On the link between emotion, attention and content in virtual immersive environments

Quentin Guimard, Florent Robert, Camille Bauce, Aldric Ducreux, Lucile Sassatelli, Hui-Yin Wu, Marco Winckler, Auriane Gros

ON THE LINK BETWEEN EMOTION, ATTENTION AND CONTENT IN VIRTUAL IMMERSIVE ENVIRONMENTS

Quentin Guimard[⋆], *Florent Robert*[⋆†], *Camille Bauce*[⋆], *Aldric Ducreux*[⋆], *Lucile Sassatelli*[⋆§], *Hui-Yin Wu*[†], *Marco Winckler*[⋆†], *Auriane Gros*[‡]

[⋆]Université Côte d'Azur, CNRS, I3S; [†]Université Côte d'Azur, Inria;
[‡]Université Côte d'Azur, CHU de Nice, CoBTeK; [§]Institut Universitaire de France

ABSTRACT

While immersive media have been shown to generate more intense emotions, saliency information has been shown to be a key component for the assessment of their quality, owing to the various portions of the sphere (viewports) a user can attend. In this article, we investigate the tri-partite connection between user attention, user emotion and visual content in immersive environments. To do so, we present a new dataset enabling the analysis of different types of saliency, both low-level and high-level, in connection with the user's state in 360° videos. Head and gaze movements are recorded along with self-reports and continuous physiological measurements of emotions. We then study how the accuracy of saliency estimators in predicting user attention depends on user-reported and physiologically-sensed emotional perceptions. Our results show that high-level saliency better predicts user attention for higher levels of arousal. We discuss how this work serves as a first step to understand and predict user attention and intents in immersive interactive environments.

***Index Terms*—** 360° videos, saliency maps, emotions, physiological signals, gaze

1. INTRODUCTION

Immersive media and environments are on the rise with increased affordability of virtual reality (VR) equipment and the deployment of popular platforms for 360° streaming or advanced interaction in the *Metaverse*[1]. This new type of content questions the design of compelling immersive experiences, as well as the technical choices for storage and distribution. Visual quality assessment by humans is key to enable efficient storage and distribution of content with high perceptual quality [1]. To apply quality assessment-driven processes (such as compression, streaming decisions, etc.) to any video content, it is crucial to automate quality assessment with quality estimators associated to the content. For immersive media specifically, quality assessment depends on the sequence of viewports attended by the human rater, which can be significantly different from one rater to the other. For example, Xu et al. [2] consider content saliency and viewport preference to extend the PSNR and SSIM metrics to 360° content. The accuracy of quality estimators for immersive content therefore strongly depends on the accuracy of content-based saliency estimators predicting patterns of human attention. Saliency is hence a key component of quality assessment of immersive media. On the other hand, immersive content have been shown to elicit more intense emotions compared with regular flat-screen presentations [3, 4, 5].

In this article, we investigate how emotions are associated with head motion to impact the accuracy of content-based saliency estimators. To the best of our knowledge, this is the first article to investigate the tri-partite connection between user attention, user emotion and visual content in immersive environments. Our contributions are:

- A new dataset from user experiments recording head and gaze movements, along with self-reports and continuous physiological measurements of emotions. The stimuli are 360° videos selected to enable the analysis of different types of saliency in connection with the user's state. We verify the consistency of our results. The dataset is publicly available[2].
- An investigation into how emotions affect the accuracy of saliency estimators, both with low-level and high-level saliency, in 360° videos. Our results show that high-level saliency better predicts user attention for higher levels of arousal.

2. RELATED WORK

Sensing and analyzing emotions in immersive environments has spurred interest (see, e.g., [3, 4, 5, 6]), more recently coupled with motion recordings and analysis [7, 8, 9, 10, 11].

Human emotions are commonly decomposed along two main dimensions: valence, representing the negative or posi-

This work has been partly supported by the French government, through the UCA JEDI and EUR DS4H Investments in the Future projects ANR-15-IDEX-0001 and ANR-17-EURE-0004. This work was partly supported by EU Horizon 2020 project AI4Media, under contract no. 951911 (https://ai4media.eu/).

[1] https://www.cnbc.com/2021/12/27/metas-oculus-virtual-reality-headsets-were-a-popular-holiday-gift.html

tive nature of an emotion (unpleasant-pleasant), and arousal, representing the intensity of the perceived emotion (calm-excited) [12]. The first reference database [7] providing emotional ratings and motion recordings of 360° videos is made of 73 VR videos on which 95 users rated valence and arousal using the self-assessment manikin (SAM) tool [13] after experiencing each video. Their head positions were continuously recorded. A dataset of self-reported emotions of 19 users watching thirty-six 360° images is collected by Tang et al. [8], with eye motion recorded. However, ratings made in retrospect cannot represent the variety of states a user goes through during the experience [6], limiting potential analyses and interpretations. Recent works have therefore proposed tools enabling a continuous collection of self-reports inside the immersive environment [9, 10]. The data collected in these recent works also comprise physiological measurements of heart rate and electrodermal activity (EDA, as skin conductance), which has been shown to reliably represent user instantaneous arousal [14]. Understanding how different types and levels of emotions correspond to specific types of motion has already been investigated [7, 8, 11]. Results from Li et al. [7] show some level of correlation between (time) average arousal and average pitch angle, and between yaw angle standard deviation and valence, while results from Tang et al. [8] show a significant impact of negative images on eye behavior.

While above works have focused on the analysis of user emotion and motion based on coarse-grained categorization of the entire content (high/low positive/negative valence and high/low arousal), other works have focused on the impact of specific regions on the user's attention, described with low-level (LL) saliency or emotional aspects. LL saliency refers to pixel-level features (e.g., edges, luminance, motion). Cerf et al. [15] showed that human eye movements are influenced both by LL and high-level (HL) saliency (related to higher semantic concepts such as objects and faces). Chaabouni et al. [16] showed that normalizing fixations density with LL saliency significantly improves the interest estimators based on gaze data. Hedger et al. [17] re-examined previous results suggesting that emotional faces in an image attract more user attention/fixations outside awareness. They showed that facial expressions had no effect on attentional allocation, which can instead be explained by the higher LL saliency.

In this article, we present a first step towards understanding the connection between user emotion and predictability of motion from content saliency. Specifically,we analyze how the accuracy of LL and HL saliency estimators depend on the user's self-reported and physiologically-sensed emotional perceptions.

3. MATERIAL AND METHODS

Dataset We present the dataset we have collected to analyze saliency accuracy in Sec. 4. This dataset is publicly available[2]. It is composed of user head and eye movements recorded while watching 360° videos in a VR headset, along with users' EDA and heart rate (HR) during viewing, and valence and arousal ratings collected at the end of every clip. The user experiment has been approved by the university ethics committee.

Stimuli In order to investigate the differences in accuracy of LL and HL saliencies depending on user emotions, we select seven videos from the database provided by Li et al. [7] that meet two criteria. First, the videos must span an as broad as possible range of (valence, arousal). Video details are provided here[2]. Second, HL and LL saliencies must not always overlap. LL and HL saliencies are computed on 100 "patches" made of overlapping projections centered on points uniformly sampled from each frame of the video. LL saliency is computed using the Itti model [18] combined with optical flow between consecutive frames on these individual patches. Inspiring from Chopra et al. [19], HL saliency is obtained from YOLOv4 object detector on these same patches, object bounding boxes are used as binary saliency maps. These patches are then back-projected on the equirectangular frame, by addition of the overlapping patches. To select videos where the HL and LL saliencies do not overlap systematically but are rather balanced inside and outside objects, we compare (i) the number of pixels inside and outside objects, and (ii) the per-pixel LL saliency (ranging between 0 and 255), computed as the total LL saliency inside and outside objects normalized with the corresponding number of pixels. Fig. 1 exemplifies that in video 12. The number of pixels with such minimum LL saliency inside and outside objects is equivalent over time, as is the per-pixel LL saliency in both areas. Fig. 2 shows a frame where regions with high LL saliency can be seen outside the detected objects.

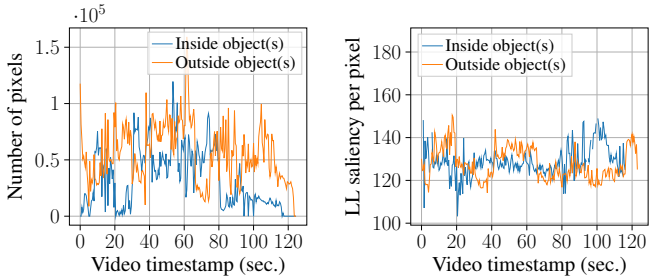

Fig. 1: HL and LL saliency characterization of Video 12. Left: number of pixels inside and outside objects. Right: average LL saliency per pixel inside and outside objects.

Methodology Recordings of head and eye movements have been made with a FOVE headset, equipped with an eye-tracker with a 120Hz acquisition rate, and tethered to a desktop computer. The video is played in an application

[2] https://gitlab.com/PEM360/PEM360

Fig. 2: HL and LL saliency visualization for frame 1545 of video 12. Left: the frame. Center: HL saliency (detected objects are elephants). Right: LL saliency.

developed in Unity with the FOVE SDK. Recordings of EDA and optical pulse have been made with a Shimmer3 GSR+ unit with a frequency range of 15.9Hz. All of the measurements were resampled to 100Hz for analysis. The lab experiment involved 31 users (10f, 20m, 1nb; 18-29 years old, M=24, SD=3.26). First, participants were presented a pre-questionnaire to assess their background with VR and checking for visual deficiencies. Next, the VR experiment systematically started with a low-arousal (relaxing) video (ID 32) to bring EDA and HR levels to a user-relative baseline. Then, the remaining six VR videos were experienced in a random order by every user. Finally, after each video, the headset was taken off and the SAM scale presented for arousal and valence rating. An at least 1-min break outside the headset was observed between every video. All the videos were played without sound.

Preliminary analysis After visual inspection to remove erroneous EDA data, the phasic component is extracted using cvxEDA as implemented by Neurokit [20]. The skin conductance response (SCR) is finally obtained as the absolute value of the first derivative of the phasic component [10]. The reliability of arousal and valence ratings is assessed by the intra-class correlation coefficient (ICC), where a class of measurements corresponds to a stimulus (360° video). ICC estimates based on mean ratings with a two-way mixed effects model are 0.96 (95% CI 0.87-0.99) for arousal and 0.88 (95% CI 0.72-0.98) for valence. The median root square difference of valence-arousal ratings with the corresponding values available in the original dataset [7] is 1.17 (range: 1-9), showing the agreement between both. Finally, we look at the correspondence between SCR and arousal ratings. Similar to Toet et al. [9], for each video, we compute the mean (resp. median) of SCR across users. The results show that the video rankings according to mean arousal and to mean SCR are very close (not shown here due to space limitation).

4. ANALYSIS

Our objective is to compare the accuracy of both types of saliency maps, HL and LL, to match the users' fixations over every frame of the 360° video. To do so, we compute the normalized scanpath saliency (NSS), which measures the amount of saliency around fixations [21]. We consider segments of 5 sec. to average the saliency maps of all frames and aggregate the user's fixations in this interval, hence obtaining an NSS value for both saliency types $\text{NSS}^{HL}_{u,v,i}$ and $\text{NSS}^{LL}_{u,v,i}$ for every user u, video v, and interval i. The averages over intervals (resp. users) are denoted by $\text{NSS}_{u,v}$ and NSS_v, respectively. We analyze the association between $\text{NSS}^{HL}_{u,v}$ and $\text{NSS}^{LL}_{u,v}$ with mean centered SCR denoted $\text{cSCR}_{u,v}$ and graded arousal $\text{GA}_{u,v}$. SCR is centered per user with $\text{cSCR}_{u,v} = \text{SCR}_{u,v} - E_v[\text{SCR}_{u,v}]$ because the preliminary analysis has shown that the absolute levels of SCR vary significantly across users, but intra-user variations across videos are consistent with the ordering of each user's arousal ratings.

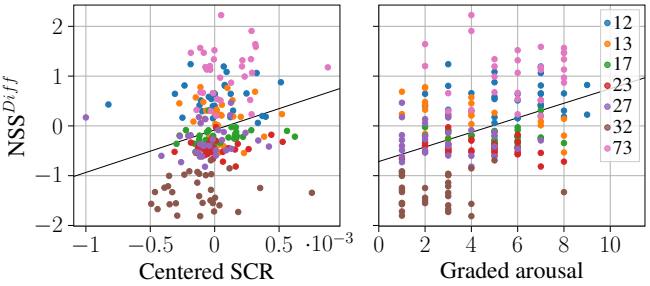

Fig. 3: $\text{NSS}^{Diff}_{u,v}$ against $\text{cSCR}_{u,v}$ and $\text{GA}_{u,v}$ for every user u and video v. The black line shows a linear regression model fitted on the data.

To analyze the difference in accuracy of both types of saliency depending on the user's arousal, we consider in Fig. 3 the difference $\text{NSS}^{Diff}_{u,v} = \text{NSS}^{HL}_{u,v} - \text{NSS}^{LL}_{u,v}$ plotted against $\text{cSCR}_{u,v}$ (left) and graded arousal $\text{GA}_{u,v}$ (right) for all u, v, the points being colored per video. The major finding is the increasing trend of NSS^{Diff} with EDA and graded arousal. Specifically, the PCC between NSS^{Diff} and EDA cSCR is 0.25 ($p < 10^{-3}$), and the PCC between NSS^{Diff} and graded arousal $\text{GA}_{u,v}$ is 0.41 ($p < 10^{-9}$). These estimates are obtained over 217 (u, v) samples. According to Walline [22, Appendix 6C, page 79], such levels of correlation are significant for 123 and 44 samples, respectively (see [23]).

We then analyze the same associations averaged per video in Fig. 4, where the x-axis of the first row is cSCR_v and that of the second row is GA_v, with v in the set of video indices. The columns are numbered from the left. We first confirm from the leftmost column that ordering and appearance of NSS^{Diff}_v against EDA or graded arousal are close. Second, we observe a clear increasing trend confirming the above positive significant correlation results.

To investigate the reasons for this trend, we decompose NSS^{Diff}_v into its individual components NSS^{HL}_v and NSS^{LL}_v depicted in columns 2 and 3. Owing to the similarity of trends against EDA and graded arousal, we conduct the analysis only on the latter. We first observe an increasing trend of NSS^{HL}_v. It could have been even clearer considering that underwater objects in video 12 (brown dot) are often missed by the ob-

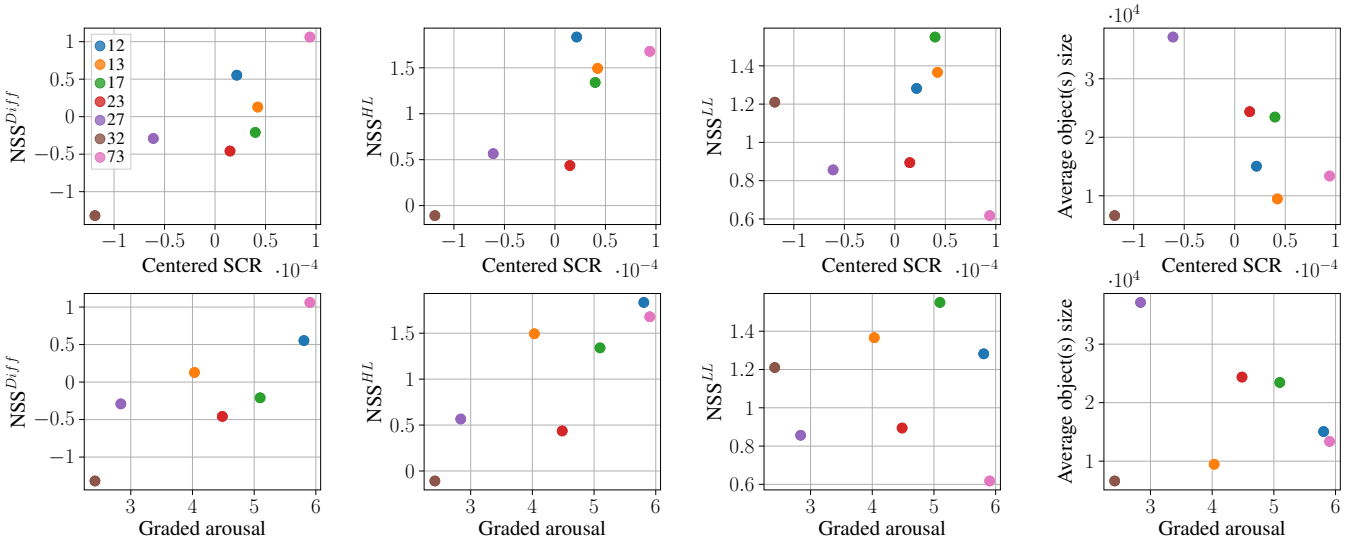

Fig. 4: From left to right: NSS_v^{Diff}, NSS_v^{HL}, NSS_v^{LL} and average number of pixels inside objects against cSCR_v (top) and GA_v (bottom) for all videos v.

ject detector (large shark), hence under-estimating NSS_{12}^{HL}. We can then question whether this increase is due to users focusing more when more aroused, or to intrinsic features of the videos, where larger objects would appear in higher arousal videos. We verify in the last column that the increase in NSS^{HL} with arousal cannot be entirely attributed to a relatively larger area occupied by objects. Second, column 3 shows no clear trend. The variation in NSS^{LL} does not appear to be related to EDA or graded arousal.

We can therefore conclude that the increasing trend of NSS^{Diff} with arousal is mainly due to higher NSS^{HL} for higher-arousal videos. A first conclusion we may draw is that the relative weight of HL saliency should vary in a saliency model depending on the user's arousal state. Sensing the user's state may hence help predict their attention.

5. DISCUSSION

We cannot claim causation on whether users focus because they are more aroused by the content, or if they are more aroused because they focus on objects. A first question is whether significantly different levels of arousal occur for users on the same video. This is part of future work. This would mean that the video content alone is not informative enough to adapt the relative weights of HL and LL saliency to users. On the contrary, if the video content is sufficient, then one can think of leveraging arousal (physiological or subjective) measurements in quality assessment sessions to serve as an auxiliary loss to train (deep) saliency models.

While arousal and valence are major dimensions to describe user emotion of a given content like a video, the richer experience of an immersive and possibly interactive environment is described over various additional dimensions, particularly presence, immersion, agency, engagement, flow, usability, skill or judgement [24]. Recently, valence, arousal and agency have been shown to interact in non-trivial ways to produce presence [25]. In 6DoF environments, which we are currently investigating for rehabilitation scenarios and where engagement, skills and judgments are major outcomes, it is crucial to adapt the environment's content to provide proper adaptive guidance to the user. This requires an understanding and the prediction of the user's attention and intents, which depend on the user's emotional state. This work in 3DoF immersive low-interaction environment hence serves as a baseline for immersive interactive environments.

6. CONCLUSION

In this article, we have first introduced a new dataset of user head and gaze movements in 360° videos with valence and arousal ratings, and continuous physiological measurements of skin conductance and heart rate. The stimuli have been specifically selected to enable a spatio-temporal analysis in relation to content, user motion and emotions. We have presented first results comparing HL and LL saliency accuracy depending on user arousal, showing that the accuracy of HL saliency increases when user arousal increases.

Next steps will consist in investigating finer temporal associations of saliency and attention locations with arousal/EDA, as well as structural modeling of possible interacting factors (e.g., valence, fear, agency) in the production of head and gaze patterns. Also, the accuracy of more refined saliency models such as deep neural networks explicitly or implicitly combining saliency levels will be assessed to better understand motion predictability depending on the user's emotional state.

7. REFERENCES

[1] Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hossfeld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, Bob Lawlor, Patrick Le Callet, Sebastian Möller, Fernando Pereira, Manuela Pereira, Andrew Perkis, Jesenka Pibernik, Antonio Pinheiro, Alexander Raake, Peter Reichl, Ulrich Reiter, Raimund Schatz, Peter Schelkens, Lea Skorin-Kapov, Dominik Strohmeier, Christian Timmerer, Martin Varela, Ina Wechsung, Junyong You, and Andrej Zgank, "Qualinet White Paper on Definitions of Quality of Experience," Mar. 2013, Qualinet White Paper on Definitions of Quality of Experience Output from the fifth Qualinet meeting, Novi Sad, March 12, 2013.

[2] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan, "Assessing visual quality of omnidirectional videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 12, pp. 3516–3530, 2019.

[3] Rosa María Baños, Cristina Botella, Isabel Rubió, Soledad Quero, Azucena García-Palacios, and Mariano Luis Alcañiz Raya, "Presence and emotions in virtual environments: The influence of stereoscopy," *Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society*, vol. 11 1, pp. 1–8, 2008.

[4] Anna Felnhofer, Oswald D. Kothgassner, Mareike Schmidt, Anna-Katharina Heinzle, Leon Beutl, Helmut Hlavacs, and Ilse Kryspin-Exner, "Is virtual reality emotionally arousing? investigating five emotion inducing virtual park scenarios," *Int. J. Hum.-Comput. Stud.*, vol. 82, no. C, pp. 48–56, oct 2015.

[5] Federica Pallavicini, Alessandro Pepe, and Maria Eleonora Minissi, "Gaming in virtual reality: What changes in terms of usability, emotional response and sense of presence compared to non-immersive video games?," *Simulation & Gaming*, vol. 50, no. 2, pp. 136–159, 2019.

[6] Jan-Niklas Voigt-Antons, Eero Lehtonen, Andres Pinilla Palacios, Danish Ali, Tanja Kojic, and Sebastian Möller, "Comparing Emotional States Induced by 360° Videos Via Head-Mounted Display and Computer Screen," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, 2020, pp. 1–6.

[7] Benjamin J. Li, Jeremy N. Bailenson, Adam Pines, Walter J. Greenleaf, and Leanne M. Williams, "A Public Database of Immersive VR Videos with Corresponding Ratings of Arousal, Valence, and Correlations between Head Movements and Self Report Measures," *Frontiers in Psychology*, vol. 8, pp. 2116, Dec. 2017.

[8] Wei Tang, Shiyi Wu, Toinon Vigier, and Matthieu Perreira Da Silva, "Influence of Emotions on Eye Behavior in Omnidirectional Content," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, Athlone, Ireland, May 2020, pp. 1–6, IEEE.

[9] Alexander Toet, Fabienne Heijn, Anne-Marie Brouwer, Tina Mioch, and Jan B. F. van Erp, "An Immersive Self-Report Tool for the Affective Appraisal of 360° VR Videos," *Frontiers in Virtual Reality*, vol. 1, pp. 552587, Sept. 2020.

[10] Tong Xue, Abdallah El Ali, Tianyi Zhang, Gangyi Ding, and Pablo Cesar, "CEAP-360VR: A Continuous Physiological and Behavioral Emotion Annotation Dataset for 360 VR Videos," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[11] Tong Xue, Abdallah El Ali, Gangyi Ding, and Pablo Cesar, "Investigating the Relationship between Momentary Emotion Self-reports and Head and Eye Movements in HMD-based 360° VR Video Watching," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan, May 2021, pp. 1–8, ACM.

[12] Lisa Feldman Barrett, "Discrete emotions or dimensions? The role of valence focus and arousal focus.," *Cognition and Emotion*, vol. 12, no. 4, pp. 579–599, 1998, Place: United Kingdom Publisher: Taylor & Francis.

[13] Margaret M. Bradley and Peter J. Lang, "Measuring emotion: The self-assessment manikin and the semantic differential," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[14] Wolfram Boucsein, *Electrodermal activity, 2nd ed.*, Electrodermal activity, 2nd ed. Springer Science + Business Media, New York, NY, US, 2012, Pages: xviii, 618.

[15] Moran Cerf, Jonathan Harel, Wolfgang Einhaeuser, and Christof Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems*, J. Platt, D. Koller, Y. Singer, and S. Roweis, Eds. 2007, vol. 20, Curran Associates, Inc.

[16] Souad Chaabouni and Frederic Precioso, "Impact of Saliency and Gaze Features on Visual Control: Gaze-Saliency Interest Estimator," in *Proceedings of the 27th ACM International Conference on Multimedia*, Nice France, Oct. 2019, pp. 1367–1374, ACM.

[17] Nicholas Hedger, Matthew Garner, and Wendy J. Adams, "Do emotional faces capture attention, and does this depend on awareness? Evidence from the visual probe paradigm.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 45, no. 6, pp. 790–802, June 2019.

[18] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[19] Lovish Chopra, Sarthak Chakraborty, Abhijit Mondal, and Sandip Chakraborty, "PARIMA: Viewport Adaptive 360-Degree Video Streaming," in *Proceedings of the Web Conference 2021*. 2021, pp. 2379–2391, ACM.

[20] Dominique Makowski, Tam Pham, Zen J. Lau, Jan C. Brammer, François Lespinasse, Hung Pham, Christopher Schölzel, and S. H. Annabel Chen, "NeuroKit2: A python toolbox for neurophysiological signal processing," *Behavior Research Methods*, vol. 53, no. 4, pp. 1689–1696, feb 2021.

[21] Olivier Le Meur and Thierry Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods*, vol. 45, no. 1, pp. 251–266, Mar. 2013.

[22] Jeffrey J. Walline, "Designing Clinical Research: an Epidemiologic Approach, 2nd Ed.," *Optometry and Vision Science*, vol. 78, no. 8, 2001.

[23] UCSF, "Sample size calculators for designing clinical research," https://sample-size.net/correlation-sample-size/, 2021.

[24] Katy Tcha-Tokey, Olivier Christmann, Emilie Loup-Escande, and Simon Richir, "Proposition and Validation of a Questionnaire to Measure the User Experience in Immersive Virtual Environments," *International Journal of Virtual Reality*, vol. 16, no. 1, pp. 33–48, Jan. 2016.

[25] Crescent Jicol, Chun Hin Wan, Benjamin Doling, Caitlin H Illingworth, Jinha Yoon, Charlotte Headey, Christof Lutteroth, Michael J Proulx, Karin Petrini, and Eamonn O'Neill, "Effects of Emotion and Agency on Presence in Virtual Reality," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, Yokohama Japan, May 2021, pp. 1–13, ACM.