USER-INDEPENDENT CLASSIFICATION OF 2D VERSUS 3D MULTIMEDIA EXPERIENCES THROUGH EEG AND PHYSIOLOGICAL SIGNALS

Eleni Kroupi[1]*, Philippe Hanhart*[1]*, Jong-Seok Lee*[2*]*, Martin Rerabek*[1]*, and Touradj Ebrahimi*[1]

[1]Multimedia Signal Processing Group (MMSPG)
École Polytechnique Fédérale de Lausanne (EPFL)
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland

[2]School of Integrated Technology
Yonsei University
162-1 Songdo-dong, Yeonsu-gu, 406-840 Incheon, Republic of Korea

ABSTRACT

3D image and video have become popular since they can enhance the Quality of Experience (QoE) by adding the depth dimension to the traditional 2D media. In order to design and optimize the human-centric 3D multimedia processing techniques, it is important to understand the QoE perceived by users for given contents, as well as ways 2D and 3D media affect them. The goal of this paper is twofold. First, to investigate perceived QoE of 2D and 3D multimedia contents based on subjective ratings, and second, to investigate the way in which 2D and 3D multimedia technologies affect physiological signals. In particular, we construct subject-independent classification systems to predict 2D versus 3D multimedia experiences based on electroencephalography (EEG) and peripheral physiological signals including heart rate and respiration, from which the potential of EEG for this purpose is demonstrated.

1. INTRODUCTION

With the advances of the 3D imaging technologies, 3D video contents have gained in popularity. Addition of depth cues in 3D contents is supposed to enhance users' multimedia experience when compared to 2D. However, low 3D quality may not only reduce the added value of the depth dimension, but also cause adverse problems such as visual fatigue or dizziness. Therefore, measuring and maximizing users' quality of experience (QoE) of 3D contents have been considered critical in designing successful 3D applications.

Traditionally, users' perceived quality is measured via subjective quality assessment, where perceived quality of selected visual stimuli is obtained from a number of subjects. The subjects have to explicitly rate the quality of each stimulus in a pre-defined rating scale. In contrast, efforts toward understanding perceived quality without users' explicit response have been made recently. For example, it was shown that 3D visual fatigue can be identified to some extent by observing the electroencephalogram (EEG) response [1] or cortical activities measured by fMRI [2]. Such implicit, direct monitoring of subjects' brain responses is an interesting alternative to explicit rating, and also can exclude subjective bias occurring during the rating activity.

While EEG monitoring techniques have been mainly used for assessment of 3D visual fatigue, this paper explores the feasibility of identifying difference in users' experience for 2D and 3D contents through EEG and peripheral physiological signals including electrocardiogram (ECG) and respiration. We conduct subjective experiments, in which 2D and 3D multimedia contents are presented to users, and both explicit subjective ratings and implicit EEG and physiological responses are captured. We first investigate influences of different aspects of 2D and 3D media, such as the rendering mode, image quality, and content, on depth perception by analyzing the subjective ratings. Then, we construct subject-independent classification systems that can distinguish the type of media (i.e., 2D vs. 3D) based on EEG or peripheral physiological signals. The performance of these modalities is then presented and compared.

The rest of the 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 signal classification. Finally, conclusion is given in Section 4.

*Refers to the corresponding author.

The research leading to these results has been performed in the framework of Swiss National Foundation for Scientific Research (FN 200020-132673-1, FN 200020-149259-1, FN IZK0Z2-150904, and FN 200020-143696-1) and the Basic Science Research Program through the National Research Foundation of Korea funded by the Ministry of Science, ICT and Future Planning, Korea (2013R1A1A1007822).

2. DATA COLLECTION

2.1. Participants

Sixteen naive subjects (5 females, 11 males) took part in our experiments. They were between 19 and 30 years old with an average of 23.8 years of age. All subjects were screened for correct visual acuity (no errors on 20/30 line), color vision and stereo vision using Snellen, Ishiara and Randot 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 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

The dataset was composed of eight video contents: one for the training and seven for the tests. All contents were shot during the 2012 edition of the Montreux Jazz Festival (MJF) (protected by copyright), with two RED SCARLET-X mounted on a Genus Hurricane Rig. All video sequences were recorded in REDCODE RAW (R3D) format, DCI 4K resolution (4096×2160 pixels), at 25 fps, and had a duration of about one minute. Stereo audio was recorded in PCM format, sampled at 48 kHz, 24 bits. Table 1 describes the contents and their characteristics. The recorded video sequences were cropped and downsampled to Full HD resolution (1920×1080 pixels) and then compressed with H.264/MPEG-4 AVC. Two different quantization parameters (QP) were selected: QP=2 for high quality (HQ) and QP=35 for low quality (LQ). For each content, four different versions were considered: 2D HQ, 3D HQ, 2D LQ, and 3D LQ, leading to a total of 28 video sequences, 14 of which in 2D and 14 in 3D.

2.3. Monitor, sound system and environment

To display the video stimuli, a HD 46" Hyundai S465D polarized stereoscopic monitor was used. The laboratory setup was controlled in order to ensure the reproducibility of results by avoiding involuntary influence of external factors. The test room was equipped with a controlled lighting system with a 6500K color temperature and an ambient luminance at 15% of the maximum screen luminance. For the audio playback, the PSI A14-M professional studio full range speakers were used.

2.4. Physiological signal acquisition

The EEG signals were recorded from 256 electrodes placed at the standard positions on the scalp. An EGI's Geodesic EEG System (GES) 300 was used to record, amplify, and digitalize the EEG signals while the participants were watching the stimuli. Additionally, two standard ECG leads were placed on the lower left ribcage and on the upper right clavicle, as well as two respiratory inductive plethysmography belts (thoracic and abdomen). All signals were recorded at 250 Hz.

2.5. Experimental protocol

The participants were seated at a distance of 3.2 times the picture height, corresponding to roughly 1.8 meters from the stereoscopic monitor, as suggested in [3]. All video sequences were viewed with 3D glasses. The experiments were conducted in three sessions. A fifteen-minute break was given between two sessions in order to avoid subject fatigue and lack of attention. Nine video sequences were presented in the first and second sessions, and ten in the last one, leading to a total of 28 video sequences, and thus, to a total of 28 trials.

Each trial consisted of a ten-second baseline period and a stimulus period. The biosignals recorded during the baseline period were used to remove stimulus-unrelated variations from the signals obtained during the stimulus period. During the baseline period, 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 sequence was 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 [4].

Regarding the self-assessed ratings, subjects were asked to evaluate the video sequences in terms of four different aspects, namely perceived overall quality, content preference, sensation of reality, and perceived depth quantity. 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 perceived overall quality and content preference, "no presence" and "very strong presence" for sensation of reality", and "no depth" and "a lot of depth" for perceived depth quantity.

Once a trial was over, the next baseline period was recorded and the next video sequence was randomly selected and presented. The procedure was repeated until all 28 video sequences were presented and rated, leading to 28 trials. Although the experiments lasted for almost two hours, including the training and set up, the subjects did not report fatigue.

Table 1. Characteristics of the contents used in our experiments.

Content	Description and characteristics
Training	Rock band playing at the Auditorium Stravinski. Dark. Bright spots. Shot from the back of the auditorium.
Jazz	Jazz band playing at the Funky Claude's Lounge at the Opening Party. Wide shot.
Rock	Rock band playing at the Auditorium Stravinski. Dark. Bright spots. Shot from the back of the auditorium.
Stage	MJF general manager on stage introducing the next artist. Very dark. In French. Wide shot.
Speech1	MJF general manager giving a speech at the Opening Party. In French. Mid shot.
Speech2	Speech at the Opening Party. In French. Mid shot.
Outdoor	Crowd walking on the street near the lake. Lot of depth. Wide shot.
Interview	Interview of Quincy Jones. Medium close up.

3. ANALYSIS

To detect and remove subjects whose ratings appear to deviate significantly from others, outlier detection was performed. During the training session, examples of the lowest and highest quality levels (LQ and HQ) were shown, in order to guide subjects to bound their own perceived overall quality ratings more or less similarly. Since quality was the only factor in which subjects could be trained, the outlier detection was performed only on the scale of the perceived overall quality ratings. The outlier detection was applied according to the guidelines described in Section 2.3.1 of Annex 2 of [5]. In this study, no outliers were detected.

3.1. Subjective rating analysis

The mean opinion score (MOS) and associated 95% confidence interval (CI) were computed for each test stimulus, assuming a Student's *t*-distribution of the subjective ratings, to represent explicit estimates of perceived depth quantity. Figure 1 shows the resulting MOS and CI for perceived depth quantity. As it can be observed, for a given quality level, perceived depth is higher for 3D stimuli when compared to 2D stimuli. Similarly, high quality sequences generally obtained higher ratings for perceived depth quantity when compared to their corresponding low quality versions. However, the difference in terms of perceived depth between 3D LQ stimuli and 2D HQ stimuli is not significant as the CIs considerably overlap in all contents. This observation shows that monocular depth cues are powerful in high quality video sequences. As content *Stage* is very dark, the perceived 3D effect was not very strong and the perceived depth quantity was rated relatively low.

To investigate quantitatively whether the objective factors, such as the rendering mode, actual quality level, and content have a significant influence on perceived depth, an ANOVA analysis was performed on the subjective ratings. In particular, the null hypothesis was that the rendering mode, quality level, and content do not influence perceived depth quantity. The null hypothesis was rejected for all cases, $p < 0.001$, indicating that the effects of the rendering mode, actual quality level, and content on perceived depth quantity were significant. However, the interactions among these three different factors were not significant, $p > 0.3$.

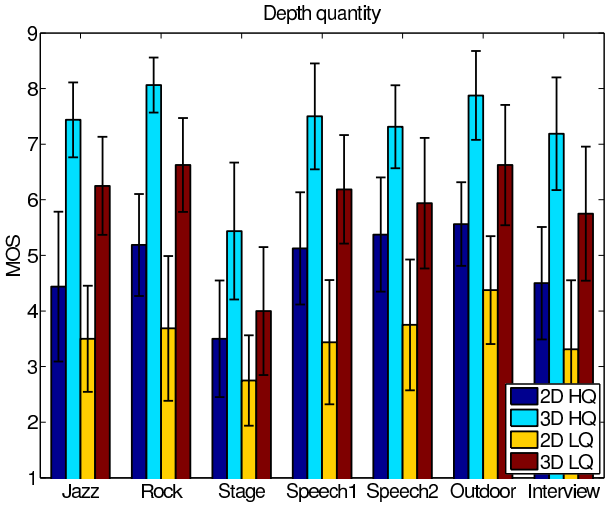

Fig. 1. Mean opinion scores: depth quantity.

Also, to understand the impact of the perceptual factors, such as sensation of reality, content preference, and perceived overall quality, on perceived depth quantity, the correlation between the MOS for all four factors was measured using the Pearson correlation coefficient. Table 2 reports the correlation coefficients. The results show that there is a strong correlation between perceived depth quantity and sensation of reality ($\rho > 0.88$). On the other hand, the correlation between perceived depth quantity and content preference ($\rho < 0.16$) is very weak. Thus, apparently content *per se* impacts on depth perception, but content preference does not. The correlation between perceived depth quantity and perceived overall quality is rather low ($\rho < 0.42$), but significantly different from zero, $p = 0.027$. Thus both perceived and actual quality impact on perceived depth, according to both the ANOVA and the correlation analyses.

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

	Content preference	Sensation of reality	Depth quantity
Overall quality	0.3392	0.7308	0.4172
Content preference	-	0.3017	0.1527
Sensation of reality	-	-	0.8835

3.2. Physiological signal analysis

In this section, the pre-processing steps to remove the artifacts, the feature extraction methods and the classification results are presented.

3.2.1. Preprocessing

EEG electrodes in which muscle activity was discernible were rejected manually, leading to a total 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. Remaining artifacts were removed by cubic interpolation. All signals were visually inspected to make sure they did not contain further artifacts. EEG signals were initially referenced to the Cz electrode and re-referenced to the common average.

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 heart 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 [6]. 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 nonuniform R-R intervals, the HR was regularly resampled at 4 Hz rate. Then, the respiration drift was removed using a morphological operator.

Both respiratory signals (abdomen and thoracic) were filtered by a second-order bandpass filter with cutoff frequencies in a range of 0.1-1 Hz, approximated by a Butterworth polynomial.

Only the last 45 seconds of all signals were used in our analysis, considering that stabilization and adaptation to 3D contents may take some time.

3.2.2. Feature extraction

Regarding the EEG signals, the frequency power of the signals was extracted for frequencies between 4 and 47 Hz, using the Welch's method with windows of 128 samples. The mean trial power was then divided by the mean baseline power, in order to extract the power changes without considering the pre-stimulus period. These power changes were captured for different frequency bands, namely theta band (4-7 Hz), alpha band (8-13 Hz), beta band (14-29 Hz) and gamma band (30-47 Hz). Apart from the power division, a novel distance metric between trial and baseline power is also used as a feature, namely the Wasserstein distance [7]. This metric is less sensitive to the location of the frequency peaks, but provides instead, information about the global structure of the frequency domain.

Physiological signals of living organisms, such as EEG signals, appear to vary over time in a complex manner. These temporal variations result from intrinsic disturbances and actions, such as the activity of an organism. In physiological signals these fluctuations are non-periodic. In the current study, the normalized length density (NLD) [8] index is extracted in order to capture the self-similarities of the EEG, HRV and respiration signals during reality perception processes.

Regarding the peripheral signals, mean, standard deviation, and mean absolute values of the first and second derivatives were extracted for both HRV and respiration signals [9]. Regarding the frequency domain, the power of the very low frequency (VLF, 0.0033-0.04 Hz), the low frequency (LF, 0.04-0.15 Hz), high frequency (HF, 0.15-0.4 Hz), and the LF/HF ratio were extracted from HRV [10]. Finally, the power of three different bands was extracted from the respiration signals (0.1-0.2 Hz, 0.2-0.3 Hz, 0.3-0.4 Hz).

3.2.3. Classification

A support vector machine (SVM) classifier with radial basis function (RBF) kernel was trained, and was used to predict 2D and 3D multimedia experiences, in a subject-independent classification scheme. The RBF kernel was used due to the fact that this kernel considers the possible non-linear relationships between class labels and features. The LIBSVM package was used for this study [11]. The parameters of the classifier were optimized for each trial, using a grid search in a leave-one-subject-out cross-validation scheme [11]. As a feature selection method, Fisher's linear discriminant, which is defined as

$$J(f) = \frac{|\mu_1 - \mu_2|}{\sigma_1^2 + \sigma_2^2}, \quad (2)$$

was used to estimate the most significant features. In eq. (2), μ and σ are the mean and standard deviation for each

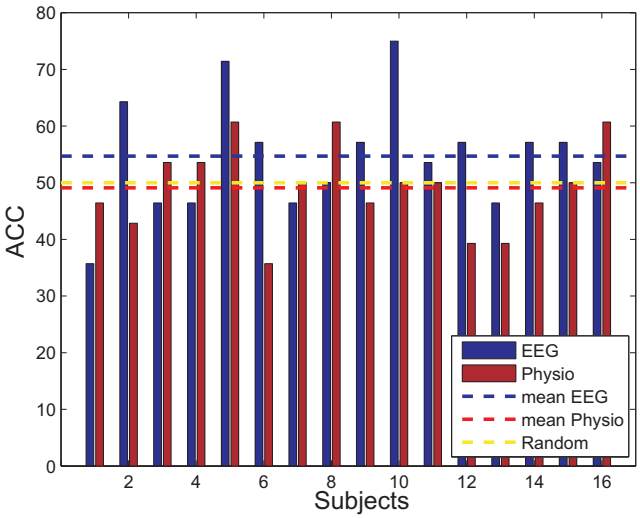

Fig. 2. ACC for EEG and peripheral physiological signals for each subject.

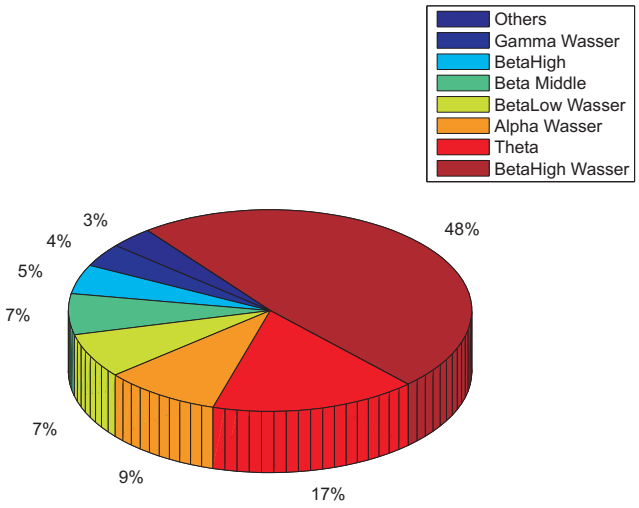

Fig. 3. Best EEG features for all subjects.

feature f, respectively. The training and testing of the classifier were carried out in a leave-one-subject-out cross- validation scheme. To evaluate the performance of the classifiers the classification accuracy (ACC) was computed, due to completely balanced classes. ACC is considered as an accurate metric to assess the performance of a classifier with balanced classes.

3.2.4. Results

The classification was performed on two balanced classes, with 14 samples each, namely 2D and 3D. Figure 2 presents the ACC both for the EEG and for the peripheral signals. Regarding the EEG signals, the mean ACC is significantly higher than random (mean $ACC = 54.69$, $p < 0.05$, a t-test was applied to test the significance), indicating that EEG-based classification is possible and can be used to automatically recognize 2D from 3D perception, in a subject-independent classification framework, independently of the video quality. In fact, a possible reason why the accuracy of the classifier is low may be due to the difference in quality levels among the sequences. Indeed, the subjective-score analysis showed that the difference in terms of perceived depth between 3D low quality and 2D high quality is not significant (Section 3.1). Regarding the peripheral signals, classifying automatically 2D versus 3D is not significantly different from random guess. These results were expected due to the fact that 2D versus 3D sequences are expected to influence the brain in a subject-independent way and create patterns, but are not expected to cause subject-independent patterns and changes to respiration or heart rate signals. Heart rate and respiration may convey rendering-related information but in a more subjective way.

Figure 3 presents the most significant EEG features according to the Fisher's linear discriminant (eq. (2)). The most significant features were estimated as follows; in the leave-one-subject-out cross validation scheme, each time the best feature set was selected based on the training subjects. The overall most significant features were obtained once all subjects were a test subject, and were based on how frequently each feature belonged to the best feature set. The results show that the Wasserstein distance between the high beta power of a trial and a neutral baseline is the most significant feature (almost 50% of the times belonged to the best feature set).

It has been established that beta band is highly associated with cognition, and reflects emotional characteristics [12]. Hence, cognitive and emotional processes seem to take place and be responsible for the automatic discrimination between 3D and 2D video.

4. CONCLUSION

In this paper we presented the results of experiments in which 2D and 3D sequences are presented to users, while their subjective ratings on various aspects are captured, and their EEG, ECG and respiration signals are recorded. Regarding the analysis on the subjective ratings, we showed that for a given quality level the perceived depth is higher for 3D that for 2D sequences, but for high 2D quality and low 3D quality the perceived depth is not significantly different. We also showed that the actual and perceived quality level, rendering mode, and content influence significantly the perceived depth. Regarding 2D versus 3D automatic classification from EEG and peripheral signals, EEG-based

classification is possible and can be used to automatically recognize 2D from 3D perception, in a subject-independent classification framework, independently of the video quality. However, classifying automatically 2D versus 3D using peripheral signals is not significantly different from random.

5. REFERENCES

[1] H.-C. O. Li, J. Seo, K. Kham, and S. Lee, "Measurement of 3D visual fatigue using event-related potential (ERP): 3D oddball paradigm," in *Proc. 3DTV Conf.*, Istanbul, Turkey, 2008, pp. 213–216.

[2] D. Kim, Y. J. Jung, E. Kim, Y. M. Ro, and H. W. Park, "Human brain response to visual fatigue caused by stereoscopic depth perception," in *Proc. Int. Conf. Digital Signal Processing*, Corfu, Greece, 2011, pp. 1–5.

[3] ITU-R BT.2021, "Subjective methods for the assessment of stereoscopic 3DTV systems," International Telecommunication Union, August 2012.

[4] ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, April 2008.

[5] ITU-R BT.500-13, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, January 2012.

[6] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Trans. Biomedical Engineering*, vol. 32, no. 3, pp. 230–236, 1985.

[7] L. Oudre, J. Jakubowicz, P. Bianchi, and C. Simon, "Classification of periodic activities using the Wasserstein distance," *IEEE Trans. Biomedical Engineering*, vol. 59, no. 6, pp. 1610–1619, 2012.

[8] A. Kalauzi, T. Bojic, and L. Rakic, "Extracting complexity waveforms from one-dimensional signals," *Nonlinear Biomedical Physics*, vol. 3, no. 8, 2009.

[9] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[10] K. C Bilchick and R. D Berger, "Heart rate variability," *Journal of Cardiovascular Electrophysiology*, vol. 17, no. 6, pp. 691–694, 2006.

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Trans. Intelligent Systems and Technology*, vol. 2, no. 3, pp. 27:1–27, 2011.

[12] W. J. Ray and H. W. Cole, "EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes," *Science*, vol. 228, no. 4700, pp. 750–752, 1985.