 35-44. The group included 43.48% females and 56.52% males. Participants had a high education level (which we associated to a factor of technology use) 56.52% had master's degree, 39.13% had a bachelor's degree, whereas others accounted for 4.35%.

The study was realized in a dedicated room with no distractions and optimal light conditions, where participants were comfortably seated on a swivel chair, in front of a 20" display playing videos at full screen. Subjects were asked to watch the videos, which were played in a random order. After each task, they were presented with a short questionnaire asking them to evaluate the video in terms of (1) perceived duration (length in seconds), (2) perceived engagement (on a Likert scale), (3) clarity of

the content (on a Likert scale), (4) quantity of content (on a Likert scale). Also, they were asked to categorize the video according to the model described in Figure 1.

4 Results and discussion

A total of 296 videos were displayed to participants. Data from ROIs were acquired and answers to the questionnaires after individual videos were collected. Table 1 summarizes responses to the survey. Interestingly, they show strong correlation with the type of category and with the heatmaps (see Figure 2). Moreover, results have a strong internal consistency. Although the duration of each video was 60 seconds, viewers perceived them differently. In accordance with studies about perception of time in relation to the engagement of an experience, content perceived as less engaging was rated as lasting longer than one minute, whereas users associated a shorter duration with videos that they considered more interesting. Video 4 (contemplation), which featured a lecture, rated as the longest, compared to videos 13 (exploration, featuring a roller-coaster ride), and two of the tracking videos, 10 and 9 (consisting of a hip-hop dance and a film-like computer graphics-generated scene, respectively), which were perceived as the shortest. In addition to content and engagement, there could be other reasons for this difference. For instance, as the audio track was removed, viewers might have perceived videos in which sound was is a prominent component (e.g., concert and ballet, such as videos 6 and 7, respectively) as lasting longer, because of the different experience compared to an actual situation. Also, information elicited by cues in the videos and associated with memories, beliefs, and personal history, might influence the perceived duration: as an example, video 1 (contemplation), containing a still image performed better than video 4 (contemplation). In addition to quantifying duration in seconds, users were asked to evaluate whether the video was too long or too short, using a Likert scale. The results were consistent with the data for perceived duration. In addition, users' responses on engagement have a strong negative correlation with the duration, as the most engaging videos were perceived to last less than the least interesting ones.

Table 1. Viewers' responses for 360° videos

Video	Perceived duration	Engagement	Content
1	59.25 ±11.04	1.95 ±1.05	4.65 ±0.67
2	59.77 ±10.74	2.09 ±0.97	4.41 ±0.67
3	59.35 ±11.21	2.09 ±0.95	4.39 ±0.78
4	61.00 ±10.46	2.35 ±1.23	3.6 ±0.99
5	59.52 ±9.86	2.86 ±0.96	3.14 ±0.91
6	57.38 ±9.30	2.86 ±0.96	3.43 ±0.98
7	57.61 ±8.38	3.74 ±0.96	2.61 ±0.84
8	58.86 ±8.99	2.82 ±1.22	3.55 ±0.86
9	52.39 ±8.38	4.09 ±1.16	2.13 ±0.92
10	54.35 ±7.43	4.43 ±0.66	1.91 ±0.79
11	56.90 ±9.55	3.43 ±1.21	2.57 ±1.08
12	55.87 ±8.87	2.96 ±1.02	3.00 ±1.00
13	50.00 ±6.22	4.26 ±0.81	2.61 ±1.03
14	57.50 ±8.69	3.82 ±0.91	3.09 ±0.97
15	58.10 ±9.93	3.86 ±0.79	2.81 ±1.17

16	57.50 ±8.69	3.91 ±0.87	2.68 ±1.17

In addition, we asked participants to rate video clarity and to evaluate whether it was easy to interpret the scene or content was confusing. Videos 13 (exploration of a roller coaster ride), 9 (tracking of a movie scene), and 6 (observation of concert) were the ranked first, whereas video 12 (a ride on a Super Mario Bros cart) was regarded as the most confusing video. This is because the POV was moving on a car, but the viewing angle was not aligned with the direction of the cart. As a result, participants were changing their VA to match the direction faced by the car, otherwise they faced the sides or the back of the car. Video 12 is an example of overlapping between the tracking and exploration categories, because it involves a moving POV, though viewers have to change VA to track the direction of the car. However, most of the videos featured simple content or stories.

Furthermore, participants were asked to describe the amount of content, to evaluate whether there was too much (values closer to 1) or too little (values closer to 5) happening, or if the quantity of information in the video was appropriate (values close to 3). Although there are some outliers, responses from participants show correlation with the categories of the videos. Data from user interaction patterns confirm the information explicitly stated by subjects, though groups account for the sign of correlation.

Table 2. Spearman's R statistics analysis for 360° videos

Factors	Perceived duration	Engagement	Content
Perceived duration	——	-0.781***	0.797***
Stated duration	-0.804***	0.917***	-0.876***
Engagement	-0.781***	——	-0.91***
Clarity	-0.31	0.391	-0.415
Content	0.797***	-0.91***	——

Note: *** means $p<0.001$

Table 3 shows the internal consistency of our results, and it summarizes the relationship between duration, engagement, and content. From our findings, we identified negative correlation between perceived duration and stated duration (videos perceived as lasting less are perceived as shorter), and between perceived duration and engagement (less engaging videos are perceived as longer), whereas perceived duration and content show positive correlation (if content is obvious, then the video is perceived as boring). The negative sign of the correlation between the perceived duration and the stated duration is due to the sorting utilized in the scale.

Heatmaps obtained from multiple VA tracks of different users (see Figure 2) revealed insightful information about user experience with the immersive video and their interaction patterns with the content of 360° scenes. Heatmaps showed interesting similarities due to natural overlapping occurring between adjacent categories.

From our findings, we can conclude that both perceived and stated duration have a positive correlation with data points describing the VAs and, specifically, with the dispersion (d) in ROIs shown in the heatmaps (Figure 2). Moreover, depending on the category, the dispersion coefficient represents specific features of the video, which, in turn, can be utilized to modulate the experience. Specifically, in contemplation videos, when there are no specific events, dispersion reflects the presence of visual cues in the scene (e.g., roads, forests, and trails) that attract the user, because they might reveal some action. Conversely, in observation videos, dispersion is moderate and it is triggered by events, such as, an applause, or by elements that instantaneously become POIs. Changes in VAs are elicited by movement of items in the scene that do not belong to the main POI, and their effect is stronger when they are humans. In tracking videos, dispersion is elicited by the interaction of elements in the scene: patterns in VAs are less noisy when the story involves movement of a single element, or multiple elements move within one ROI. Finally, for exploration videos, if the POV moves, dispersion in VAs is negatively correlated with the speed of the POV.

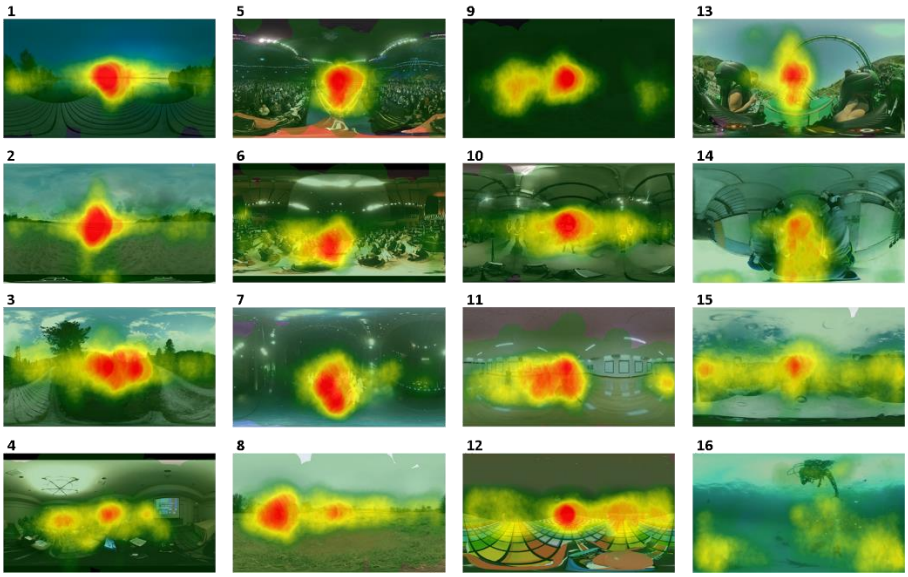

Fig. 2. Heatmap visualization of the videos utilized in the experiment: tracks of VAs from multiple users have been utilized to generate attention maps. The four categories are represented in the columns: contemplation (1-4), observation (5-8), tracking (9-12), and exploration (13-16). Features that include movement (quantity of changes in pixels) result in changes in ROIs and in different concentration (red indicates high persistence of the VA). Videos are organized in order of their dispersion coefficient (d), described in Figure 3.

Based on our findings, we argue that the categorization achieved by [9] can be further improved by incorporating our model, which has the benefit of being more abstract and, potentially, content agnostic. This, in turn, might enable automatic identification of ROIs based on correlation between a limited training set of user interaction patterns and the corresponding features extracted from frame-by-frame pixel changes.

Consequently, this would enable implementing compression codecs and streaming protocols that could deliver ROI-based variable-bitrate quality.

Moreover, as shown in Figure 3, users have different patterns in exploring the scene, and visual cues have a different role depending on the amount of information and on the speed of movement and elements change in the scene. For instance, in exploration videos, the relevance of visual cues and even moving elements decreases with the speed of the POV, whereas the importance of implicit and distal visual cues increases in contemplation videos. In observation and in tracking videos, changes in VAs indicate lower levels of engagement. Especially in the latter category, dispersion in viewers' ROIs might reveal issues in the spatial organization of content, or in the design of the scene.

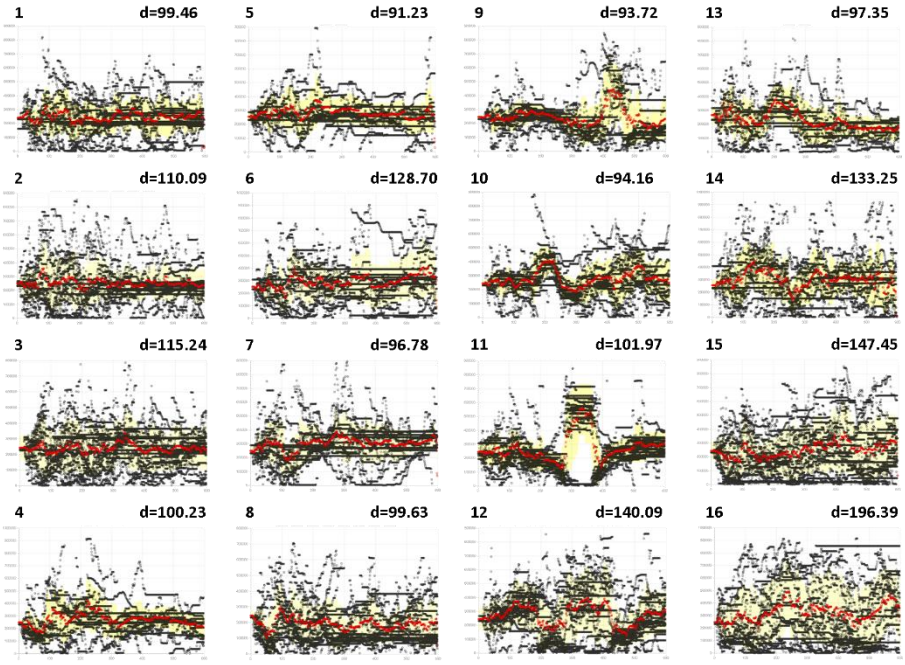

Fig. 3. Scatterplot visualization of ROIs over time, represented as the pixel in the middle of the region of interest at any given time. Videos are arranged as in Figure 2.

5 Conclusion

In the recent years, great emphasis has been given to technological features of 360° video (e.g., resolution, format, and devices), whereas less attention has been dedicated to content. Nevertheless, the increasing community of producers, directors, and video makers - both professional and amateur - are starting to investigate how to fully understand and unlock the potential of 360° video.

In this paper, we focused on the usability of 360 video by evaluating how users interact with immersive scenes. To this end, we proposed a model for categorizing immersive videos based on the number and type of points of interest, and we detail the results of a study focusing on how users change their viewing angles and interact with regions of interest, to analyze interaction with different categories of video and different types of content. From our findings, we can confirm the robustness of our categorization system, which could be utilized, together with rotation patterns in viewing angles and with dispersion in ROIs to predict viewers' engagement.

Our work aimed at establishing a framework that can be utilized in future studies to investigate aspects involved in the production, recording, and consumption of 360° video in the areas such as, journalism, narrative storytelling, entertainment, and interaction.

References

1. HUAWEI iLab, "VR data report," HUAWEI Report. (2016). https://mp.weixin.qq.com/s/tcsm9NIECa7d1L7gZekrrQ.
2. The MPEG Virtual Reality Ad-hoc Group. (2016). "Summary of survey on virtual reality," in ISO/IEC JTC 1/SC 29/WG 11 N16542.
3. Kaasinen, E., Roto, V., Hakulinen, J., Heimonen, T., Jokinen, J.P., Karvonen, H., Keskinen, T., Koskinen, H., Lu, Y. and Saariluoma, P. (2015). Defining user experience goals to guide the design of industrial systems. Behaviour & Information Technology, ahead- of-print: 1-16.
4. Sheikh, A., Brown, A., Watson, Z., & Evans, M. (2016). Directing attention in 360-degree video.
5. Bao, Y., Wu, H., Zhang, T., Ramli, A. A., & Liu, X. (2016,). Shooting a moving target: Motion-prediction-based transmission for 360-degree videos. In Big Data (Big Data), 2016 IEEE International Conference on (pp. 1161-1170). IEEE.
6. Bao, Y., Zhang, T., Pande, A., Wu, H., & Liu, X. (2017). Motion-prediction-based multicast for 360-degree video transmissions. In Sensing, Communication, and Networking (SECON), 2017 14th Annual IEEE International Conference on (pp. 1-9). IEEE.
7. Bailenson, J. N., Blascovich, J., Beall, A. C., & Loomis, J. M. (2003). Interpersonal distance in immersive virtual environments. Personality and Social Psychology Bulletin, 29(7), 819-833.
8. Wilcox, L. M., Allison, R. S., Elfassy, S., & Grelik, C. (2006). Personal space in virtual reality. ACM Transactions on Applied Perception (TAP), 3(4), 412-428.
9. Almquist, M., & Almquist, V. (2018). Analysis of 360° Video Viewing Behaviours.
10. Köpp, C., von Mettenheim, H. J., & Breitner, M. H. (2014). Decision analytics with heatmap visualization for multi-step ensemble data. Business & Information Systems Engi-neering, 6(3), 131-140.
11. Yeap, Emma, &I., Uy. (2014). "Marker Clustering and Heatmaps: New Features in the Google Maps Android API Utility Library." Google Geo Developers. Accessed April
12. ArcGIS, E. S. R. I. (2012). 10.1. Redlands, California: ESRI.
13. DeBoer, M. (2015). Understanding the heat map. Cartographic Perspectives, (80), 39-43.
14. Tula, A. D., Kurauchi, A., Coutinho, F., & Morimoto, C. (2016). Heatmap Explorer: an interactive gaze data visualization tool for the evaluation of computer interfaces. In Proceed-ings of the 15th Brazilian Symposium on Human Factors in Computer Systems (p. 24). ACM.