4th International Workshop on Perceptual Quality of Systems (PQS 2013)
2-4 September 2013, Vienna, Austria

Recommendations for the Subjective Evaluation of Sensory Experience

Benjamin Rainer, Christian Timmerer, and Markus Waltl

Multimedia Communication (MMC) Research Group,
Institute of Information Technology
Alpen-Adria-Universität Klagenfurt, Austria

firstname.lastname@itec.aau.at

Abstract

Selecting and adopting the appropriate assessment method for conducting subjective quality assessments is a challenging task. The method decides whether the assessment is successful in delivering the correct answers to previously set up hypotheses. Therefore, in this paper we provide recommendations on test methods used in the domain of Sensory Experience. The proposed test methods comprise single stimulus and double stimulus methods. These test methods were used in previous studies and are presented in combination with the results of the subjective quality assessments with which they were used. Furthermore, we briefly outline our test setup, test design, and test content for assessing Sensory Experience which have been validated through conducted assessments.

Index Terms: Subjective Evaluation Methods, Sensory Experience

1. Introduction

In the past decade more and more multimedia content is becoming available in stereoscopic video format which shall enable an increased immersive user experience. However, still it mainly stimulates two senses, namely hearing and seeing. More and more researchers started enhancing multimedia presentations to stimulate additional senses. For example, in [1], video content is enhanced with additional ambient light effects to assess whether the additional light effects increase the viewing experience. In [2][3][4], multimedia content is enriched with additional olfactory effects. The authors in [2] enriched images with olfaction and three user studies for categorization, tagging, and recalling the tagged images. In [3], the influence of olfactory on the information recall is investigated. Therefore, different video sequences with the corresponding smell were presented to the participants. The results state that olfaction has a negative impact on the QoE. Furthermore, in [4], the authors investigated the synchronization threshold for olfactory-enhanced multimedia concluding that an asynchrony of up to 20 seconds may be tolerated by the users for certain video sequences.

In [5], we have conducted a subjective quality assessment on the influence of multiple sensory effects on the Quality of Experience (QoE) which showed promising results. Hence, our research focuses on stimulating multiple senses which goes beyond 2D/3D content. In particular, we enrich existing content with additional sensory effects (e.g., wind, vibration, light) such as defined in the MPEG-V standard [6] and together with appropriate devices (e.g., [7]) it allows to stimulate potentially all human senses. In this context, these additional descriptions are referred to as Sensory Effect Metadata (SEM) and the QoE is referred to Sensory Experience.

The International Telecommunication Union (ITU) defines a number of different evaluation methods for audio and/or video content targeting a single quality factor (e.g., video quality, transmission fidelity) [8][9]. In this paper, we provide recommendations on how to conduct subjective quality assessments for Sensory Experience including the selection of the test content, test conditions, setup, and evaluation methodologies we have used in previous subjective quality assessments.

The remainder of this paper is organized as follows. Section 2 discusses the test condition, test setup, and the test design that were used to conduct our subjective quality assessments in the field of sensory effects. Section 3 briefly describes our sensory effect dataset. Section 4 discusses our subjective quality assessments adopting a single stimulus method whereas Section 5 discusses our assessments using a modified double stimulus method including results on the reliability. Section 6 concludes the paper.

2. Test conditions, Setup, and Design

The ITU provides recommendations on test conditions for subjective quality assessments [8]. For subjective quality assessments in the domain of Sensory Experience, these recommendations provide a good starting point.

2.1. Test Conditions

Our subjective quality assessments are conducted in a room which fulfills the recommendations of the ITU for subjective quality assessments [8]. Therefore, we selected a room where the light conditions can be dimmed to the recommended candela value. The participants were seated in front of the display in a distance of around one meter. The keyboard and mouse were placed approx. 25 cm away from the front edge of the desk. The desk had a height of 72 cm.

2.2. Test Setup

In our user tests, we used displays with a size of 19”, 24”, and 26” with a resolution of 1280x1024, 1400x1050, and 1680x1050, respectively.

For rendering of sensory effects, special equipment (e.g., vibration chair, fans) is needed such as the amBX system [7] but other devices may be also appropriate. We used the amBX system, which comprises two fans with 5,000 rounds per minute (RPM), 2.1 sound system (two sound speaker RGB lights and a subwoofer), a wall washer which includes three RGB lights and a wrist rumbler. The lights provide 16 million colors.

Our experimental setup is depicted in Figure 1. The devices should be positioned around the display, i.e., left and right sound

10.21437/PQS.2013-23

Figure 1. Experimental Setup of amBX Devices.

speaker lights next to the display, the wall washer behind the display in an elevated position where the upper edge of the wall washer should end with the upper edge of the display. Furthermore, the wall washer should be placed in front of a wall such that it illuminates the wall behind the display but subjects shall not notice the actual device. The fans should be positioned next to the lights and the wrist rumbler should be put on the thighs of the participant to receive a more realistic vibration experience. The subwoofer cannot be seen in the figure as it is positioned below the table. It has to be mentioned that the subwoofer can amplify the vibration effect.

2.3. Test Design

Figure 2 depicts our general test design for conducting subjective quality assessments. In the first stage, the participants have to read the introduction which explains the purpose of the actual experiment. Furthermore, it provides information on the assessment itself, e.g., how many video sequences will be shown, how does the rating possibility look like, the rating scale may be described, and how long the assessment will last. Furthermore, a disclaimer is presented which the participants have to accept in order to participate in the subjective quality assessment. The disclaimer indicates that persons with audio/visual impairments should not take part in the experiment. Additionally, it is stated that persons with epilepsy are not allowed to take part in the subjective quality assessment. After the introduction, a pre-questionnaire is presented asking the participant to provide some demographic and educational information. The main idea of having a pre-questionnaire is that it should allow for an exhaustive demographic analysis of the participants.

Depending on the type of the assessment and whether the assessment requires a training phase our test design includes the possibility to have a separate training phase after the introduction. In our experiments, the training phase should eliminate the surprise effect and help the participants to become familiar with the stimulus presentation and the rating possibility.

After the training phase, the main evaluation starts. In our subjective quality assessments, we adhere to the recommendations of the ITU [8][9] regarding the test methods and the test design. We only modified the presentation of the stimuli from a grey background to a black background because the black background accentuates sensory effects better. For the rating phase, we used the proposed mid-grey background and intuitive rating mechanisms like a slider with labels or option boxes that allow a categorical scale.

Figure 2. Test Design.

When the main evaluation has finished a post-questionnaire is presented to the participants. This questionnaire asks the participants whether they have already participated in a similar experiment or not. This question may help in assessing the influence of experienced participants on the ratings in comparison to inexperienced participants. Furthermore, the post-questionnaire includes the possibility to give feedback. The feedback of participants may point out whether there are design issues which cannot be identified by statistical analysis.

3. Sensory Effect Dataset

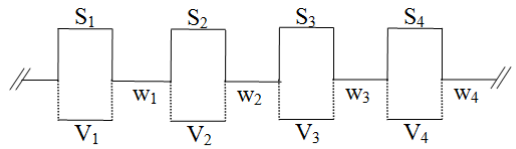
For conducting subjective quality assessments, the test content plays an important role and its selection is tightly coupled with the goal of the assessment. Thus, test content has to be selected carefully and has to fit the goals of the actual subjective quality assessment.

In [10], we presented our sensory effect dataset. During the selection of appropriate multimedia content for conducting subjective quality assessments in the domain of Sensory Experience, we noticed that the standard test content such as foreman, carphone, etc. are not suitable to be enriched with sensory effects because they do not offer enough possibilities to enrich the actual content by sensory effects. For example, these video sequences do not provide scenes where sensory effects may accentuate the current happening. Another problem of the standard test content is its duration. In the domain of Sensory Experience, video sequences may last longer because the effects should be sound and valid to the user such that no effect ends abrupt. The authors of [11] have shown that using video sequences longer than 10 seconds do not have a negative influence on the results. On the other hand, it has been shown that video sequences longer than 30 seconds elicit the effect of a more realistic viewing experience.

Therefore, we collected in total 76 video sequences from the genres: action (38), documentary (12), sports (8), news (5), and commercial (13). Each video sequences was annotated with the use of our Sensory Effect Video Annotation (SEVino) tool [12]. For verifying that the annotated effects are sound and valid, we established an in-house review process, where the first coarse annotation was refined until it fitted the needs of the reviews. We also took different quality version for the content into account. The dataset comprises single video sequences more than once in different quality version. Furthermore, we tried to cover the most important genres for providing a dataset for subjective quality assessments in the domain of Sensory Experience. The dataset is available at the Sensory Experience Laboratory (SeLab) at [13].

4. Single Stimulus Evaluations

In [14], we conducted a subjective quality assessment using a single stimulus method to assess whether sensory effects are able to compensate low video bitrates. Therefore, the video sequences were presented in different bitrates (resulting in different qualities) with sensory effects (wind, vibration and light) and without sensory effects. For this purpose, we selected the Earth



S_1, S_2, S_3, S_4 … Test sequences (reference and modified)
V_1, V_2, V_3, V_4 … Voting for corresponding sequence
w_1, w_2, w_3, w_4 … Grey screen (5 – 10 seconds) with overall voting for corresponding sequence

Figure 3. Combined Single Stimulus Evaluation Method [14].

Figure 4. SSCQE Voting Device and Mapping to Voting Scale [14].

(documentary) and Babylon A. D. (action) video sequences from our dataset [10].

4.1. Single Stimulus Test Method

Figure 3 depicts the test method for the single stimulus based subjective quality assessment. We used a combination of the Single Stimulus Continuous Quality Evaluation (SSCQE) and the Absolute Category Rating (ACR) with Hidden Reference (HR). Thus, the participants had the possibility to rate during the stimulus presentation and afterwards, they were asked to rate the overall perceived quality of the presented stimulus.

Figure 4 depicts the voting device and the rating scale. The combined single stimulus method allows the participant to state the QoE during a stimulus presentation and, additionally, to give an overall rating after the stimulus presentation. The continuous voting was done by pressing the appropriate buttons on the voting device. Thus, once a participant pressed a button the rating taken for the time period until the participant pressed another button. Furthermore, the single stimulus method provides the possibility to assess the users' QoE in a home viewing condition.

4.2. Experiment Results

In total, 24 students (11 female and 13 male), aged between 18 and 37 years took part in this subjective quality assessment. Three were identified as outliers and those were not taken into account for the analysis of the results according to [9]. The Mean Opinion Score (MOS) and the 95% Confidence Interval (CI) are calculated using the ratings after each stimulus presentation (i.e., w_i in Figure 3).

Figure 5 illustrates the MOS for the Earth video sequence with sensory effects and without sensory effects. It can be seen that with lower quality versions of the content sensory effects have a positive influence on the QoE such that participants had a higher QoE with sensory effects than without sensory effects for all versions of the Earth video sequence.

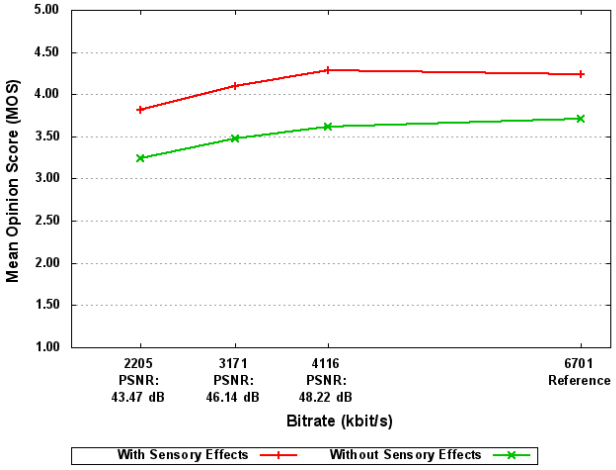


Figure 5. MOS vs. PSNR/Bit-rate for Earth [15].

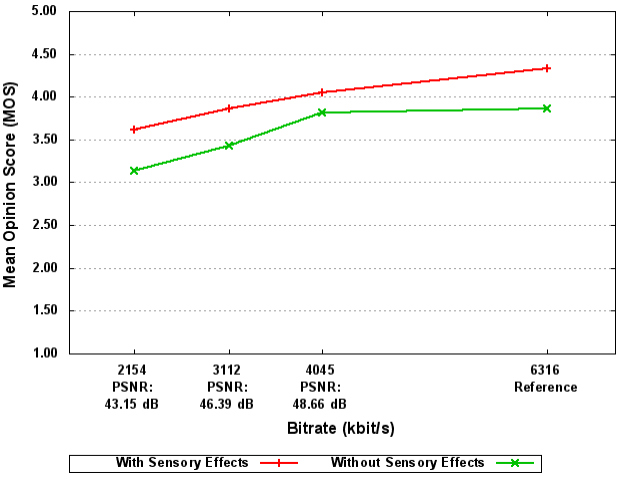


Figure 6. MOS vs. PSNR/Bit-rate for Babylon A.D. [15].

Figure 6 depicts the MOS for the Babylon A.D. video sequences and paints the same picture as Figure 5. Again, the QoE with sensory effects is higher for all versions of the Babylon A.D. video sequence than without sensory effects.

4.3. Discussion on Combined Single Stimulus

As we used a voting device for the SSCQE part of the subjective quality assessment, participants stated that the voting during the actual stimulus presentation stressed them resulting in a mentally overload during the actual stimulus presentation. Therefore, the ratings cannot be reliably used for deducing conclusions from the SSCQE method. Due to this circumstance, we expect that the handling of the buzzer was not intuitive enough and the content included to many scene changes to be used with the SSCQE method. Therefore, we suggest providing other mechanical devices for a continuous rating possibility such as mechanical sliders.

However, the single stimulus methods recommended by [9] can be used without modifications in the domain of Sensory Experience because they assess the absolute rating on a single presentation of the video sequence with or without sensory effects.

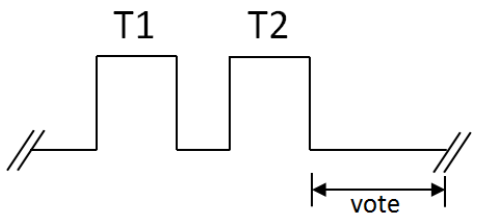


Figure 7. DCR Evaluation Method [5].

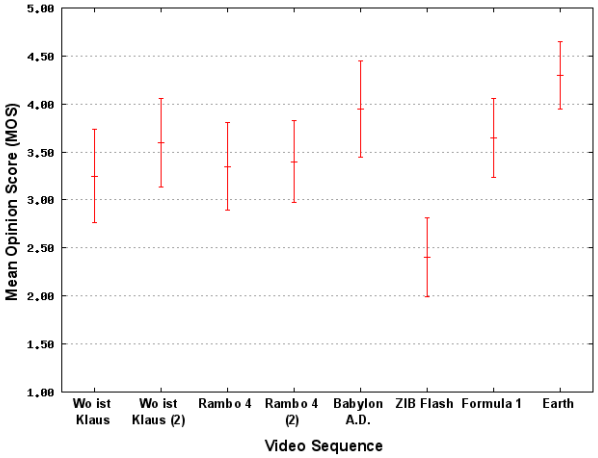


Figure 8. Results with the Use of the DCR Evaluation Method [5].

5. Double Stimulus Evaluations

In [5] and [16], we adopted the Degradation Category Rating (DCR) and the Double Stimulus Continuous Quality Scale (DSCQS), respectively. Therefore, we picked up the existing recommendations of both methods and adopted them to our needs. The integral part of both evaluation methods is the double stimulus presentation, where the impaired or the test condition is displayed before or after the reference and after the presentation of both stimuli, the participants have to rate the quality with respect to the reference.

5.1. DCR Test Method

Figure 7 illustrated the used DCR method in [5]. The DCR provides a voting after showing the second video sequence by using a categorical impairment scale. We exchanged this scale to a categorical enhancement scale, where 1 depicts "very annoying", 2 "annoying", 3 "imperceptible", 4 "little enhancement" and 5 "big enhancement" to let participants judge on the enhancement of QoE thanks to sensory effects. As we evaluate the Sensory Experience and, thus, the reference sequence and the test sequence can be clearly distinguished (e.g., sensory effects are present or not), we did not allow for changes in the order of the sequences for a single test presentation. Therefore, at T1 the reference was presented and at T2 the test condition was presented. Between T1 and T2 a grey screen was presented to the participants as recommended by [8]. After both stimuli presentations, the participants had the possibility to rate their perceived QoE.

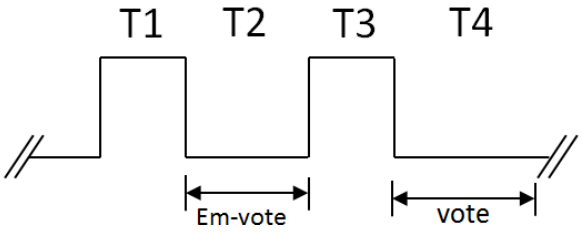


Figure 9. Modified DSCQS [16].

5.2. DCR Experiment Results

The study presented in [5] should give us clarity about the impact of sensory effects on the QoE. The idea was to assess whether sensory effect enhance the viewing experience or not depending on the genre. Therefore, we selected eight video sequences from our dataset from the genres commercial, action, news, documentary, and sports. The selected video sequences are depicted by the x-Axis of Figure 8. In this experiment, 25 participants took part (12 female and 13 male), five participants were identified as outliers. Figure 8 depicts the MOS and the 95% CI for each video sequence. The results show that for nearly each video sequence the MOS is above 3. The only exception is the genre news where sensory effects had a negative impact on the QoE. As the CI is very wide it is difficult to draw a general conclusion on the enhancement of the QoE due to presence of sensory effects. However, for the genres represented by the selected video sequences except for news, sensory effects had a positive impact on the QoE.

5.3. Discussion on Combined Single Stimulus

The DCR method may be used without any modifications for assessing sensory effects. With our other user studies we experienced that the selection of the rating scale is crucial for later statistical analysis. Therefore, we suggest selecting a rating scale and sticking to it for similar subjective quality assessments. Mapping ratings obtained by an ordinal scale to an interval scale introduces some error that is not negligible (e.g., transforming a categorical scale to a continuous scale with an interval of [0,100]).

5.4. DSCQS Test Method

In [16], we conducted subjective quality assessments with the modified DSCQS described in the previous section. The subjective quality assessments were conducted in Austria at the Alpen-Adria-Universität Klagenfurt (AAU) and Australia. In Australia the subjective quality assessment was conducted by two universities, namely the University of Wollongong (UoW) and the Royal Melbourne Institute of Technology (RMIT). For conducting this subjective quality assessment we used a continuous rating scale within an interval of [0,100], thus, we adopted the DSCQS method. Due to the purpose of a subjective quality evaluation, we added after each stimulus presentation a possibility to rate the perceived emotions and their intensity on a continuous rating scale with the interval of [0,100] in addition to the overall rating possibility after the second stimulus presentation. Figure 9 depicts the modified DSCQS with the additional rating possibilities. T1 represents the presentation of the video sequence without sensory effects. T2 illustrates the

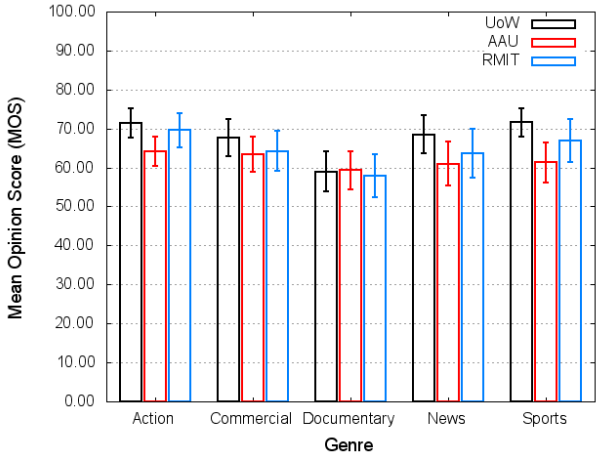


Figure 10. Enhancement of QoE for each Genre for all three User Studies [16].

rating for emotions and their intensity. T3 depicts the presentation of the same video sequence with sensory effect and, finally, T4 states the rating for emotions and their intensity for the video sequence with sensory effects. After the emotion rating, the rating for the perceived QoE is shown. The continuous scale for the perceived QoE is divided into 0 – 20 "very annoying", 20 – 40 "annoying", 40 – 60 "imperceptible", 60 – 80 "little enhancement", 80 – 100 "big enhancement. The categories were explained in the introduction of the subjective quality assessment.

With this modified DSCQS, we combined a single stimulus method with a double stimulus method by adding additional rating possibilities after each stimulus presentation. This was necessary for assessing the emotional response with the reference condition and the test condition for each video sequence. This allows us to conduct statistical analyses on the difference of the emotional response between the video sequence without sensory effects and with sensory effects.

In this paper, we focus on the validation of the modified DSCQS by investigating whether the combination of single stimulus and double stimulus has an impact on the reliability of the results retrieved by the double stimulus fragment of the resulting stimulus method.

5.5. DSCQS Validation and Experiment Results

For the video sequences, we selected three video sequences for each of the following genres from our dataset: action, commercial, documentary, news, and action. The user study was conducted under the same ambient conditions at all three locations. In total, 68 (36 female and 32) people participated in the experiment, where 26 (18 female and 8 male) participated in the study conducted at AAU, 21 (6 female and 15 male) at UoW and 21 (12 female and 9 male) at RMIT. In our case, the empirical kurtosis and variance are used for the test statistics. The screening according to the β_2-Test did not reveal any outliers in the conducted user studies [9].

For determining whether the modified DSCQS delivers reliable and robust results, we used the ratings from the two studies from Australia. The reason for excluding the user study conducted in Austria is that cultural influences may have an impact on the ratings. There are several ways of determining whether a test method delivers reliable data. One possibility to check the reliability of a test method is to repeat the test with the same participants and, therefore, the correlation may serve as test statistics to check whether the rating of each participant has significantly changed. The drawback of this method is that it relies on the behavior of each participant. Thus, we used ratings from RMIT and UoW with different participants but from the same state.

For identifying differences within the genres, we used the F-Test prior to the student's t-test [17]. The test statistics for the F-Test is given by the hypothesis that the variance form two samples are equal (H0) [17]. The F-Test revealed that the variances are equal within the genres. For none of the genres, the student's t-test revealed a significant difference for the means between the ratings from UoW and RMIT with the following values for p, t and $\alpha = 0.05$ for each genre: action $p = 0.54$ and $t = -0.61$, commercial p=0.34 and $t = 0.961$, documentary $p = 0.77$ and $t = 0.29$, news $p = 0.24$ and $t = 1.19$, sports $p = 0.16$ and $t = -1.41$. If the single stimulus would have caused a bias in the ratings of the double stimulus, we would have seen a significant difference between the random variables within the same genre because the samples are drawn from the same population.

Figure 10 depicts the results for the subjective quality assessment conducted by AAU, UoW, and RMIT. It can be seen that for nearly all genres the MOS points are above 60 which indicates an enhancement of the QoE. Only with the documentary the MOS scores are exactly 60 (AAU) or below. Nevertheless, we may conclude that sensory effects enhance the QoE. Some genres benefit more than others from sensory effects. Another interesting point is that the MOS from AAU are always below RMIT and UoW except for the genre documentary which was slightly above the Australian ratings but which is negligible.

5.6. Discussion on modified DSCQS

The statistical analysis stated that there is no significant difference within the genres. This shows us that the combination of the single stimulus and double stimulus methods still deliver reliable results because the means and the variances do not significantly differ from each other within every genre. Thus, we may consider this combination as appropriate for assessing the enhancement of the QoE with sensory effects. The additional rating possibilities after each stimulus presentation added by the single stimulus had no influence on the ratings assessed by the double stimulus rating phase.

6. Conclusions

We presented the test methods used in our conducted subjective quality assessments. The single stimulus methods recommended by the ITU can be used without modifications. The double stimulus methods have proven useful and we had only to modify the DSCQS method to fit our needs. Therefore, we have shown that modifying the DSCQS method did not have any impact on the reliability of the results of our conducted user studies. Thus, we can conclude that the single stimulus test methods recommended by the ITU recommendations can be used without any modification in the domain of Sensory Experience. As already mentioned, on the one hand we had to modify the DSCQS, but on the other hand the DCR method fitted the need of assessing the influence of sensory effects on the QoE. For the

case that modifications are necessary, it has to be verified that the modified test method still delivers reliable results. Additionally, we have provided results from our previous conducted subjective quality assessments. Furthermore, we presented the results of previously conducted subjective quality assessment and we can conclude that sensory effects enhance the QoE for most of the tested genres. Additionally, sensory effects help in rendering video quality degradations imperceptible up to a certain extent.

Acknowledgments: This work was supported in part by the EC in the context of the ALICANTE (FP7-ICT-248652), SocialSensor (FP7-ICT-287975), and QUALINET (COST IC 1003) projects and partly performed in the Lakeside Labs research cluster at AAU.

7. References

[1] Begemann, S. H. A., "A Scientific Study on the Effects of Ambilight in Flat-Panel Displays", http://bit.ly/pxbZYA, 2005.

[2] Brewster, S., McGookin, D., and Miller, C., "Olfoto: Designing a Smell-based Interaction", *In Proc. of the SIGCHI (CHI'06)*, ACM, New York, USA, 653-662, 2006.

[3] Ghinea, G. and Ademoye, O. A., "Olfaction-enhanced Multimedia: Bad for Information Recall?", *In Proc. of ICME 2009*, pp. 970-973, 2009.

[4] Ademoye, O. A, and Ghinea, G., "Synchronization of Olfaction-Enhanced Multimedia", *IEEE Transactions on Multimedia*, vol. 11, no. 3, pp. 561-565, 2009

[5] Waltl, M., Timmerer, C., and Hellwagner, H., "Increasing the User Experience of Multimedia Presentation with Sensory Effects", *In Proc. of WIAMIS 2010*, Desenzano del Garda, Italy, 2010.

[6] ISO/IEC 23005, Information Technology – Media Context and Control, 2011.

[7] amBX UK Ltd., http://www.ambx.com

[8] ITU-R Rec. BT.500-13, "Methodology for the subjective assessment of the quality of television pictures", 2012.

[9] ITU-T Rec. P.910, "Subjective video quality assessment methods for multimedia applications", 2008.

[10] Waltl, M., Timmerer, C., Rainer, B., and Hellwagner, H., "Sensory Effect Dataset and Test Setups," *Proceedings of the 4th International Workshop on Quality of Multimedia Experience (QoMEX'12)*, Yarra Valley, Australia, pp. 115-120, Jul. 2012.

[11] Froehlich, P., Egger, S., Schatz, R., Muehlegger, M., Masuch, K., and Gardlo, B., "QoE in 10 Seconds: Are Short Video Clip Lengths Sufficient for Quality of Experience Assessment?", *In Proc. of QoMEX 2012*, Yarra Valley, Australia, pp. 242-247, Jul. 2012.

[12] Waltl, M., Timmerer, C., and Hellwagner, H., "A Test-Bed for Quality of Multimedia Experience Evaluation of Sensory Effects", *1st Int'l Workshop on Quality of Multimedia Experience (QoMEX'09)*, San Diego, USA, pp. 145-150, Jul. 2009.

[13] Sensory Experience Lab (SELab), http://selab.itec.aau.at (Last access: July. 2013).

[14] Waltl, M., Timmerer, C., and Hellwagner, H., "Improving the Quality of Multimedia Experience through Sensory Effects", *In Proc. of QoMEX 2010*, Trondheim, Norway, Jun. 2010.

[15] Waltl, M., "The Impact of Sensory Effects on the Quality of Multimedia Experience", *PhD Thesis*, Klagenfurt, Austria, March 2013.

[16] Rainer, B., et al., "Investigating the Impact of Sensory Effects on the Quality of Experience and Emotional Response in Web Videos", *In Proc. of QoMEX 2012*, Yarra Valley, Australia, pp. 278-283, Jul. 2012.

[17] Rice, J. A., "Mathematical Statistics and Data Analysis", *Duxbury Advanced Series, Brooks/Cole*, ISBN 9780534399429, 2009.