VALIDATION OF A NOVEL APPROACH TO SUBJECTIVE QUALITY EVALUATION OF CONVENTIONAL AND 3D BROADCASTED VIDEO SERVICES

Jesús Gutiérrez , Pablo Pérez , Fernando Jaureguizar , Julián Cabrera , and Narciso García

ABSTRACT

Transmission errors are the main cause of degradation of the quality of real broadcasted video services. Therefore, knowing their impact on the quality of experience of the end users is a crucial issue. For instance, it would help to improve the performance of the distribution systems, and to develop monitoring tools to automatically estimate the quality perceived by the viewers. In this paper we validate a subjective evaluation approach specifically designed to obtain meaningful results of the effects of degradations caused by transmission errors. This methodology has been already used in our previous works with monoscopic and stereoscopic videos. The validation is done by comparing the subjective ratings obtained for typical transmission impairments with the proposed methodology and with the standard method Absolute Category Rating. The results show that the proposed approach could provide more representative evaluations of the quality of experience perceived by end users of conventional and 3D broadcasted video services.

Index Terms— quality of experience, subjective evaluation, 3DTV, visual comfort, ACR

1. INTRODUCTION

Nowadays, in addition to traditional video broadcasting systems (e.g., terrestrial, satellite, and cable), there is a wide offer of video services delivered over packet-based networks, from IPTV over networks controlled by service providers, to streaming services over the internet.

Therefore, audiovisual services are these days much more accessible to consumers, whose quality expectations have increased as a result of that variety of offers. Thus, knowing the quality perceived by the end users, and how the impairments that affect the video through the transmission chain influence their Quality of Experience (QoE), is a crucial aspect for those sectors involved in the development of video services. This knowledge helps them in the deployment and maintenance tasks to guarantee the performance of the distribution systems.

In this sense, the most interesting techniques to know the QoE of the users are the objective metrics aimed at estimating automatically the perceived quality [1]. However, the development of this type of algorithms is still an open issue due to their limited performance, especially in the case of no-reference metrics [2], which, in fact, are the most interesting in broadcasting video services since they could be used for monitoring tasks [3]. Therefore, subjective assessment tests are still considered the most reliable way to evaluate the users' QoE, and they are an essential tool to measure the performance of the video distribution systems. Although these methods are not applicable in practical cases, they provide useful information to develop and evaluate objective metrics to be used in broadcasting scenarios.

Many subjective assessment methodologies have been proposed to evaluate audiovisual quality, including some methods standardized by international organizations, such as the International Telecommunication Union (ITU), the Moving Picture Experts Group (MPEG), or the Video Quality Experts Group (VQEG). However, this kind of assessment techniques for audiovisual content is still focus of research, especially with the aim of designing methods that provide representative and meaningful results in relation with the quality perceived by the end users of broadcasted video services.

Furthermore, this research activity has nowadays increased with the arrival of 3D video services to the consumer market, not only in Blu-ray, but also in broadcasting channels. Since various new factors influence the visual experience of the users of 3D video services [4], many efforts are being carried out to develop methodologies to subjectively evaluate the QoE related to those services.

Taking this into account, in this paper we analyze the validity of a novel subjective methodology to evaluate the QoE of the end users of video distribution systems. With this objective we compare the standard Absolute Category Rating (ACR) methodology with an evaluation procedure proposed, and successfully used, in our previous works [5], both with monoscopic and stereoscopic videos. This approach aims at estimating the impact on the quality of typical transmission events that may cause degradations on the audiovisual content. To obtain reliable conclusions, the tests are designed to respect, in the way it is possible, real home viewing conditions. Moreover, the methodology could

be applied to evaluate 3D video content, allowing the estimation of factors related to the stereoscopic perception like visual comfort.

The rest of the paper is structured as follows. In section 2, a review of the existing approaches for subjective evaluation of the quality provided by video distribution systems is presented, considering both conventional and 3D video. Section 3 describes the details of the proposed methodology. In sections 4 and 5 the settings of the subjective experiment and the results are presented respectively. Finally, in section 6 the conclusions of the work are exposed.

2. RELATED WORK

The strong efforts carried out by the research community in the last years designing methodologies for subjective evaluation of the quality of multimedia services led to the standardization of some techniques that provide reliable results in certain studies. One of the most popular recommendations related to this issue is the ITU-R BT.500 [6], which proposes diverse methodologies for different application cases.

For example, double-stimulus methods, in which the unimpaired original content is shown to the observers as a reference to evaluate the test content, are described to be applied especially for the evaluation of coding artifacts (e.g., blockiness, blurring, etc.). In fact, Double-Stimulus Impairment Scale (DSIS) and Double-Stimulus Continuous Quality Scale (DSCQS) have been extensively used in many studies, like in the subjective tests of the proposals for the new High-Efficiency Video Coding (HEVC) standard [7].

On the other hand, single-stimulus methodologies are usually considered more appropriate for evaluation of impairments caused by errors in the transmission of the content (e.g., packet-losses, jitter, etc.). This is due to the fact that in these methods only the test content is shown to the observers, because presenting a reference could distance the observers from real viewing conditions. A popular example of these techniques is the Single-Stimulus Continuous Quality Evaluation (SSCQE) [6]. It is based on the continuous evaluation of the video quality using an electronic device, like a slider, to record the ratings, allowing the evaluation of time-varying and fluctuant degradations.

Other widely used single-stimulus methodologies were described in the recommendation ITU-T P.910 [8]. For example, ACR allows the evaluation of short test sequences (around 10s) that are rated by the observers during an evaluation period after the clip in which the screen turns grey. A variation of this method includes unimpaired sequences, called *hidden references* (giving the name to the method, ACR-HR), which are also evaluated as tests sequences, to compare the scores of the reference and the impaired sequences. These methodologies have been extensively used in studies of coding and transmission impairments [9].

Apart from standardized methodologies, which recommend, in addition to the evaluation procedures, the environmental conditions to run the tests correctly, some other approaches have been proposed with the aim of getting representative results of what viewers perceive in their households. For example, a different study is presented in [10], were opinions about the effects of typical degradations in digital video services were directly collected from the consumers by means of surveys. In addition, this study investigated the acceptance of the consumers regarding those impairments, taking into account economic aspects. These aspects are very difficult to be analyzed, especially in laboratory environments, however, they are very important because the quality demands of the consumers, and QoE in consequence, are highly related to the price of the service.

Another methodology in which the subjective evaluation is carried out at home is proposed in [11], to avoid the influence that controlled test conditions could have on the subjective results. To stress the importance of carrying out the evaluations in real-life environments, the authors evaluated the same degradations with the proposed method and the ACR standard methodology, and compared the obtained results.

In contrast to the variety of methodologies for subjective evaluation of conventional video services, and the deep knowledge acquired with the extensive research during the last decades, nowadays a strong effort is being made to identify the most appropriate ways to subjectively evaluate the QoE of the users of 3D video services. Although ITU published a recommendation for subjective evaluation of stereoscopic content [12], it is merely based on the ITU-R BT.500 [6] and only contemplates a few factors related to 3D QoE, such as image quality and depth perception. Therefore, some organizations are working towards standard methodologies, like VQEG under the project 3DTV [13].

Even so, some subjective tests have been already presented evaluating the performance of coding methods [14] and the effects of transmission errors [15], using methodologies based on those existing for traditional video, trying to take into account intrinsic factors related to the 3D visual experience, like visual discomfort.

Considering all this, we proposed in our previous work a subjective testing approach with the aim of contributing to achieve a reliable methodology for the evaluation of the QoE perceived by the users of broadcasting video systems [5]. As described in the next section, it was designed to provide representative results of what they perceive in their homes, and is also valid for 3D video content.

3. PROPOSED SUBJECTIVE METHODOLOGY

The main objective is to know the effects of transmission impairments on the QoE perceived by the end users of broadcasted video services. Therefore, the ideal manner would be to obtain direct subjective reports from their homes. Obviously, this is not feasible, but it is clear that the

subjective assessment tests designed to analyze these effects should emulate real viewing conditions. This way, the results obtained from the subjects will be representative of what they would perceive in their real-life experiences, and will help in the design of monitoring tools to automatically estimate the perceived quality [3].

Taking this into account, one of the main aspects is the test material to use. Normally, the studies carried out according to the standard methodologies consider the use of short sequences whose content is usually not interesting for the observers. This way, it is difficult to obtain meaningful results from the experiments, since in real situations, viewers watch long sequences, and they are not focused on detecting degradations, but on the content itself. These aspects have been taken into account in [11], where full-length movies were used to evaluate transmission impairments. However, in this case the evaluation of the degradations is provided when the movie is finished (after approximately 120 minutes), which may be too much to remember the impact of each distortion. In contrast, in the standard methodologies, like ACR, the evaluation is provided after the short test sequence, which guarantees that no impairments will be forgotten. Another possible approach is the SSCQE methodology. However, the requirement of a device to record the evaluations is a drawback, and typical transmission errors in digital video do not require a strictly continuous evaluation.

Taking this into account, our proposed methodology is based on the use of long sequences, but evaluating the degradations after they take place [5]. Then, in order to not break the continuity of the video (as it would happen using a similar method to ACR since the evaluation is done during a period in which the screen turns grey), after each inserted degradation, a period without any impairment is left while the observers can rate. This is depicted in Fig. 1. The test sequences are divided into segments of certain duration (in this test we set it to six seconds). The impairments under study are introduced in alternate segments (T_i), and evaluated during the following segment, which is left unimpaired. In addition, at the beginning of the sequence, the first segment is left unimpaired in order to provide to the subjects a reference of coding quality. In some cases, no degradations were inserted in some test segments, being hidden references. The total duration of the sequences could be set depending on the particular conditions of the test, taking into account the limited duration of the subjective experiment sessions.

For its simplicity, the proposed way to recollect the scores of the subjects is using questionnaires. Although
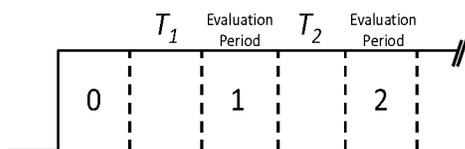
some observers could need to avert their gaze from the screen to write their score, during the evaluation period there are no impairments in the test video. To indicate in which box they should write a cross, the unimpaired segments have a small number printed on the right-bottom corner. The evaluation is done according to the five-grade impairment scale recommended in [6].

As aforementioned, this methodology is also applicable to the evaluation of the impact of transmission errors on the QoE related to 3D video. In fact, a method to evaluate also the visual discomfort caused by these types of degradations is proposed. Specifically, the observers are normally asked to write a cross on the box corresponding to the rate of impairment scale and to the evaluated segment. Then, in the case that the observed impairment caused them some kind of discomfort or annoying visual difficulties to correctly perceive the 3D effect, they should write a single line instead of a cross. This way it is possible to identify which impairments could cause visual discomfort.

4. SUBJECTIVE TEST SETUP

4.1. Environment and equipment

The experiments were carried out in a room at the Universidad Politécnica de Madrid with controlled ambient light. The sequences were displayed in a stereoscopic 42" full-HD Panasonic television set, and the observers used the corresponding active shutter glasses of the set. According to the recommendations of the manufacturer, the viewing distance was set to three times the height of the screen.

4.2. Analyzed impairments

The degradations to analyze in this type of tests should be similar to those caused by typical transmission errors in real broadcasting video systems. Taking into account the most common degradations reported by the end users [10], and following with our previous studies [3][5], the impairments that have been simulated are described in Table I.

Specifically, different patterns of video packet losses were considered, causing macroblocking effects with different severity. These effects are the result of a typical error concealment technique of the decoder that substitutes the lost macroblocks by the corresponding from previous frames that have been correctly decoded. Also, video freezes were taken into account simulating different durations, and in combination with audio loss (outage events). In addition, other typical degradations were analyzed, which are caused by the strategies used by service providers to reduce the amount of data to transmit when the performance of the network is reduced. Commonly, in conventional video services the two main possibilities are to discard frames reducing the framerate, or to reduce the video coding quality. Besides, when dealing with 3DTV other approaches could be considered, for example, to switch to a monoscopic version of the transmitted video [15], or to reduce the coding bitrate using asymmetric coding [16].

Fig. 1. Diagram of the structure of the test sequences

TABLE I. PROPERTIES OF THE IMPAIRMENTS

ID	Type	Description
R1	Bitrate reduction	Reduction of the coding bitrate to the half.
R2	Bitrate reduction	Reduction of the framerate to the half.
R3	Bitrate reduction	Switch to 2D. Equivalent to reduce the original global bitrate to the half.
R4	Bitrate reduction	Asymmetric coding equivalent to reduce the original coding bitrate to the half.
E1	Video packet loss	Loss of one whole b-frame.
E2	Video packet loss	Affecting one whole I-frame, during 1/3 of the GOP.
E3	Video packet loss	Affecting one whole I-frame, during 1 GOP.
V1	Video freeze	Duration of 1s.
V2	Outage	Duration of 2s. The video continues with the corresponding frame to the actual moment.
V3	Outage	Duration of 2s. The video continues with the subsequent frame of the frozen image.

4.3. Test material

In this work only stereoscopic sequences were used, but the methodology has been already used with traditional video [5]. The main properties of the source videos used in the experiments are shown in Table II. Both sequences were encoded using H.264/AVC standard, and a bitrate of 12Mbps. The coding configuration consisted of closed GOPs of approximately one second of duration and IbBbP structure (where b represents B frames used as reference). To simulate the transmission of the sequences over IP networks, RTP packetization was used.

TABLE II. PROPERTIES OF THE TEST SEQUENCES

ID	Sequence	Resolution	Framerate
SRC1	Movie	1920x1080p	24fps
SRC2	Sports	1280x720p	50fps

The sequences have an approximate duration of five minutes. During this time, the content of SRC1 was quite diverse, having dialog scenes and action intervals with high motion activity. On the other hand, SRC2 was a football match.

To generate the test sequences to be evaluated with the proposed methodology, each source video was divided into segments according to the structure described in the previous section. Then the errors described in subsection 4.2 were randomly introduced in alternate segments of the sequences.

On the other hand, in the case of the sequences to be evaluated using ACR methodology, they were composed of impaired segments extracted from the two test sequences created for the other methodology. Two different test sequences were generated, having both of them the same test segments but in different randomized orders to compensate contextual effects. At the beginning of the sequences a message indicating the start of the test was displayed. In addition, the message in grey background between each impaired segment contained a number indicating the observers in which box of the questionnaire they had to mark the evaluation. The length of the impaired segments and the messages was six seconds, which is the same duration used for test and evaluation segments in the proposed methodology.

4.4. Observers

A total of 36 observers participated in the tests, recruited among the students and personal of the university. Their ages were ranged between 20 and 44, with an average age of 25. All of the observers have normal color vision, and visual and stereoscopic acuity.

The ratings provided by the subjects were screened to detect possible random evaluations, or misunderstandings with the evaluation procedures. This was done controlling the scores provided to the hidden references and discarding those observers that did not provide high values [8].

4.5. Assessment sessions

The evaluation of the test sequences was carried out in different sessions of approximately 30 minutes. At the beginning of each session a detailed explanation of the experiment was provided to the subjects, including the aim of the tests to analyze the impact of transmission errors in 3DTV, the type of errors, and especially, the basis of the assessment methodologies. The rest of the sessions consisted of a presentation of the two test sequences to be evaluated with the proposed methodology. Then, the two sequences generated to be evaluated following the ACR procedure were displayed. A maximum of five subjects participate in each session.

5. RESULTS

5.1. Comparison between ACR and the proposed methodology

The results obtained in the tests were statistically analyzed to carry out the comparison between both assessment methodologies. Specifically, the Mean Opinion Scores (MOS) of the evaluations provided by the observers for each one of the considered degradations were computed. In the case of the ACR evaluations, since each test segment was evaluated twice by each observer, the mean value was taken as the resulting MOS.

In Fig. 3, the results obtained from all the impairments considering the degradations inserted in both source videos are shown, except for the reduction of framerate (R2). Since the source sequences have different framerates, the effects of this degradation should be analyzed separately for each video, as depicted in Fig. 2.

As it can be seen in Fig. 3, slightly better scores were obtained with the proposed methodology in the cases of macroblocking, which can be caused by the fact that they could be less perceptible, since a long continuous sequence is used, the observers are not focused on detecting the impairments, but on the content of the video.
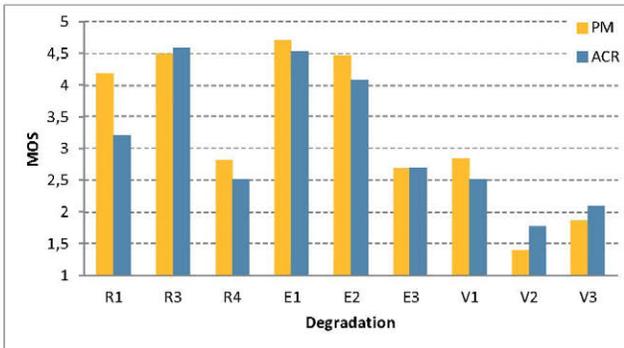
Fig. 3. Comparison of the results with ACR and the proposed method

Fig. 2. Comparison of the results for framerate reduction

A similar conclusion could be extracted from the comparison in the cases of coding bitrate reductions (R1 and R4). In these cases, the higher values obtained with the proposed methodology could be also due to the fact that the observers perceive the reduction with respect to the original bitrate as it would happen in real-life situations. On the contrary, with ACR, observers only evaluate the coding quality of the segment without any contextual reference to the previous content of the video.

Interesting results were obtained for the video freeze (V1) during one second (duration of a GOP), which could be caused by a packet loss affecting the header information and avoiding the decoding of the whole GOP. As it can be seen, the results are also better in the case of the proposed method, and similar to the scores for a video packet loss causing macroblocking propagated along all the GOP. This could be caused because in 3D, the effects of macroblocking are very annoying causing binocular rivalry [5].

Concerning the outage degradations causing longer freezes and also the loss of audio (V2 and V3), better ratings were obtained with the ACR methodology. This could be due to the severe temporal interruption of a long continuous sequence that the observers are following in the case of the proposed methodology. In ACR, the observers just evaluate an error stop event within a six second segment, with no contextual reference to the rest of the sequence.

Finally, almost identical values where obtained for the degradation based on the switching to a monoscopic version of the 3D video (R3). In fact, it mainly caused annoyance due to the transition itself, more than the loss of 3D effect, which is reflected in the high MOS values.

In Fig. 2, the results obtained with the two evaluation methodologies for the framerate reductions (R2) and for each sequence are shown. As it can be seen, very similar results were obtained since the effect of discontinuous transitions between adjacent frames is equally annoying in both cases. The higher framerate of SRC2 causes the better ratings in this case.

To analyze the statistical significance of the differences between the evaluations with ACR and the proposed method, a non-parametric analysis was carried out. This was because the Kolmogorov-Smirnov [17] test was firstly
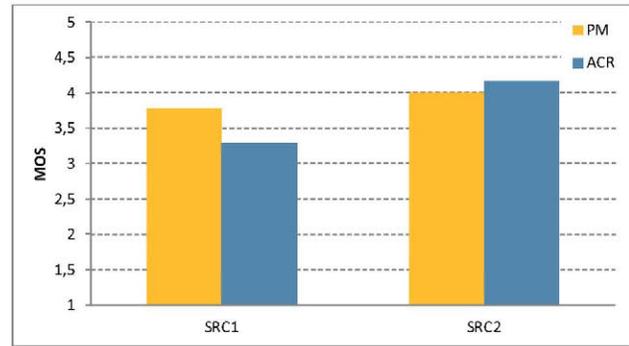
applied to the sets of obtained values, and it shown that they were not normally distributed at the 5% significance level. Therefore, since the sets of values were not exactly paired, the Mann-Whitney test [17] was applied. As a result, statistical significance was obtained, at the 5% significance level, in the cases of bitrate reduction (R1), and error stop (i.e., video freeze and audio loss) continuing the playback skipping to the frame corresponding to the actual instant (V2). Especially the effects of these types of degradations are highly dependent on the contextual reference to the content previously seen. For the other degradations, the fact that there is no statistical significance could be interpreted in the sense that the proposed method is as valid as ACR.

5.2. Visual comfort

The results obtained from the evaluations of visual discomfort are represented in Fig. 4, where it is shown the percentage of the observers that expressed that they felt visual difficulties in the analyzed degradations. As it can be seen, macroblocking effects and asymmetric coding of the stereo views (R4) are the most annoying impairments in this sense. In the case of macroblocking effects it has been shown that these impairments could cause binocular rivalry and, consequently, visual discomfort [5]. Regarding asymmetric coding, when the bitrates used for each stereo view cause a considerable loss of structural information of the images, binocular rivalry could also be produced to the observers.
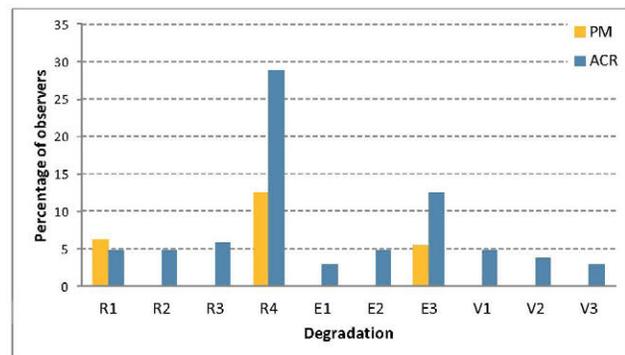

Fig. 4. Percentage of observers that felt visual discomfort

Another aspect extracted from the tests is that in the case of ACR more observers felt visual discomfort. This could be probably due to the fact that the human visual system should adjust the stereoscopic perception in each test segment. In fact, the visualization of the message segments (without depth effect), and the evaluation of test segments with probably very different disparity could cause visual discomfort. On the other hand, with the proposed methodology, the adjustment of the 3D perception by the visual system is smoothly done during the sequence. However, similar effects could appear in the cases when the observers need to avert their gaze from the screen to write on the questionnaire. Thus, more appropriate ways to recollect that scores are focus of further research.

6. CONCLUSIONS

Nowadays, knowing the effects of transmission impairments on the QoE of the end users of broadcasting video services is a key issue. To get that knowledge, subjective assessment tests are considered the most reliable methods. However, the results obtained from these tests should be meaningful and representative of what viewers perceive in real-life viewing situations. Therefore, subjective tests should not drastically distance the observers from those real viewing conditions.

In this paper, a validation of a subjective methodology designed taking into account this requirement is presented. In fact, a comparison with the standard ACR methodology is carried out, considering typical transmission impairments that take place in video delivery networks. The results show that the proposed methodology could reflect better the effects of some degradations on the QoE perceived by the end users in their homes, especially in the cases in which the impact of the impairments have some contextual relation with the rest of the sequence, like in the case of coding quality reductions or outage events.

Moreover, the proposed methodology could be applied to evaluate the impact of transmission errors in 3D video, taking into account also visual discomfort. In fact, this methodology seems to be more adequate to evaluate visual comfort than ACR, since using a continuous sequence avoids severe changes in the accommodation of the visual system to correctly perceive the 3D effect.

Future research work will be focused on the analysis of more impairment patterns and new degradations, in addition to more sequences, to obtain more solid conclusions about the effects of transmission errors and the performance of the methodology. Finally, the improvement of the evaluation methodology will be studied in order facilitate to the observers the way to provide their opinions regarding, not only video quality, but also other factors related to 3D content, such as visual comfort and depth perception.

7. ACKNOWLEDGMENT

8. REFERENCES

[1] S. Chikkerur, V. Sundaram, M. Reisslein, and L. Karam, "Objective video quality assessment methods: Performance comparison", *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165-182, Jun. 2011.

[2] S. Hemami and A. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design", *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 469-481, Aug. 2010.

[3] P. Pérez, J. Gutiérrez, J. Ruiz, and N. García, "Qualitative monitoring of video quality of experience", *Proc. IEEE International Symposium on Multimedia*, Dana Point, USA, pp. 470-475, Dec. 2011.

[4] W. Chen, J. Fournier, M. Barkowsky, and P. Le Callet, "New requirements of subjective video quality assessment methodologies for 3DTV", *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, USA, Jan. 2010.

[5] J. Gutiérrez, P. Pérez, F. Jaureguizar, J. Cabrera, and N. García, "Subjective assessment of the impact of transmission errors in 3DTV compared to HDTV", *Proc. 3DTV Conference*, Antalya, Turkey, pp. 1-4, May 2011.

[6] ITU-R, "Methodology for the subjective assessment of the quality of television pictures", Rec. BT.500-12, Sep. 2009.

[7] V. Baroncini, J. R. Ohm, and G. Sullivan, "Report of subjective test results of responses to the joint call for proposals on video coding technology for High Efficiency Video Coding (HEVC)", Doc. JCTVC-A204, Apr. 2010.

[8] ITU-T, "Subjective video quality assessment methods for multimedia applications", Recommendation P.910, Apr. 2008.

[9] M. Barkowsky, M. H. Pinson, R. Pépion, and P. L. Callet, "Analysis of freely available subjective dataset for HDTV including coding and transmission distortions", *Int. Workshop on Video Processing and Quality Metrics for Consumer Electronics*, Scottsdale, USA, Jan. 2010.

[10] G. W. Cermak, "Consumer opinions about frequency of artifacts in digital video", *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 336-343, Apr. 2009.

[11] N. Staelens et al., "Assessing quality of experience of IPTV and video on demand services in real-life environments", *IEEE Transactions on Broadcasting*, vol. 56, no. 4, pp. 458-466, Dec. 2010.

[12] ITU-R, "Subjective assessment of stereoscopic television pictures", Recommendation BT.1438, 2000.

[13] VQEG, 3DTV project, http://www.its.bldrdoc.gov/vqeg/

[14] ISO/IEC JTC1/SC29/WG11, "Report of subjective test results from the call for proposals on 3D video coding technology", MPEG2011/N12347, Nov. 2011.

[15] M. Barkowsky, K. Wang, R. Cousseau, K. Brunnström, R. Olsson, and P. Le Callet, "Subjective quality assessment of error concealment strategies for 3DTV in the presence of asymmetric transmission errors", *Proc. International Packet Video Workshop*, Hong Kong, China, Dec. 2010.

[16] G. Saygili and C. Göktu, "Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3D video streaming", *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 593-601, Jun. 2011.

[17] H. Coolican, *Research methods and statistics in psychology*, 2nd ed. Hodder & Stoughton, 1994.