Detection of Strong and Weak Moments in Cinematic Virtual Reality Narration with the Use of 3D Eye Tracking

Paweł Kobyliński
Laboratory of Interactive Technologies
National Information Processing Institute
Warsaw, Poland
e-mail: pawel.kobylinski@opi.org.pl

Grzegorz Pochwatko
Virtual Reality and Psychophysiology Lab
Institute of Psychology, Polish Academy of Sciences
Warsaw, Poland
e-mail: grzegorz.pochwatko@psych.pan.pl

***Abstract*—Cinematic Virtual Reality (CVR) is a medium growing in popularity among both filmmakers and researchers. The medium brings challenges for movie and video makers, who need to narrate in a different way than in traditional movies and videos to keep viewers' attention in the right place of the 360-degree scene. In order to ensure an adequate pace of development, tools are needed to conduct systematic, reliable and objective research on narration in CVR. In the short paper, the authors for the first time fully report results of the initial empirical test of their recently developed Scaled Aggregated Visual Attention Convergence Index (*sVRCa*). The index utilizes 3D Eye Tracking (3D ET) data recorded during a CVR experience and allows measuring and describing the effectiveness of any system of attentional cues employed by a CVR creator. The results of the initial test are promising. The method seems to substantially augment the process of detection of strong and weak moments in CVR narration.**

Keywords–cinematic virtual reality; omnidirectional video; 3D eye tracking; visual attention; narration.

I. Introduction

Cinematic Virtual Reality (CVR) is a medium growing in popularity among both filmmakers and researchers. The recent rapid growth has been made possible by synergic progress in technology related to omnidirectional video cameras, VR headsets, computer hardware, software, and internet bandwidth. At this stage, some creators try to transfer old and tested narrative methods from traditional movies, while others have realized that they need to develop a new film language. Both educational and artistic CVR creators, who have sought advice regarding problems with guiding visual attention, have approached the authors of this short paper. Hence, these encounters constituted the real-life inspiration for this reported work in progress.

Since the viewpoint has been moved to the center of the movie set, participants, not directors, are in charge of the visual focus. This means the viewers are very probable to miss an important element of a story. Therefore, one has to rely more on light, spatial sound and arrangement of the scenery while building good narration. Setting actors around the camera and acting itself must be also different and thought over well.

In order to ensure an adequate pace of development, tools are needed to conduct systematic, reliable and objective research on narration in CVR. Visual attention is the crucial factor to measure if one wants to assure audience retention and make effectively entertaining, persuasive, or educational VR movies [1][2]. In the current short paper, the authors for the first time fully report results of the initial empirical test of their recently developed Scaled Aggregated Visual Attention Convergence Index (*sVRCa*). The quantitative index utilizes 3D Eye Tracking (3D ET) data recorded during a CVR experience. Theoretical basis of the *sVRCa*, among other variants of the Visual Attention Convergence Index (*VRC*), has been described in detail in [3] and [4].

The rest of the paper is structured as follows. Section II positions this work in relation to the other similar works in the literature. Section III presents the scaled aggregated visual attention convergence index. Section IV presents and discusses the empirical test. The work is concluded in Section V.

II. Relation to other Work

The method tested by the authors is by no means the first attempt to tackle the problem of measurement and improvement of CVR quality by addressing the issue of visual attention. Substantial work has been done to develop [5]-[7] and benchmark [8][9] computational models of visual attention prediction in order to advance optimization of such technical aspects as video compression [10], immersive media distribution formats [11], cashing for streaming [12], and artifact detection [6]. Attempts have also been made to analyze thoroughly the real viewing behavior in CVR [7][8][10][12]-[15].

On the other hand, some niche research is focused explicitly on the problem of storytelling in CVR. Ways of directing attention in CVR are proposed in [16]-[18]. The quality of narration is evaluated by viewers in [19]. Subjective questionnaires and recorded head orientation were used to assess the quality of video cuts and storytelling in [15]. Audience retention was analyzed in comparison with an "average YouTube video" in [20].

In contrast to the mentioned attempts, the authors report here a concise, timelined, near-continuous value, based solely on the measured positions of gaze fixations and computed without the need for prior computation of saliency maps or saliency optimization (neither theoretically- nor empirically-driven) [8][10][14]. The authors do not propose another measurement of video quality intended to improve its computational properties, neither by means of predicting nor mathematical optimization of visual attention. Moreover, the proposed method is designed to act differently than methods based on entropy measures [14]. Low values of the utilized *sVRCa* index do not necessarily relate to visual attention scattered randomly around the 360-degree scene (an unrealistic scenario in the case of narrated videos). Instead, the method may detect moments in which the values of index remain low, despite the order present in the visual attention pattern when different viewers look at distinct objects located at the opposite sites of the 360-degree video (a realistic scenario).

To the authors' best knowledge, the tested measure is the first reported one that is both objective in a quantitative, data-driven manner and, at the same time, indented to relate simply to a narration line chosen subjectively by a CVR video maker.

III. Scaled Aggregated Visual Attention Convergence Index

Values of the *sVRCa* index tell us if several people looked at the same or rather different virtual areas of the 360-degree scene during a chosen, short time interval. The index is based on Euclidean distances in 3D space and aggregates information about gaze fixations from a group of CVR experience participants.

The values are scaled to the range between 0 and 1, which is convenient for between-experiment comparisons. The authors have decided to fine-tune the simplified scaling proposed in [9] in order to make it more realistically constrained. The improved formula takes the assumption that the index takes approximately zero value when there are only two points of viewers' focus located at a maximum possible distance from each other, at the opposite sides of the virtual scene:

$$sVRCa \approx 1 - \frac{\sqrt{2}}{n} \frac{\sqrt{\sum_{i,j=1}^{n} \boldsymbol{D}_{ij}^{\,2}}}{2r} \in [0,1] \qquad (1)$$

D is the $n \times n$ distance matrix calculated from 3D positions of n detected gaze fixations (corrected for the headset positions in virtual space). In the case of CVR, the 360-degree video is displayed on the inner surface of a virtual sphere. r denotes the radius of the sphere (see [3] for details). In the future, it might be reasonable to find the exact scaling formula for sphere-constrained CVR experiences by means of mathematical optimization.

The full procedure requires computation of the index values for subsequent short time intervals and ordering the values into time series covering the whole time span of a 360-degree video. The time intervals should be long enough to catch enough fixations to enable calculation of the *sVRCa* values and short enough to approximate the precision of continuous measurement. Half-second intervals met the assumptions in the reported test.

The interpretation of the *sVRCa* values is relative to the intentions of a CVR maker. If the CVR designer had indented to focus people on a specific area or object at a given moment and the *sVRCa* peaked at a high level at that moment, it means success (provided the viewers did not focus on something else, which should be verified with the use of the same 3D ET data). If the creator had wanted the viewers to explore the scene at a given moment, high levels of the visual attention convergence at that moment mean failure and low levels mean success.

IV. Empirical Test

A. Materials

1) 360-degree immersive video

An educational video, aimed at a younger audience, contained a number of short scenes explaining professional work on a traditional 2D film set. The background commentary by a lector described the basic principles of setting the camera, lighting, and working with sound. Other narrative means included mainly changes of scenes, changes of lighting, and positioning of actors and props around the 360-degree scene. The film lasted about 17 minutes.

The tested CVR video and data collection were funded by the National Center for Film Culture (Lodz, Poland). The authors of the current paper had been asked to assess the efficiency of narration in this specific educational video, which gave them the first opportunity to test their recently developed methodology.

2) Software

To operate the experiment (displaying the CVR video, recording 3D ET and headset position data), the VIZARD 6 ENTERPRISE was used. All 3D ET data handling steps (i.e., fixation detection [21][22], data processing and analysis) were executed with the use of R [23] scripts coded from scratch. Gazes of angular speed below 30 degrees per second and lasting longer than 150 milliseconds were classified as fixations.

3) Equipment

HMD HTC VIVE has been equipped with a dedicated SMI eye tracker. 3D ET data were recorded synchronously with information about the location of the headset in 3D virtual space. The sampling frequency has been synchronized with the headset screen refresh rate of 90 Hz.

B. Participants

92 school children, 36 girls (Mage=14.1) and 56 boys, (Mage=14.2) participated in the study with the consent of a parent or a guardian. An ethical committee approved the procedure.

C. Context

The participants watched the 360-degree video separately, one by one, in controlled laboratory conditions. Only two people were present in the laboratory room: a participant and a trained experimenter. No external sounds or tactile stimuli distracted the viewing experience.

D. Results

Figure 1 illustrates the changes in the smoothed (Simple Moving Average over a 5 s window) and non-filtered *sVRCa* values (computed for half-second intervals) over the time span of the entire CVR educational video. Such visualization enables looking at the dynamics of the visual attention convergence results from the bird's eye view and grasping the general attentional pattern shaped by the properties of the narration employed in the immersive video.

We can observe that the visual attention convergence started from very low level; the viewers were looking around and not focused at any specific area of the 360-degree scene.

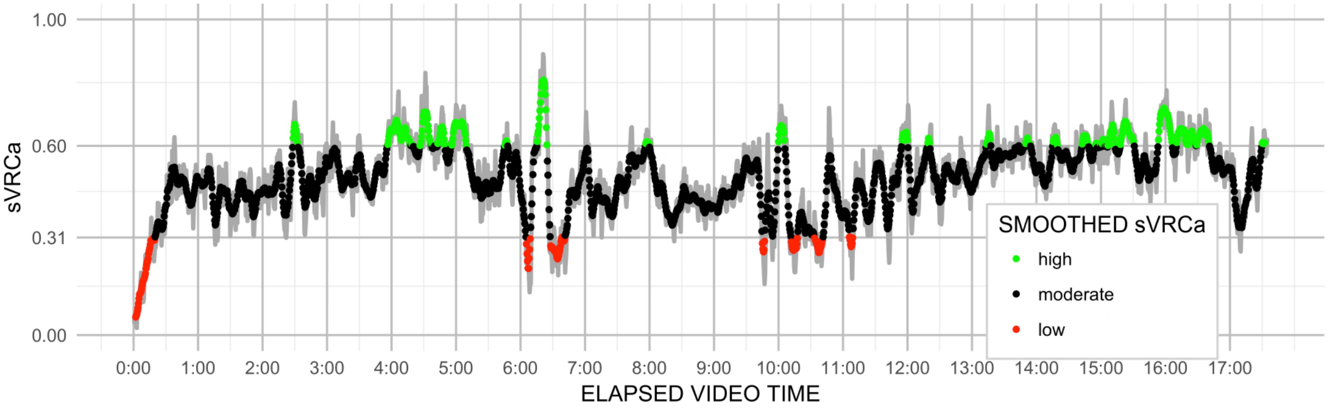

Figure 1. Changes in the values of both the non-filtered (grey) and smoothed (green, black, red) Scaled Aggregated Visual Attention Convergence Index (*sVRCa*) over the time span of the entire CVR educational video.

Then, the visual attention steadily converged towards moderate levels. Further in the video, there were relatively short fragments that provoked high and low levels of the visual attention convergence, between longer fragments characterized by moderate levels.

Table I presents chosen few examples of automatically detected high and low peaks in *sVRCa* time series within the detected fragments of the immersive CVR educational video. High peaks represent high levels of the visual attention convergence between the CVR participants at a given moment of the video. Low peaks represent low levels of the visual attention convergence. Colors denote values chosen for example *ex post* qualitative interpretation (Figures 2-4).

The procedure detected 24 high peaks within 24 short high-value fragments, as well as 7 low peaks within 7 short low-value fragments. The mean length of the automatically detected high-value video fragments was 7 seconds, exactly the same as in the case of the fragments with low *sVRCa*.

TABLE I. CHOSEN EXAMPLES OF DETECTED VIDEO FRAGMENTS AND PEAKS IN THE SCALED AGGREGATED VISUAL ATTENTION CONVERGENCE INDEX (*SVRCA*) TIME SERIES

Video fragment begins at:	**Video fragment ends at:**	**Median *sVRCa***	**Peak *sVRCa***	**Peak type**	**Peak at:**
06:16	06:25	0.75	0.89	**max**	**06:21**
15:18	15:31	0.64	0.75	**max**	**15:23**
04:43	04:50	0.63	0.72	**max**	**04:47**
10:34	10:41	0.27	0.22	**min**	**10:40**
06:28	06:41	0.28	0.19	**min**	**06:35**
06:06	06:09	0.22	0.14	**min**	**06:09**

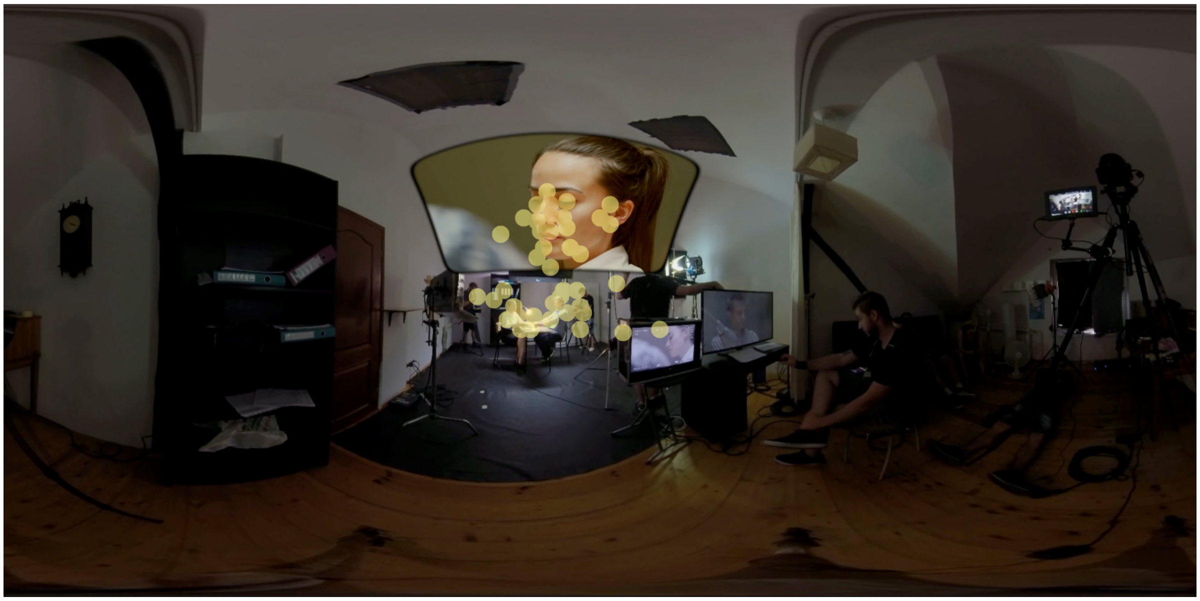

Figure 2. The 4:47 frame corresponds to a high peak in the *sVRCa* (0.72). Yellow dots represent gaze fixations.

The cut-off thresholds for the smoothed *sVRCa* time series were calculated by dividing the empirical range of the original (non-smoothed) *sVRCa* values (0.02 to 0.89) into three even sub-ranges. 0.60 was the resulting value of the threshold above which the smoothed *sVRCa* values were classified as high (green in Figure 1), 0.31 was the value of the threshold below which the smoothed sVRCa values were regarded as low (red in Figure 1). The resulting detection of short video fragments allowed fast and simple determination of (quasi) local maxima and minima (in non-smoothed *sVRCa* time series) within the fragments.

E. Example Ex Post Interpretation of Detected Peaks

The *sVRCa* enabled us to determine candidates for strong and weak moments of the CVR video in the context of narration.

The 4:47 frame (green in Table 1) can be interpreted as a strong moment. A high value of the *sVRCa* (0.72) corresponds to participants' attention focused on actors in the depicted set and on an additional screen positioned above the set (Figure 2). All the distractors, members of the crew, camera, equipment, etc., were hidden in low light areas (we had actually two sets here: the set of the CVR, embracing the "set" of the depicted traditional 2D movie and all the surroundings; similarly actors of the CVR included "actors" and "crew" of the depicted 2D movie).

Another strong moment corresponds this time to a low *sVRCa* value (0.19). The 6:35 frame (blue in Table 1) represented a situation from the beginning of a scene. The stage was being prepared to show the role of lighting in the movies (Figure 3). There were many important elements around a participant. The CVR creator might have expected participants to look around the scene to figure out where the most important elements were.

Low *sVRCa* indicates weak moments as well. In the 10:40 frame (red in Table 1, index value: 0.22, Figure 4), instead of focusing subsequently on elements of lighting being described by the lector one at a time, participants focused on all of the elements of lighting and other, unrelated elements, like moving camera, members of the crew, etc.

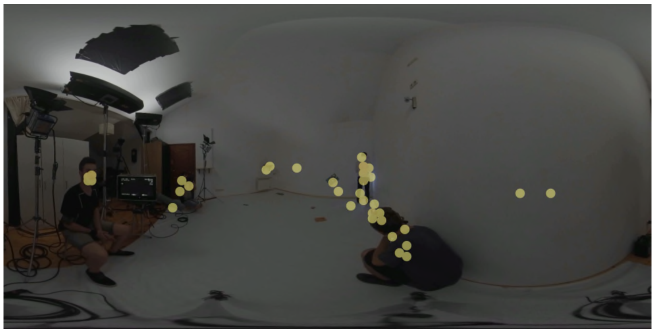

Figure 3. The 6:35 frame corresponds to a low peak in the *sVRCa* (0.19). Yellow dots represent gaze fixations.

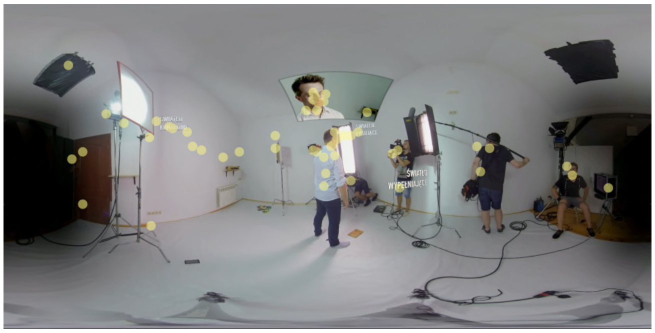

Figure 4. The 10:40 frame corresponds to a low peak in the *sVRCa* (0.22). Yellow dots represent gaze fixations.

V. Conclusion

The initial results of the test are promising. The method seems to substantially augment the process of detection of strong and weak moments in CVR narration. It delivers both the bird's eye view on the changes in reaction to narration and detailed information allowing either automated or point-by-point analysis of specific cuts, fragments, and moments in the immersive 360-degree video.

The authors propose a human-oriented measure, values of which reflect effectiveness of the process of attention directing along a narration line intended by a CVR experience creator. The method measures empirically the inter-viewer convergence in visual attention in order to give CVR creators feedback regarding whether they managed to converge the visual attention or whether they managed to dissipate it at any given moment of the video, according to their original creative intentions. Ideally, it is a CVR creator who should decide the qualitative interpretation of the quantitative measurement.

From a purely technology-oriented perspective, the need for the qualitative interpretation of the quantitative *sVRCa* values might be perceived as a limitation of the proposed method. However, the authors stand on the ground that, in the case of artistic pursuits, it is an artist, not a software system (not even a scientist), who should stay free to draw final conclusions from scientific data and be responsible for all the decisions as to the changes in the narration.

Regarding the next research steps, the authors find it indispensable to correlate the *sVRCa* time series with scenarios for videos, formulated *ex ante* by cooperating video artists in terms of precisely timelined sequence of cuts, attention-guiding cues, and other narrational tricks. Such a triangulation with data extracted from timelined scripts will not only advance the work on the methodology described in the paper. Above all, it will allow fully iterated feedback given by researchers to video makers in the real-life circumstances of CVR creation. It is even conceivable that the iterative feedback paradigm might be further developed into a semi-automatic software system that suggests ways to edit a video according to a desired flow of the *sVRCa* values [24].

The introduction of methodology, such as proposed and described in the paper, seems necessary to help CVR makers in the process of development and validation of the emerging

CVR narration language and means of expression [18]. There are no major objective obstacles for applying the proposed methodology in practice. Headsets equipped with eye trackers are recently available on the consumer market and it is even possible to approximate visual attention data directly from headsets' positions and rotations [25].

The authors hope the method could serve not only as feedback for CVR creators, but also as a criterion for other visual attention measures, physiological indices of attention (e.g., Heart Rate Variability (HRV)) or declarative, quantitative, and qualitative measures.

Acknowledgment

The tested CVR video and data collection were funded by the National Center for Film Culture (Lodz, Poland). The authors gratefully acknowledge the permission to use the data for scientific and publication purposes.

References

[1] J. Blascovich et al., "Immersive virtual environment technology as a methodological tool for social psychology," Psychol. Inq., vol. 13, pp. 103–124, April 2002.

[2] D. M. Markowitz, R. Laha, B. P. Perone, R. D. Pea, and J. N. Bailenson, "Immersive virtual reality field trips facilitate learning about climate change," Front. Psychol, vol. 9, pp. 2364 (1–20), Nov. 2018, doi:10.3389/fpsyg.2018.02364.

[3] P. Kobylinski and G. Pochwatko, "Visual Attention Convergence Index for virtual reality experiences," Human Interaction and Emerging Technologies. IHIET 2019. Adv. Intell. Syst., vol. 1018, Springer, July 2019, pp. 310–316, doi:10.1007/978-3-030-25629-6_48.

[4] P. Kobylinski, G. Pochwatko, and C. Biele, "VR experience from data science point of view: how to measure inter-subject dependence in visual attention and spatial behavior. Intelligent Human Systems Integration 2019. IHSI 2019. Adv. Intell. Syst., vol 903. Springer, Jan. 2019, pp. 393–399, doi: 10.1007/978-3-030-11051-2_60.

[5] I. Bogdanova, A. Bur, and H. Hügli, "Visual attention on the sphere," IEEE T. Image Process., vol. 17, pp. 2000–2014, Nov. 2008, doi:10.1109/TIP.2008.2003415.

[6] S. Croci, S. Knorr, L. Goldmann, and A. Smolic, "A framework for quality control in cinematic VR based on Voronoi Patches and saliency," 2017 International Conference on 3D Immersion (IC3D 2017), IEEE, Jan. 2018, pp. 1–8, doi:10.1109/IC3D.2017.8251907.

[7] M. Xu, C. Li, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," IEEE T. Circ. Syst. Vid, vol. 29, pp. 3516–3530,, Dec. 2019, doi:10.1109/TCSVT.2018.2886277.

[8] J. Gutiérrez-Cillán, E. J. David, A. Coutrot, M. P. Da Silva, and P. Le Callet, "Introducing UN Salient360! Benchmark: a platform for evaluating visual attention models for 360-degree contents," 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX 2018), IEEE, Sept. 2018, pp. 1–3, doi:10.1109/QoMEX.2018.8463369.

[9] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," 2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX 2018), IEEE, June 2018, pp. 1–6, doi:10.1109/QoMEX.2018.8463418.

[10] E. Upenik and T. Ebrahimi, "Saliency driven perceptual quality metric for omnidirectional visual content," 2019 IEEE International Conference on Image Processing (ICIP 2019), IEEE, August 2019, pp. 4335–4339, doi:10.1109/ICIP.2019.8803637.

[11] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," Proceedings Volume 9970. Optics and Photonics for Information Processing X, SPIE, Sept. 2019, pp. 9970–11, doi:10.1117/12.2235885.

[12] N. Carlsson and D. Eager, "Had you looked where I'm looking: cross-user similarities in viewing behavior for 360-degree video and caching implications," Available from: https://arxiv.org/abs/1906.09779 (accessed Dec. 2019).

[13] I. Bogdanova, A. Bur, H. Hügli, and P.-A. Farine, "Dynamic visual attention on the sphere," Comput. Vis. Image Und., vol. 114, pp. 100–110, Jan. 2010, doi:10.1016/j.cviu.2009.09.003.

[14] V. Sitzmann et al., "Saliency in VR: how do people explore virtual environments?," IEEE T. Vis. Comput. Gr., vol. 24, pp. 1633–1642, Jan. 2018, doi:10.1109/TVCG.2018.2793599.

[15] C. O. Fearghail, C. Ozcinar, S. Knorr, and A. Smolic, "Director's cut - analysis of aspects of interactive storytelling for VR films," Interactive Storytelling. ICIDS 2018. Lect. Notes Comput. Sc., vol. 11318, Springer, Nov. 2018, pp. 308–322, doi: 10.1007/978-3-030-04028-4_34.

[16] S. Rothe, D. Buschek, and H. Hussmann, "Guidance in cinematic virtual reality - taxonomy, research status and challenges," Multimodal Technologies and Interaction, vol. 3, pp. 1–23, March 2019, doi:10.3390/mti3010019.

[17] A. Sheikh, A. Brown, Z. Watson, and M. Evans, ".Directing attention in 360-degree video," IBC 2016 Conference, IET Digital Library, Sept. 2016, pp. 29(9)–29(9), doi:10.1049/ibc.2016.0029.

[18] J. S. Pillai and M. Verma, "Grammar of VR storytelling: narrative immersion and experiential fidelity in VR cinema," The 17th International Conference on Virtual-Reality Continuum and Its Applications in Industry (VRCAI 2019), ACM, Nov. 2019, pp. 34(1)–34(6), doi: 10.1145/3359997.3365680.

[19] U. Świerczyńska-Kaczor, M. Żelazowska, M. Kotlińska, and J. Wachowicz, "Online interactive storytelling: evaluation of the viewer experience of 360-degree videos," Journal of Economics and Management, vol. 36, pp. 105–122, Feb. 2019, doi: 10.22367/jem.2019.36.06.

[20] D. Dowling, C. O. Fearghail, A. Smolic, and S. Knorr, "Faoladh: a case study in cinematic VR storytelling and production" Interactive Storytelling. ICIDS 2018. Lect. Notes Comput. Sc, vol. 11318, Springer, Nov. 2018, pp. 359–362, doi:10.1007/978-3-030-04028-4_42.

[21] A. T. Duchowski, Eye Tracking Methodology: Theory and Practice, Springer, 2007.

[22] D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols,". ETRA '00 Proceedings of the 2000 Symposium on Eye Tracking Research and Applications, ACM, Nov. 2000, pp. 71–78, doi:10.1145/355017.355028.

[23] R Core Team, A Language and Environment for Statistical Computing, 2019, Available from: https://www.R-project.org (retrieved Dec. 2019).

[24] B. Huber, H. V. Shin, B. Russell, O. Wang, and G. J. Mysore, "B-script: Transcript-based B-roll video editing with recommendations," CHI '19: Proceeedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, May 2019, pp. 81(1)–81(11), doi:10.1145/3290605.3300311.

[25] E. Upenik and T. Ebrahimi, "A simple method to obtain visual attention data in head mounted virtual reality," 2017 IEEE International Conference on Multimedia & Expo Workshops, IEEE, July 2017, pp. 73–78, doi:10.1109/ICMEW.2017.8026231.