Multimodal Dataset on the Sense of Presence using physiological signals

Anne-Flore Perrin, He Xu, Eleni Kroupi, Martin Rerabek, Touradj Ebrahimi
Multimedia Signal Processing Group (MMSPG)
Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland
{name.surname}@epfl.ch

ABSTRACT

Measuring the perceived sense of immersion of a set of videos with respect to content, quality, resolution and sound system, is not an easy task because of the subjectivity of the human perception. A way to express the overall perceived quality of a video is to assess the Quality of experience (QoE), especially the Sense of Presence (SoP). The latter, not so explored, is the principle we would like to assess. To be independent of the human way to explicitly express feelings and experiences, the analysis is carried out using physiological signals such as electroencephalography (EEG), electrocardiography (ECG) and respiration.
The aim pursued is the creation of a data set containing all the above-mentioned signals in order to study the SoP. The analysis demonstrates that data set is well constructed and functional for the assessment of the SoP.

1. INTRODUCTION

The Sense of Presence (SoP) is the main quality metric for virtual environment as attested in [15, 3]. According to [18], the SoP, also called immersiveness level in this paper, refers to the subjective experience of having left the real world and being now "present" in a virtual environment. Its definition is explicated in [19, 12] which also explains how to measure it.

The human subjective and explicit assessment is, for instance, emotional, cultural and educational-dependent [4, 5, 6, 7]. Thus a valuable survey of the SoP can not only rely on subjective ratings. Based on [10, 11, 16], subjects'physiological signals such as brain activity (electroencephalography (EEG)), heart activity (electrocardiography (ECG)) and respiration are objective data adequate to assess the SoP complementarily to the subjective rates.

Multimedia content provided for TV can be viewed as small virtual environment. Indeed, sensory cues for virtual environment usually consists primarily of visual stimuli, often but not always accompanied with audio stimuli. Digital television technologies aim to provide higher quality multimedia experiences, possibly with various Quality of service (QoS). Therefore the SoP can be investigated to understand and to further analyze the Quality of experiences (QoEs).

This paper presents a novel database that captures the differences in user-experience during multimedia stimuli with various immersiveness levels. EEG and peripheral physiological signals including ECG and respiration, as well as subjective ratings are required during the experiment.

An investigation of the experience transcribed in The subjective ratings analysis shows that some QoS properties are correlated with the SoP. Finally the constructed subject-independent classification system distinguishes the various immersiveness levels based on EEG and/or peripheral physiological signals.

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 subjective rating analysis and user-independent physiological classification. Finally, conclusion is given in Section 4.

2. DATA COLLECTION

2.1 Participants

Eight females and twelve males participated in the study. They were from 18 to 30 years old (23 average and standard deviation years of age). The 20 subjects were screened for correct visual acuity (no errors on 20/30 lines) and color vision using Snellen and Ishiara charts respectively [8, 17]. They all provided written consents forms. Before each experiment, oral instructions were provided to the participants to explain their tasks. Additionally, a training session was organized to allow participants to familiarize with the assessment procedure. The content shown in the training session was selected by expert viewers in order to include examples of all evaluated aspects.

2.2 Audio-visual stimuli

Video stimuli are derived from nine video sequences extracted from four open source movies published by the Blender Foundation (Big buck bunny, Elephant dream, Sintel and Tears of Steel)[1]. A supplementary sequence content was chosen for the training session.
The one-minute selected video contents have the highest audio, spatial end temporal energy and are related to the scene

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

ACM MM 2015 , Brisbane, Australia

cuts. The twenty seven video stimuli shown during an experiment are the combination of the nine video contents with the three levels of immersiveness described below.
Low, middle and high immersiveness levels were defined based on the audio sound system, the video quality\level of compression and the resolution. The table 1 illustrates the settings of QoS creating three level of immersiveness.

	Immersiveness levels		
	Low	Middle	High
Audio	No Audio	Stereo	Surround
Quality (QP)	36	20	20
Resolution	SD	HD	UHD

Table 1: Immersiveness levels settings

[TBD] The video stimuli order could impact the data and so the results. Thus the sessions are built to allow this study : The first session will display the video stimuli from the lowest immersiveness level stimuli to the ones with the highest level. The second session order is middle immersiveness level stimuli, followed by the low immersiveness level stimuli and then the high immersiveness level stimuli. The last session will display the video stimuli from the video with the highest immersiveness level to ones with the lowest. Thus the study of the change of devices is allowed and besides these constraints, the video order is different for each volunteer, thanks to a pseudo-random function.

2.3 Monitor, sound system and environment

Professional high-performance 4K/QFHD LCD reference 56-inch monitor Sony Trimaster SRM-L560[2] was used to display video stimuli. As recommended in [13], the viewing distance was set at 1.6H (H - Height of the screen). The Altec Lansing 5.1 THX speaker system, super subwoofer was used as audio sound system. The laboratory setup has been thought to provide a quiet environment and fix the ambient light to ensure subjects comfort during bright and dark scene as well as during the resting periods.

2.4 Physiological signal acquisition

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 digitized the EEG signals while the participants were watching the stimuli. The heart activity is recorded from two standard ECG electrode placed on the lower left rib cage and the upper right clavicle. Two respiratory inductive plethysmography belts (thoracic and abdomen)are used to acquire the respiration. All signals were recorded at 250 Hz.

2.5 Experimental protocol

The experiments consist of three sessions intersected by ten-minutes breaks in order to avoid subject fatigue and lack of attention. Nine video stimuli (coming from the nine sequences) were presented in each session leading to a total of 27 video stimuli, and thus, to a total of 27 trials. The order of the SoP levels of the video stimuli is set as follows. For the first session, the three first stimuli have a low immersiveness level (IL), the three next a middle IL and the three last ones a high IL. Respectively for the second and third session the order of the IL is middle, low, high and high, middle and low. 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 pseudo-randomly selected and presented. After the video sequence was over, the subjects were asked to provide their self-assessed ratings for the particular video sequence without any restriction in time, following the Absolute Category Rating (ACR) evaluation methodology [9].

Regarding the self-assessed ratings, subjects were asked to evaluate the video sequences in terms of five different aspects, namely interest in the video content, perceived video quality, interest in audio content, immersiveness level and surrounding awareness. A 9-point rating scale was used that ranged from 1 to 9, with 1 representing the lowest value, and 9 the highest value of each aspect. 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 the immersion level and "no conscience of my environment" and "full conscience of my environment" for the surrounding awareness.

Once a trial was over, the next baseline period was recorded and the next video sequence was pseudo-randomly selected and 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.

An illustration of a session is presented figure 1.

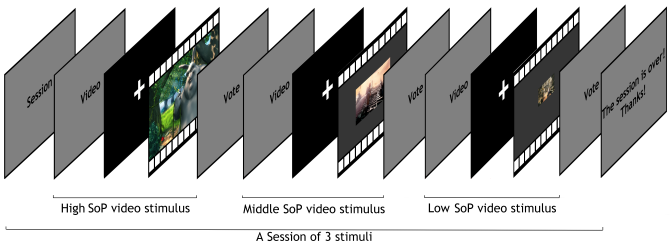

Figure 1: Example of a 3 video stimuli session progress

3. ANALYSIS

The training session was used to illustrate the low, middle and high SoP levels in order to guide subjects to bound their own perceived overall ratings more or less similarly. To ensure that the ratings do not deviate significantly across subjects, the detection and elimination of outliers was performed. The assessed IL is the factor in which subjects were trained, thus the outliers detection was based on the scale of the perceived overall quality ratings. The outliers detection was applied according to the guidelines described in Section 2.3.1 of Annex 2 of [2]. In this study, no outliers were detected.

[2] *http://pro.sony.com/bbsccms/assets/files/cat/mondisp/brochures/di0195_srm1560.pdf*

3.1 Subjective ratings analysis

The analysis conducted on the subjective rates includes score distribution histograms, box plots, Mean opinion scores (MOS) and associated 95% Confidence intervals (CI) and Pearson's correlations, assuming a Student's t-distribution of the subjective rates.

The first verification is to ensure that all the immersiveness levels (ILs) were experienced by the subjects. The score distribution histogram of the ratings given by all the subjects for all the trials is presented in Figure 2. The value 1 corresponds to the lowest and 9 to the highest rate.

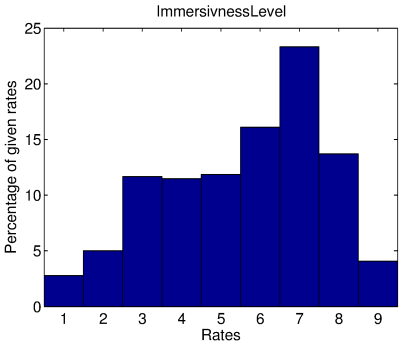

Figure 2: Score distribution histogram for the IL experienced

As it can be observed, all the IL were experienced during the experiments. More specifically, the distribution of the rates is roughly 20% for the three lowest IL, 40% for the three middle and highest IL. Thus the distribution almost describes three classes of immersiveness.

Figure 3 shows the resulting MOS and CI for the SoP experienced during stimuli.

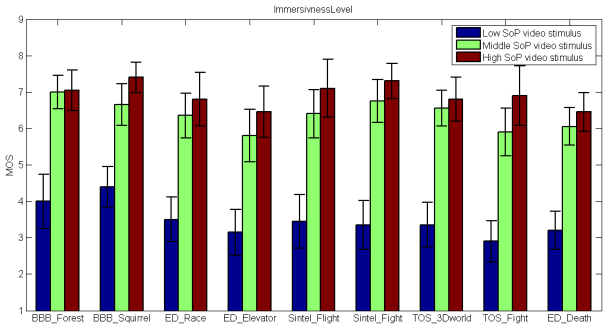

Figure 3: MOS and CI for the IL experienced

The observed MOS results valid the three IL chosen during the experiment design. In fact, the low IL is assessed around the rate 4, the middle IL at about 6.5 and the high IL at roughly 7. The high IL is always better perceived than the middle one. This latter is also always better perceived than the low IL. However, the difference between the middle and high levels is not significant as the CI considerably overlap for all contents. Nevertheless the CI attests that there is a high difference between the low IL and the two other levels. Thus a study of immersive or non-immersive QoE is possible from this database.

To understand the impact of QoS factors - such as the interest of the video and audio content, the quality and resolution of the video - and verify that the surrounding awareness is inversely related with the IL, the correlation between the MOS for all five factors was measure using Pearson correlation coefficient. Figure 4 and Table 2 illustrate and report respectively how highly the IL is correlated with the video quality and the overall correlation coefficients.

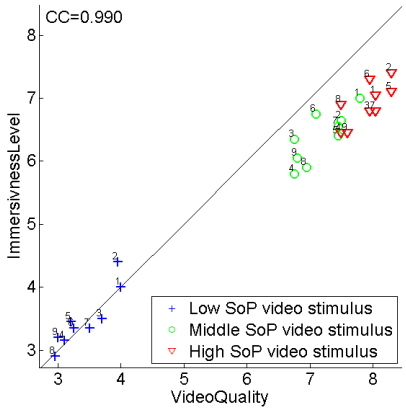

Figure 4: Correlation between the experienced SoP and the assessed quality of the video

Figure 4 depicts the correlation between the MOS rates given for the video quality and the SoP. The stimuli with the same number are coming from the same sequence. This permits the study of the evolution of sequences ratings depending on the IL. The correlation coefficient of IL and video quality is 0.99, meaning that these two characteristics are highly correlated. A huge difference is observed between the low class and middle/high class, corroborating with the previous analysis. It should be pointed out that each sequence provides a better immersive experience when its IL is increased.

	video quality	Interest in video content	Interest in audio content	Surrounding awareness
Immersiveness level	0.990	0.914	0.974	-0.986
video quality	-	0.892	0.988	-0.987
Interest in video content	-	-	0.857	-0.903
Interest in audio content	-	-	-	-0.965

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

The table 2 confirms the high correlation between IL and the video quality (cc = 0.99), and shows the influence to having sound or not (cc = 0.97). It also valid that the surrounding awareness is inversely related with the IL (cc = -0.99).

3.2 Physiological signal analysis

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

3.2.1 Pre-processing

The manualy rejection of muscle activity related EEG electrodes leads to a total of 216 electrodes for processing

and analysis. EEG signals were filtered between 3-47 Hz using a third-order Butterworth filter, in order to remove electrooculogram (EOG)and electromyogram (EMG) artifacts. [TO SET : DENOISING FUNCTION]

ECG signals were used to extract the heart rate variability (HRV), which reflects the sympathetic/parasympathetic modulation. HRV is the physiological measurement of variation in the time interval between consecutive hearts beats. In order to extract the HRV, the interval between two QRS complexes defined as R-R interval (t_{R-R})was estimated using the real-time algorithm developed by Pan and Tompkins [14]. Then the heart rate (HR, in beats per minute) was estimated as :

$$HR = \frac{60}{t_{R-R}} \quad (1)$$

The HRV is the variation of HR over time. As the HR is a time-series of non-uniform R-R intervals, the HR was regularly resampled at 4 Hz rate. [TO SET : respiration denoising]

Both respiratory signals (abdomen and thoracic) were filtered by a wavelet multivariate de-noising [1]. It combines univariate wavelet de-noising in the basis where the estimated noise covariance matrix is diagonal and non-centered Principal Component Analysis (PCA) on approximations in the wavelet domain.

In the presented results, only 19 EEG signals were kept to expedite the validation of the database.

3.2.2 *Feature extraction*

[He?]

3.2.3 *Classification*

[He?]

3.2.4 *Results*

[Me]

4. CONCLUSION

An SoP multimodal database will be made available to researchers allowing the study of the IL of a content as well as the impact of the device changes during a content visualization. The ultimate objective is to provide a reliable database leading to the supply of an higher quality multimedia experiences independently of the QoS.

The SoP is assessed thanks to the brain and heart activity as well as the respiration (body response such as sympathetic and parasympathetic activity) and subjects subjective ratings (possibly skewed by subjects related factors such as culture, education etc.)

The analysis of the database demonstrates that all the IL were experienced with a clear distinction between low and high immersive experiences during multimedia contents visualization. A shallow processing of the database (19 EEG electrode only) was conducted. This latter leads to the conclusion [CONCLUSION CLASSIFIERS].

5. REFERENCES

[1] 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. Statistical signal extraction and filtering Statistical signal extraction and filtering.

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

[3] H. Q. Dinh, N. Walker, C. Song, A. Kobayashi, and L. F. Hodges. Evaluating the importance of multi-sensory input on memory and the sense of presence in virtual environments. In *Proceedings of the IEEE Virtual Reality*, VR '99, pages 222–, Washington, DC, USA, 1999. IEEE Computer Society.

[4] 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.

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

[6] J. J. Gross and O. P. John. Individual Differences in Two Emotion Regulation Processes: Implications for Affect, Relationships, and Well-Being. *Journal of Personality and Social Psychology*, 85:348–362, 2003.

[7] A. R. Hochschild. Emotion work, feeling rules, and social structure. *American Journal of Sociology*, 85(3):pp. 551–575, 1979.

[8] S. Ishihara. Test for coiour-blindness. *Tokyo: Hongo Harukicho*, 1917.

[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. 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.

[12] 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.

[13] 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.

[14] J. Pan and W. J. Tompkins. A real-time qrs detection algorithm. *Biomedical Engineering, IEEE Transactions on*, BME-32(3):230–236, March 1985.

[15] 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.

[16] 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.

[17] S. Songden and E. Ike. Colour vision performance test. *Journal of Natural Sciences Research*, 3(11):19–24, 2013.

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

[19] 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.