TOWARDS CERTIFICATION OF 3D VIDEO QUALITY ASSESSMENT

Andrew Perkis[1], Junyong You[1], Liyuan Xing[1]
Touradj Ebrahimi[2], Francesca De Simone[2], Martin Rerabek[2],
Panos Nasiopoulos[3], Zicong Mai[3], Mahsa T. Pourazad[3]
Kjell Brunnström[4], Kun Wang[4], Börje Andrén[4]

1. Norwegian University of Science and Technology (NTNU), Trondheim, Norway
2. École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
3. University of British Columbia (UBC), Vancouver, Canada
4. Acreo AB, Kista, Sweden

ABSTRACT

Subjective quality assessment is widely used to understand and to study human perception of multimedia quality and as a basis for developing objective metrics to automatically predict the quality of audiovisual presentations. There are several recognized international protocols and procedures for reliable assessment of quality in multimedia systems and services, with emphasis on speech, audio and video modalities. However, the aspect of certification is not yet well understood in this context. This paper discusses various issues regarding certification of multimedia quality assessment. To be concrete, the discussion is illustrated by the procedure implemented to assess 3D video compression technologies within the MPEG effort for the definition of a 3D video coding standard. Selected results from four laboratories, Acreo, EPFL, NTNU and UBC, which participated in the assessment are presented. This case study is used in an early attempt to define a process for certification of subjective test campaigns, based on a cross-validation of the test results across different laboratories, towards the ultimate goal of Quality of Experience (QoE) certification.

1. INTRODUCTION

With the rapid growth of three-dimensional (3D) video technology, standardized compression algorithms for 3D video are needed. In 2011, MPEG committee issued a Call for Proposal (CfP) on 3D video coding technology with the objective to “define a data format and associated compression technology to enable the high-quality reconstruction of synthesized views for 3D displays” [1]. Both stereoscopic and auto-stereoscopic multi-view display technologies were targeted. Responding to this call, 22 proponents submitted their 3D video coding algorithms for competition. In order to analyze and to compare the performance of the proposed technologies a formal subjective quality evaluation was carried out, and a set of test video sequences, encoded with the proposed technologies, was produced. The European COST Action QUALINET (European Network on Quality of Experience in Multimedia Systems and Services) was invited by MPEG to take part in the evaluation campaign of this test material, referred to as 3DV tests in the rest of this paper.

We believe a critical issue in any subjective evaluation is the establishment of a proper certification mechanism to carry out quality evaluations, such as those performed in 3DV tests. Certification usually refers to the confirmation of certain attributes of an object, organization, person, or a process of production [2]. For example, the ISO 9000 family of standards have been designed and issued by ISO to ensure manufacturers can meet certain requirements for the quality of their management. In video quality assessment scenarios, such as in 3DV tests, a standard certificate mechanism also becomes critical. Naturally, a recognized quality assessment experiment conducted by different laboratories, using identical video content and following similar methodologies and instructions, can serve as an appropriate platform for demonstrating the certification procedure of quality assessment. In such a process, cross-laboratory analysis should be performed to find out whether or not consistent results can be obtained. To this end, four laboratories participating in 3DV test have made an effort to illustrate steps towards certification mechanisms. A cross-laboratory analysis has been performed to estimate the correlation of quality scores obtained by each laboratory and to perform a significance test. These analyses show that laboratories employing different subjects could still produce highly correlated results, as they follow similar guidelines to carry out assessments. This confirms that the participating laboratories fulfill an essential condition towards their certification.

The remainder of the paper is organized as follows. Section 2 discusses issues and important steps towards a formal certification procedure of video quality assessment. Section 3 introduces the MPEG 3DV quality test. Test results obtained in the four laboratories and relevant analyses are presented in Section 4. Finally, concluding remarks and discussions on future work are provided in Section 5.

2. TOWARDS CERTIFICATION PROCEDURE

As introduced in Section 1, certification generally refers to the confirmation of certain characteristics of an object, a person, or an organization. This confirmation is often but not always, provided by some form of external review, academic degree, or assessment. In first-party certification, an individual or organization providing a good or service, offers assurance that it meets certain claims. In second-party certification, an association to which the individual or organization belongs, provides such an assurance. Third-party certification involves an independent assessment declaring that specified requirements pertaining to a product, a person, a process or a management system have been met [2].

As an important step towards QoE certification, the European COST Action QUALINET is making an effort to better understand the concept of certification in QoE assessment scenario. The QUALINET Memorandum of Understanding states: “Observing that there are currently no European networks focusing on the concept of QoE, this certification task also aims at bringing a substantial scientific impact on fragmented efforts carried out in this field, by coordinating the research under the catalytic COST umbrella, and at setting up a European network of experts facilitating transfer of technology and know-how to industry, coordination in standardization, and certification of products and services.” [5, 6]

The general objective of certification is to assess and to guarantee uniformity of devices, processes or installations, thus allowing improving interoperability and checking quality targets. In this context, a certification process may target either products, such as laboratories and codec, or services such as quality assessment, or even some content.

Within QUALINET, the certification process may follow one or more of the following main approaches:

- Certification entity based: This approach is more centralized, as certification entities should also be certified themselves. Such an approach may, in principle, be more credible and reliable, but requires a well defined process to certify the certification entities.
- Auto-certification: This approach is not centralized. Each organization may certify itself its systems or services, based on a specific and agreed procedure. At the expense of a lighter process, the reliability and credibility of such an approach may be less certain.
- Peer-certification: This approach is not centralized neither, but requires other entities (not necessarily certified themselves) to provide confirmation or opinions about the degree of fulfillment of identified requirements by a system or service. The use of social networks and open peer review mechanisms, if properly implemented, can contribute to increase the reliability and credibility of such an approach.

In addition to the above, certification process may also involve the following elements and entities:

- Certification applicant: an institution making a certification request (to a certification entity) for a product, service or content.
- Certification entity: an institution or individual, which has either undergone a process to become an authority, or simply acts as an open peer reviewer.
- Certificate or label: a diploma or label that provides a proof to applicant, endorsing the product or service provided fulfills a well-defined set of requirements, along with other information such as the underlying conditions, including duration of the certificate.

Depending on the types of certifications to be addressed, adequate certification methodologies can be defined involving the following main steps:

- Certification request: the applicant should present to the certification entity a request, indicating the type of certification requested and providing all the information and elements necessary for this task. This information and elements will be defined in details in a procedure designed for each type of certification, e.g., laboratory facilities certification or content certification.
- Certification assessment: the certification entity will perform the appropriate steps defined, certifying or not the relevant product, service or content.
- Certificate or label: in the case of positive outcome, the applicant will be provided with a proof stating that it may use a label for its products, services or

contents, along with additional information and conditions such as the duration of the certificate.

Currently, several accreditation companies professionally perform ISO/IEC 17025 certifications by using specified procedures [7] and forms [8, 9]. If such a procedure is considered, the existing ISO/IEC 17025 standard needs to be taken into account. Since the main goal of certifications consists of guaranteeing that certain standards are met, the issue of providing a legal liability or privacy protection mechanisms might also need to be taken into consideration. More importantly, quality assessment results across different laboratories will play a kernel role in a certification procedure, e.g., certified laboratories should produce correlative results when conducting similar quality evaluations. The next sections of this paper concentrate on this last issue as a first and key step towards certification in quality assessment, with emphasis on a use case in 3D video quality evaluation.

3. 3DV TESTS – BACKGROUND, METHODOLOGY AND LABORATORY SET UP

A) Test material

The 3DV CfP defines some "classes" of test sequences, i.e. sets of spatio-temporal resolutions, and some target coding bit rates. Among them, Class A, with frame size of 1920×1088 pixels and a frame rate of 25 frames per second, and Class B, with frame size of 1024×768 pixels and a frame rate of 30 frames per second, along with four target coding bit rates, were used for the evaluation of the proponent technologies. For both stereoscopic and auto-stereoscopic codec comparisons, the test materials included four different contents in Class A (Poznan Street, Poznan Hall2, Undo Dancer, GT Fly) and four different contents in Class B (Kendo, Balloons, Lovebird1, and Newspaper). All test materials were progressively scanned and used 4:2:0 color sampling with 8 bits precision per pixel.

The video data evaluated in the subjective tests were generated from a dense set of synthesized views provided by proponents and fed uncompressed into 3D monitors thanks to a specially designed video server configuration. Particularly, two test scenarios, namely, a 2-view input configuration, to be evaluated on stereoscopic display, and a 3-view input configuration, to be evaluated on both auto-stereoscopic as well as stereoscopic display, were considered. The depth data and camera parameters for view synthesis and rendering were also provided. Readers can refer to the 3DV CfP for more details [1].

B) Proponents

By responding to the CfP, 22 proponents submitted their codec descriptions, and encoded and decoded test sequences at requested target bit rates. Two anchors, i.e., H.264/AVC and HEVC, were also included in the set of coding technologies under assessment.

C) Laboratories, hardware, software, and instrumentation set up

The 3DV tests involved 12 evaluation laboratories from around the world. Each laboratory was assigned a certain number of test sessions, either stereoscopic, auto-stereoscopic, or both, based on the availability of hardware and other facilities. All laboratories used the exact same evaluation methodology, described below, including the same monitors (a 46" Hyundai S465D polarized stereoscopic monitor and a 52" Dimenco BDL5231V auto-stereoscopic monitor, with native resolutions of 1920x1080 pixels), the same implementation of Graphical User Interface (GUI), and similar test room configuration.

The hardware and software environments used in all laboratories were designed and tested to ensure meeting well specified requirements by conducting dry runs before actual evaluations took place.

Eighteen naive viewers evaluated the quality of each test sequence. Since a maximum of 3 (5) subjects could be seated in front of a stereoscopic (auto-stereoscopic) monitor, without deteriorating the perception of the 3D rendering, several subjects could be grouped to attend a same test session. Hence, the test room set up, common in all the laboratories, included 3 to 5 subjects seating in a row, perpendicular to the center of the monitor, for the auto-stereoscopic and the stereoscopic viewings, respectively. The viewers were seated at roughly 3.5 meters from the auto-stereoscopic monitor, as required in [1], and at roughly 2.3 meters from the stereoscopic monitor, as suggested in the ITU-R BT.710 recommendation for HDTV [3]. The laboratory setup was controlled in order to ensure the reproducibility of results by avoiding as much as possible, involuntary influence of external factors. The test rooms were equipped with a controlled lighting system with a 6500K color temperature and an ambient luminance at 15% of maximum screen luminance. Each laboratory reported the details of the calibration settings used for each monitor, as well as the gender percentages and average age of their sample of viewers, and the exact number of subjects per session.

D) Test data rendering

In order to render correctly the test materials on the stereoscopic and auto-stereoscopic monitors, the following processing was performed on the raw video files received from proponents. For the auto-stereoscopic display, 28 yuv files, each containing a different view for the same video sequence, were interleaved and merged into a single avi file, using an interleaving software tool provided by Dimenco. For the stereoscopic viewing, two pre-selected yuv video sequences, corresponding to the left and right views, were cropped and horizontally shifted in order to obtain a pre-defined depth for each content (different shift parameters were set for different content) and finally interlaced (right view on top) and padded to produce a full HD resolution video.

E) Evaluation methodology: Stimulus presentation and rating scale

The Double Stimulus Impairment Scale (DSIS) evaluation methodology was selected to perform the tests. Subjects were presented with pairs of video sequences (i.e., stimuli), where the first was always an unimpaired, reference, video (stimulus A) and the second, the same content processed (stimulus B). Subjects were asked to rate the quality of each stimulus B, keeping in mind that of stimulus A. A dedicated GUI was developed for the test campaign: before each video sequence, a grey screen with the letter “A” (“B”) was shown for two seconds, informing subjects that the reference (test) stimulus would be shown. After the presentation of each pair of sequences, a grey screen with the message “Vote” was shown for five seconds. The test subjects were asked to enter their quality score for the stimulus B in paper scoring sheets during these five seconds.

An 11-grade numerical categorical scale was used [4]. The rating scale ranged from 0 to 10, with 10 indicating the highest quality, i.e., the test sequence is indistinguishable from the reference, and 0 indicating the lowest quality.

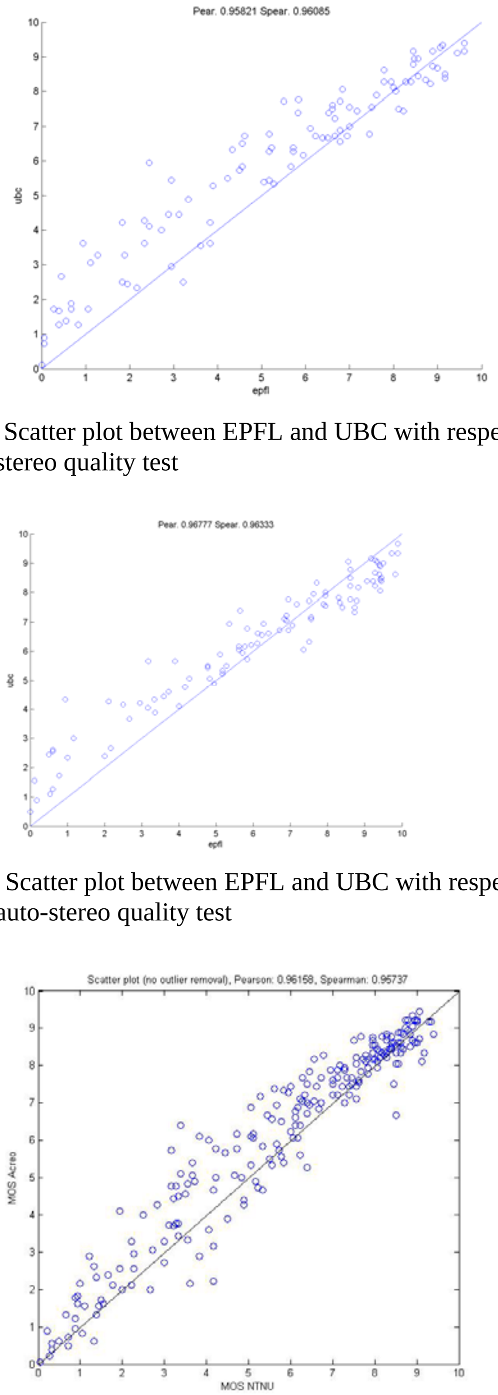
F) Screening

All subjects taking part in the evaluations underwent a screening to examine their visual acuity, using the Snellen chart, and color vision, using the Ishihara chart. Their stereo vision was also tested using the Randot test.

G) Training

Before each test session, written instructions and a short explanation by a test operator were provided to the subjects. Also, a training session was run to show the GUI, the rating sheets, and examples of processed video sequences. The training video sequences were produced using two different contents (Pantomime and Champagne) and with coding conditions similar to those used to produce the actual test materials.

It is important to stress that the same training instructions were provided to subjects in all the laboratories.



(a) Scatter plot between EPFL and UBC with respect to stereo quality test

(b) Scatter plot between EPFL and UBC with respect to auto-stereo quality test

(c)) Scatter plot between NTNU and Acreo with respect to stereo quality test

Figure 1. Scatter plots of 3DV quality test among four laboratories.

Particularly, during training, specific scores were given to the training sequences. These scores had been agreed upon across the evaluation laboratories in order to ensure close correlation and consistency of results.

H) Test sessions

A basic test session of DSIS methodology including 24 test pairs, three dummy stimuli pairs, and one reference versus reference pair, was designed. Thus, the test materials resulted in a total of: 16 sessions for each of the two classes of auto-stereoscopic data, 16 sessions for each of the two classes of 2-view stereoscopic data, 16 sessions for the Class A 3-view stereoscopic data, and 32 sessions for the Class B 3-view stereoscopic data. In each session, the stimulus pairs were presented in random orders, but never with the same video content in consecutive pairs.

4. SELECTED TEST RESULTS AND ANALYSIS

Some test sessions were performed by more than one laboratory in order to analyze inter-laboratory cross-correlations. In this section, we report the results of the test sessions performed by four laboratories, namely NTNU, Acreo, EPFL, and UBC, including some cross-validation analysis for the common sessions.

Different overlapping data within each laboratory groups were used. These included:

- Class A, 2-view stereo, 4 sessions (EPFL - UBC)
- Class A, auto stereo, 4 sessions (EPFL - UBC)
- Class B, 2-view stereo, 8 sessions (NTNU - Acreo)

Thus, cross-laboratory results analysis could be performed between EPFL and UBC, and between NTNU and Acreo.

Figure 1 shows the pairwise scatter plots and correlation coefficients on the overlapping data. It can be observed that the subjective quality results uniformly span over the entire range of quality levels from 0 to 10, which can be considered as an indication of appropriate experiment design and their implementation. More importantly, there exists a high correlation between different laboratories. The Pearson linear coefficient measures the distribution of the points around the linear trend, while the Spearman coefficient measures the monotonicity of the quality scores between different laboratories, that is, how well an arbitrary monotonic function describes the relationship between two sets of data. The results show that the data from NTNU and Acreo, as well as EPFL and UBC, are highly correlated.

Additionally, an ANOVA analysis, where two laboratories were considered as between group variables and the different Processed Video Sequence (PVS) as within group variables, was performed on the raw data, yielding a significant main effect of the “laboratory” variable on the results of the two pairs of laboratories (for instance, on the NTNU-Acreo data: F(1, 34) = 5.6, $p = 0.02 < 0.05$). One could also observe an expected significant effect of the PVS (for instance on NTNU-Acreo data: F(223, 7582) = 112.6, $p = 0.00 < 0.05$), as well as, a significant interaction between the PVS and laboratories, (For instance on NTNU-Acreo data: F(223, 7582) = 2.2, $p = 0.00 < 0.05$). Considering the data from NTNU-Acreo, a linear transformation:

$$y = 0.9375 \cdot x + 0.7423 \qquad (1)$$

(Figure 2) will make the significant main effect of laboratories disappear (F(1, 34) = 0.00, $p = 1.0 > 0.05$), but the significant main effect of PVS (F(223, 7582) = 111.7, $p = 0.00 < 0.05$) and interaction between PVS and laboratories remain (F(223, 7582) = 2.2, $p = 0.00 < 0.05$).

In addition to the above observations, a Student t-test was applied to each pair of PVS from the different laboratories, identifying the significant different PVS, shown in Figure 3 for the NTNU-Acreo comparison. It can be observed that the significant different PVSs are spread quite evenly over the entire quality range. However, when compared to NTNU, the subjects at Acreo gave significantly higher scores for the mid range qualities, and clearly lower scores at the lower end of the scale.

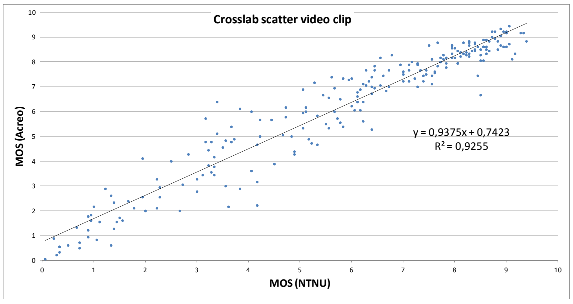


Figure 2. Scatter plot between NTNU and Acreo with estimated regression line.

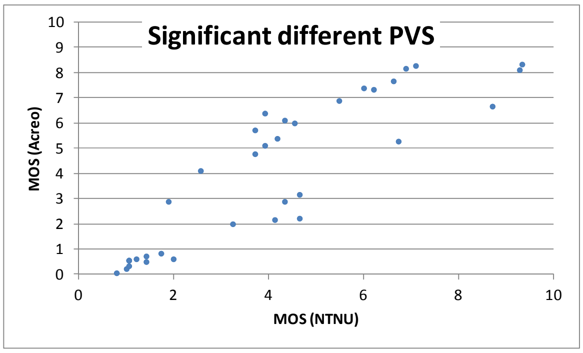
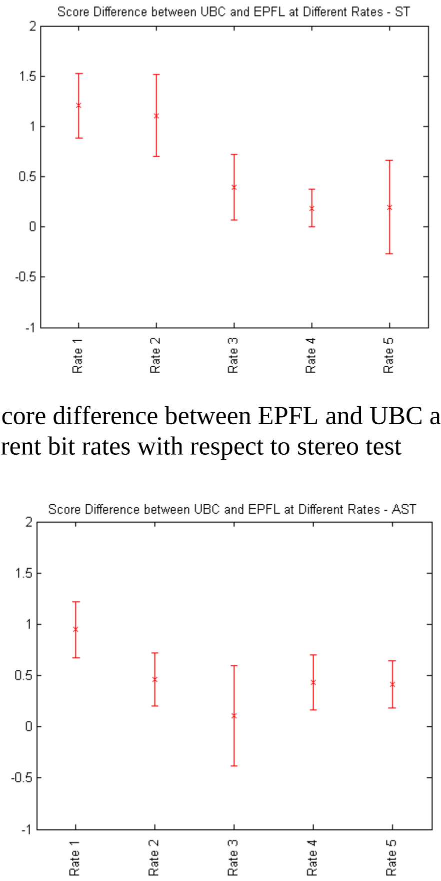


Figure 3. Scatter plot between NTNU and Acreo of significantly different PVSs.

Finally, Figure 4 compares the score difference between UBC and EPFL at different bit rate levels. Note that in this figure the video bit rate increases from “Rate 1” to “Rate 5”. As it is observed, for both the stereo and auto-stereoscopic cases, the score differences are higher at lower bit rates (i.e., low video quality), indicating that subjects had difficulties to precisely quantify low quality content.



(a) Score difference between EPFL and UBC at different bit rates with respect to stereo test

(b) Score difference between EPFL and UBC at different bit rates with respect to autostereo test

Figure 4. Score differences between EPFL and UBC at different bit rates.

This additional comparative analysis indicates that although correlations between laboratories are high and exhibit good correspondence, more complex differences exist in the voting patterns that cannot be modeled by simple transformations, like for instance the linear transformation in Figure 2.

5. CONCLUSIONS AND FURTHER WORK

This paper presented cross validation analysis from a recent MPEG 3DV quality test campaign conducted with the help of a European COST Action QUALINET. The test results, obtained from four laboratories in Europe and North America, participating in this test campaign, have been presented and analyzed. Various analyses demonstrated that different laboratories can produce similar quality assessment results when they follow appropriately selected evaluation procedures. The quality test across different laboratories with the identical video contents provides an appropriate first step for laboratory certification purpose. An effort is under way by COST Action QUALINET towards a better understanding of certification mechanism of QoE in multimedia services and systems. This could lead to the definition of a roadmap, which could hopefully help in the implementation of appropriate certification mechanisms in QoE for multimedia applications.

ACKNOWLEDGEMENT

The authors would like to acknowledge the efforts by all those involved in the MPEG 3DV tests, including the MPEG test coordinator, all laboratories which participated in the evaluations, and proponents responding to MPEG 3DV CfP. This work was performed in the framework of the COST Action IC1003, QUALINET.

6. REFERENCES

[1] ISO/IEC JTC1/SC29/WG11, “Call for proposal on 3D video coding technology,” Doc. N12036, Geneva, Switzerland, March 2011.
[2] Wikipedia for ‘certification’, May 2011.
[3] ITU-R BT. 710-4, “Subjective assessment methods for image quality in high-definition television,” International Telecommunication Union, November 1998.
[4] ITU-R BT. 510-11, “Methodology for the subjective assessment of the quality of television pictures,” International Telecommunication Union, 2002.
[5] QUALINET Memorandum of Understanding, May 2010.
[6] F. Pereira and S. Buchinger, “First thoughts on QUALINET certification,” QUALINET, July 2011.
[7] General requirements: Accreditation of ISO/IEC 17025 Laboratories, A2LA, August 2010.
[8] ISO/IEC 17025:2005 Working document, Perry Johnson Laboratory Accreditation, Inc., 2007.
[9] ISO 17025 quality forms, http://www.17025.com/quality_records.html, May, 2011.