Multimodal Dataset for Assessment of Quality of Experience in Immersive Multimedia

Anne-Flore Perrin, He Xu, Eleni Kroupi, Martin Řeřábek, Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Federale de Lausanne (EPFL)
Lausanne, Switzerland

ABSTRACT

This paper presents a novel multimodal dataset for the analysis of Quality of Experience (QoE) in emerging immersive multimedia applications. In particular, the perceived Sense of Presence (SoP) induced by one-minute long video stimuli is explored with respect to content, quality, resolution, and sound reproduction and annotated with subjective scores. Furthermore, a complementary analysis of the recorded physiological signals, such as EEG, ECG, and respiration is carried out, aiming at an alternative evaluation of human experience while consuming immersive multimedia content. Results confirm the value of the introduced dataset and its consistency for the purposes of QoE assessment for immersive multimedia. More specifically, subjective ratings demonstrate that the created dataset enables distinction between low and high levels of immersiveness, which is also confirmed by a preliminary analysis of recorded physiological signals.

Categories and Subject Descriptors

H.2.4 [**Information Systems**]: Database management—*Systems, Multimedia database*; H.5.1 [**Information Systems**]: Information Interfaces and presentation—*Multimedia Information Systems ; Evaluation/methodology, Video*

Keywords

Sense of Presence (SoP); Immersive multimedia; Quality of Experience (QoE); EEG; ECG; Respiration; Subjective assessment

1. INTRODUCTION

The Sense of Presence (SoP) also known as Immersiveness Levels (ILs) in this paper, is a desired quality metric for immersive environments [19]. According to [23], SoP refers to the subjective experience of "leaving" the intrinsic world and "being present" in a virtual environment. Detailed definition and description of SoP including its measurement can be found in [15, 23].

Multimedia technologies aim at providing higher Quality of Experience (QoE), through combination of sensory, in particular audio and visual information. For instance, audiovisual content in TV can be regarded as a limited virtual environment replicating physical reality. Indeed, sensory cues for virtual environments usually consist of visual but also audio information. Therefore, it is necessary to better investigate and understand the influence of both modalities and their impact on the QoE. The SoP, being the most significant quality metric of virtual environment, is then expected to be highly correlated to the QoE. Hence, the SoP is investigated here. Various studies in television services using subjective ratings as assessment of SoP are presented in [7].

Any explicit subjective assessment by human subjects is influenced by emotional, cultural, educational, and environmental differences across subjects [6, 8]. More importantly, explicit assessments can have an impact on the experience itself, and interfere with it. Therefore, implicit assessment based on subjects physiological signals is expected to provide additional and less biased information, when compared to explicit assessment. In [21], an objective measure of SoP using behavior or physiological responses is envisioned, nevertheless the authors used only subjective ratings. Electroencephalography (EEG) and peripheral physiological signals including Electrocardiography (ECG) and respiration have been used in various studies to assess the physiological responses to stimuli. In particular, in [12], 2D and 3D overall perceived quality is assessed using brain and peripheral signals, [11] concludes that perceived quality is related to emotional processes, whereas [10, 20] analyze emotional experiences as opposed to SoP. The authors of [1] present a similar sensor-based assessment of SoP using galvanic skin response, EEG, and facial motion tracking.

This paper presents a novel dataset that captures the differences in user experience during multimedia stimuli with various ILs. In this dataset the different ILs are induced by modifying the resolution, the amount of compression in the video, and sound reproduction. EEG and peripheral physiological signals including ECG and respiration, as well as subjective ratings are acquired during experiments.

Analysis of subjective ratings indicates that stimuli properties, such as presence of audio and its quality, are correlated with SoP. Additionally, proper application of classification algorithms can distinguish the various ILs, based on EEG and other peripheral physiological signals.

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806387.

Parameters Setting	Immersiveness Levels (ILs)		
	Low	Middle	High
Audio	No Audio	Stereo	Surround
Quality (QP)	36	20	20
Resolution	SD	HD	UHD

Table 1: Parameters setting for different ILs

The remainder of this paper is organized as follows. The next section describes how we conducted experiments to collect subjective ratings and physiological responses. Section 3 presents the results of a subjective rating analysis and of the classification using physiological signals. Finally, conclusions are provided in Section 4.

2. DATA COLLECTION

Four uncompressed (FLAC, RAW YUV 4:2:0) audiovisual contents[1] from Blender open source project[2], two originally in HD (Big Buck Bunny and Elephant Dream) and two in UHD resolutions (Sintel and Tears of Steel), were used to extract ten one-minute audiovisual sequences. Nine of these sequences were used for test stimuli preparation, whereas one sequence was used for training stimuli creation. The audiovisual sequences were selected and cut from the original contents based on the audio energy level in surround channels, as well as on spatial and temporal properties of the video. More specifically, parts of the contents with the highest values of the above-mentioned properties were used, while the original movie scene editing was respected. Low, middle, and high ILs were defined for each audiovisual sequence by experts based on the exploited audio sound system (mono, stereo, and 5.1), the video quality/level of compression (high and low QP, using x264 encoding), and the resolution (UHD, HD, and SD). The Table 1 illustrates the characteristic settings for the three defined ILs. The combination of the nine video sequences with the three ILs leads to 27 video stimuli shown to each subject during each experiment.

The professional high-performance 4K/QFHD LCD reference 56-inch monitor Sony Trimaster SRM-L560 was used to display the video stimuli. As recommended in [16], the viewing distance was set at 1.6 times the height of the screen. The Altec Lansing 5.1 THX speaker system, super subwoofer was used as audio sound system. The laboratory setup provided a quiet environment and the ambient light was set in order to ensure subject comfort during bright and dark scenes, as well as during assessment and resting periods. To record the brain activity, a 256-electrodes net was placed at the standard position on the scalp. An EGI's Geodesic EEG System (GES) 300 was used to record, amplify, and digitize the EEG signals while the participants were watching the stimuli. The heart activity was recorded from two standard ECG electrodes placed on the lower left rib cage and the upper right clavicle. Two respiratory inductive plethysmography belts (thoracic and abdomen) were used to acquire the respiration. All signals were recorded at 250 Hz.

Eight female and twelve male subjects participated in the study. They were from 18 to 30 years old (23 in average with a 3.03 standard deviation). The 20 subjects were screened for correct visual acuity (no errors on 20/30 lines) and color vision using Snellen and Ishiara charts respectively. They all provided written consent forms. Before each experiment, oral instructions were provided to the participants to explain their tasks. Additionally, a training session was used to illustrate the low, middle, and high levels in SoP to guide subjects to bound their own perceived overall ratings.

The experiments consisted of three sessions intersected by ten-minute breaks in order to prevent subject fatigue and lack of attention. The audiovisual stimuli were displayed in low-middle-high IL, middle-low-high IL, and high-middle-low IL order during first, second, and third sessions, respectively. In order to avoid boredom or fatigue, the same audiovisual sequence was displayed only once during each session. Thus, nine audiovisual stimuli were presented in each session leading to a total of 27 audiovisual stimuli forming 27 trials.

Each trial consisted of a ten-second baseline period and a stimulus period. The physiological signals recorded during the baseline period were used to remove stimulus-unrelated variations from the signals obtained during the stimulus period. During the baseline periods, the subjects were instructed to remain calm and focus on a 2D white cross on a black background presented on the screen in front of them. Once this baseline period was over, a video stimulus was presented. After the video sequence was over, subject was asked to provide his/her self-assessed ratings for the particular video sequence without any restriction in time, following the Absolute Category Rating (ACR) evaluation methodology [9]. Once a trial was over, the next baseline period was recorded and the next video sequence, whose content was randomly selected, was presented. The procedure was repeated until all 27 video stimuli were presented and rated. Although the experiments lasted for almost two hours, including the training and the set up, the subjects did not report fatigue.

Regarding the self-assessed ratings, subjects were asked to evaluate the video sequences according to five different criteria, namely, interest in the video content, perceived video quality, interest in the audio content, level of immersiveness, and awareness of their surrounding. A 9-point rating scale was used ranging from 1 to 9, with 1 representing the lowest, and 9 the highest value of each criteria. In particular, the two extremes (1 and 9) correspond to "low" and "high" for interest in video and audio content as well as the perceived video quality, "no immersion" and "full immersion" for level of immersion and "no awareness of my environment" and "full awareness of my environment" for awareness of the surrounding. The dataset, including subjective ratings, as well as recorded physiological signals, is publicly available[3].

3. ANALYSIS AND RESULTS

To ensure that the ratings did not deviate significantly across subjects, detection and elimination of outliers was performed based on the scale of the IL ratings. The outliers detection was applied according to the guidelines described in Section 2.3.1 of Annex 2 of [4]. In this study, no outliers were detected.

[1] http://media.xiph.org

[2] http://www.blender.org/foundation

[3] http://mmspg.epfl.ch/SoPMD

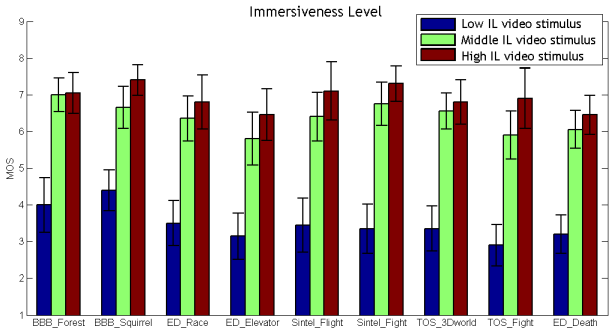

Figure 1: MOSs and CIs for the experienced IL.

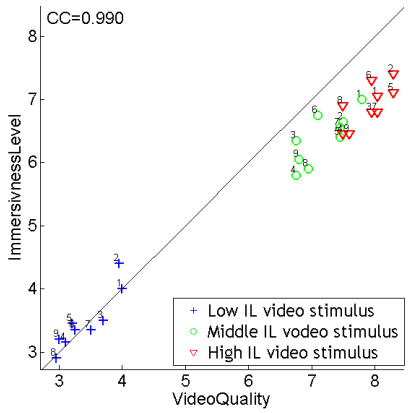

Figure 2: Correlation between the experienced SoP and the assessed quality of the video.

3.1 Subjective ratings analysis and results

The analysis conducted on the subjective ratings includes score distribution histograms, Mean Opinion Scores (MOSs) and associated 95% Confidence Intervals (CIs), as well as Pearson's correlations, assuming a Student's t-distribution of the subjective rates.

Figure 1 shows the resulting MOSs and CIs for the SoP experienced during stimuli for each content. The observed MOSs confirm that all ILs were experienced. In general, a higher IL provokes a better immersive experience as the average MOSs values show for low, middle, and high ILs corresponding to 4, 6.5, and 7, respectively. Moreover, the difference between the middle and high IL is not significant in any content as the CIs considerably overlap for all contents. However, the CIs attest that there is a significant difference between the low IL and the two other levels. These findings indicate that investigation of immersiveness is possible from this database.

To understand the impact of the sequence characteristics (interest in the video and the audio content, the quality, and resolution of the video) and verify that the awareness of surrounding is inversely related to IL, the correlation between the MOSs for all five criteria was measured using Pearson correlation coefficient. Figure 2 depicts the correlation between the MOSs rates given for the video quality and the SoP, and Table 2 presents the overall correlation coefficients. In Figure 2, the stimuli with the same number originate from the same content. This shows that, for each content, the higher the perceived quality of a video stimulus, the more immersive is the experience. The correlation coefficient of IL and video quality is 0.99, meaning that these two criteria are highly correlated. A huge difference is observed between the low and middle/high classes, corroborating with the previous analysis. It also explains the high correlation results. It should be pointed out that each sequence provides a better immersive experience when its IL is increased. The Table 2 confirms the high correlation between IL and the perceived video quality (cc = 0.99), and shows the influence of sound (cc = 0.97). It also validates that the surrounding awareness is inversely related to the IL (cc = -0.99).

	Video Quality	Video Content Interest	Audio Content Interest	Surrounding Awareness
Immersiveness Level	**0.990**	**0.914**	**0.974**	**-0.986**
Video Quality	-	0.892	0.988	-0.987
Video Content Interest	-	-	0.857	-0.903
Audio Content Interest	-	-	-	-0.965

Table 2: Pearson correlation coefficients between the ratings of different perceptual criteria.

3.2 Physiological signal analysis and results

This section presents the pre-processing steps to remove the artifacts from recordings, the feature extraction methods and the classification results.

The pre-processing steps were inspired by [13]. Regarding the ECG signals, the Heart Rate Variability (HRV) was extracted. HRV is the physiological measurement of variation in the time interval between consecutive heart beats, i.e., the variation of R-R intervals, in beats per minute. Since the HRV is a time-series of non-uniform R-R intervals, the HRV was regularly resampled at a rate of 4 Hz. The obtained HRV sequences were used for feature extraction. Features extracted from the HRV signals include the mean and variance, the heart rate, the power of the low-frequency band (0.03 - 0.12 Hz), and the power of the high-frequency band (0.12 - 0.49Hz).

Both respiratory signals (abdomen and thoracic) were filtered using a wavelet multivariate de-noising [2]. Respiration rate and average power across 0.1 to 0.4 Hz frequency band were extracted as respiration features.

EEG signals were filtered by a fourth-order Butterworth filter between 3 and 47 Hz, in order to remove Electrooculogram (EOG) and Electromyogram (EMG) artifacts. Based on the international 10-20 system configuration, 19 channels were selected and processed from a total of 256 channels. Eye-movements and blinking artifacts were removed using Independent Component Analysis (ICA). We used the concept of functional connectivity to explore the EEG signals. Functional connectivity describes the dependence across various sub-regions of the brain [5]. In this study linear granger causality was applied to the pre-processed EEG data to estimate the functional connectivity between each pair of channels.

Previous studies have shown that the brain can be interpreted as a network [3], and that related features [18] can be extracted using granger causality estimated functional connectivity maps [17]. We extracted network features, such as characteristic path length [22], global efficiency [14], clustering coefficient [22], and local efficiency [14], from the estimated functional connectivity maps. Features for each trial are concatenated into one row for classification.

		Predicted IL			
		Low	Middle	High	Total
Actual IL	Low	110	0	70	180
	Middle	70	20	90	180
	High	11	0	169	180
	Total	191	20	329	540

Table 3: Confusion matrix of classification results between ILs. Numbers in the confusion matrix represents the resulting number of trials that are classified into each classes.

A 3-class Support Vector Machine (SVM) with a Gaussian radial basis function kernel was employed to classify between low, middle, and high ILs based on the physiological signals. The feature set was constructed from fusing all the features, i.e., concatenating all EEG and peripheral features in one feature vector. The whole feature set was split into ten folds. Moreover, a radial basis function kernel parameter was selected based on one-fold cross-validation. The confusion matrix was computed to evaluate the performance of the classifier(cf. Table 3). The three classes (low, middle, and high) of IL were equally balanced (i.e., 180 instances each), so the random classification accuracy is 33%.

The confusion matrix shows that it is easy to classify low and high IL with 61% and 94% accuracy, respectively. On the other hand, the middle IL classification accuracy is only 11%. The results are consistent with the subjective analysis, indicating that high and low ILs are easier to be identified, when compared to middle IL.

4. CONCLUSION

This paper introduces a publicly available dataset of immersive multimedia contents including corresponding subjective ratings, as well as recorded physiological signals. More specifically, the dataset comprises EEG, ECG, and respiration signals, as well as subjective ratings with respect to video quality, interest in video and audio content, and Immersiveness Level (IL).

A preliminary analysis based on the subjective ratings and the physiological signals was performed. The subjective ratings analysis demonstrated that various IL were experienced. A clear distinction between low and high IL was observed, whereas the differences between middle and high IL were not significant. The results also showed a high correlation between the Sense of Presence (SoP) and the quality of the stimuli. A classifier based on EEG and peripheral features enables the clear distinction between low and high IL, which is in line with subjective ratings analysis. This leads to the conclusion that this SoP dataset is consistent and can be valuable for further analysis.

5. ACKNOWLEDGMENTS

This work has been performed in the framework of the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No. 643072 – Network QoE-Net, SNF project No. 200020-149259 – LEADME, and EUROSTARS project No. E!8307 – TOFuTV.

6. REFERENCES

[1] M. K. Abadi, J. Staiano, A. Cappelletti, M. Zancanaro, and N. Sebe. Multimodal engagement classification for affective cinema. In *Affective Computing and Intelligent Interaction (ACII), 2013*, pages 411–416. IEEE, 2013.

[2] M. Aminghafari, N. Cheze, and J.-M. Poggi. Multivariate denoising using wavelets and principal component analysis. *Computational Statistics & Data Analysis*, 50(9):2381 – 2398, 2006.

[3] D. S. Bassett and E. Bullmore. Small-world brain networks. *The neuroscientist*, 12(6):512–523, 2006.

[4] I.-R. BT.500-13. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Unio*, January 2012.

[5] A. A. Fingelkurts, A. A. Fingelkurts, and S. Kähkönen. Functional connectivity in the brain - is it an elusive concept? *Neuroscience & Biobehavioral Reviews*, 28(8):827–836, 2005.

[6] J. P. Forgas. On feeling good and being rude: Affective influences on language use and request formulations. *Journal of Personality and Social Psychology*, 76(6):928, 1999.

[7] J. Freeman, S. Avons, D. Pearson, and W. IJsselsteijn. Effects of sensory information and prior experience on direct subjective ratings of presence. *Presence*, 8(1):1–13, Feb 1999.

[8] X. Geng. Cultural differences influence on language. *Review of European Studies*, 2(2):p219, 2010.

[9] P. ITU-T RECOMMENDATION. Subjective video quality assessment methods for multimedia applications. *International Telecommunication Unio*, April 2008.

[10] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis ;using physiological signals. *Affective Computing, IEEE Transactions on*, 3(1):18–31, Jan 2012.

[11] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi. EEG correlates during video quality perception. In *Signal Processing Conference (EUSIPCO), 2014 Proceedings of the 22nd European*, pages 2135–2139, Sept 2014.

[12] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi. Predicting subjective sensation of reality during multimedia consumption based on EEG and peripheral physiological signals. In *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pages 1–6, July 2014.

[13] E. Kroupi, P. Hanhart, J.-S. Lee, M. Rerabek, and T. Ebrahimi. User-independent classification of 2D versus 3D multimedia experiences through EEG and physiological signals. *8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics-VPQM. 2014. No. EPFL-CONF-197071.*, 2014.

[14] V. Latora and M. Marchiori. Efficient behavior of small-world networks. *Physical review letters*, 87(19):198701, 2001.

[15] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff. A cross-media presence questionnaire: The ITC-sense of presence inventory. *Presence*, 10(3):282–297, June 2001.

[16] J. Li, Y. Koudota, M. Barkowsky, H. Primon, and P. Le Callet. Comparing upscaling algorithms from hd to ultra hd by evaluating preference of experience. In *The International Workshop on Quality of Multimedia Experience (QoMEX) 2014*, Singapore, Singapore, Sep 2014.

[17] W. Liao, J. Ding, D. Marinazzo, Q. Xu, Z. Wang, C. Yuan, Z. Zhang, G. Lu, and H. Chen. Small-world directed networks in the human brain: multivariate granger causality analysis of resting-state fmri. *Neuroimage*, 54(4):2683–2694, 2011.

[18] M. Rubinov and O. Sporns. Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, 52(3):1059–1069, 2010.

[19] M. Slater and S. Wilbur. A framework for immersive virtual environments (five), speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6(6):603–616, Dec. 1997.

[20] M. Soleymani, M. Pantic, and T. Pun. Multimodal emotion recognition in response to videos. *Affective Computing, IEEE Transactions on*, 3(2):211–223, April 2012.

[21] A. M. von der Pütten, J. Klatt, S. T. Broeke, R. McCall, N. C. Krämer, R. Wetzel, L. Blum, L. Oppermann, and J. Klatt. Subjective and behavioral presence measurement and interactivity in the collaborative augmented reality game timewarp. *Interacting with Computers*, 24(4):317 – 325, 2012. Special Issue on Presence and Interaction.

[22] D. J. Watts and S. H. Strogatz. Collective dynamics of "small-world" networks. *nature*, 393(6684):440–442, 1998.

[23] B. G. Witmer and M. J. Singer. Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoper. Virtual Environ.*, 7(3):225–240, June 1998.