Question(s):	**Meeting, date:**	VQEG meeting, Hillsboro, 12-16 Dec 2011
Study Group:	**Working Party:**	**Intended type of document** (R-C-TD): C
Source:	Technicolor	
Title:	Work on subjective testing methodologies for quality assessment of stereoscopic video content	

Contact:	Quan Huynh-Thu Technicolor France	Tel: +33 (0)2 99 27 90 45 Fax: +33 (0)2 99 27 30 15 Email: quan.huynh-thu@technicolor.com

Please don't change the structure of this table, just insert the necessary information.

Summary

International Recommendations for the subjective evaluation of 2D video quality exist (e.g. ITU-R BT.500, ITU-T P.910). These Recommendations specify both the test environment (i.e., lighting levels, room set-up) and the testing methodologies (i.e., presentation patterns and rating scales). It is envisaged that some or all of the existing methodologies for 2D could be applied to the quality evaluation of stereoscopic video with some appropriate modifications. Investigation is needed to identify which particular points in the methodologies need to be modified and how they need to be modified.

The following contribution identifies a list of points of investigation to define appropriate subjective testing methodologies for the quality assessment of stereoscopic video content.

1 Introduction

Amongst 3D video quality studies published recently in the literature, most of them have applied 2D methodologies as is. Most of these studies were conducted with the goal of building a dataset to develop or benchmark an objective metric. Almost none of these studies have investigated the need to adapt the 2D methodologies. None of these studies have provided experimental data nor pointed to existing literature showing that the used methodology has been shown to provide repeatable and reliable results.

Furthermore, most of these works focus solely on the visual quality. Very little details are given concerning the test environment, lab set-up, instructions to participants, practice session. These elements can influence the subjective results, in particular the instructions and practice session as subjects are not so familiar or experienced with viewing of 3D content. In addition, viewing time may influence results as judgment of quality may be impacted by visual fatigue.

Lab set-up may need to be adapted to avoid depth perception conflicts. For example, in a recent study [1], the display was positioned far enough from the wall to avoid any conflicts of the displayed 3D content with the real world. The room illumination was adjusted in such a way that, through the activated shutter-glasses, the luminance reflected from the background was 15% of the display's peak luminance (this corresponded to 50 cd/m2 without glasses). No flickering of the background light was perceived. In a similar study [2], the same subjective experiment using stereoscopic videos was conducted by two different test labs and cross-lab comparison was performed. The subjective quality scores had the same trend between the labs with a very high linear correlation, although some differences were found. In most cases, the observers in Lab1 provided higher scores than those in Lab2 for the same test sequences. An ANOVA revealed a significant difference in the main effect of the laboratories. However, a linear transformation of the

scores of one lab could be applied to remove this inter-lab difference. In both [1] and [2], the reference 2D condition obtained a similar or even higher subjective quality than the reference 3D condition. This surprising result may be due to the single-stimulus presentation, in which viewers did not perceive a real difference in quality between 2D and 3D. However, different conclusions may be obtained using a double-stimulus (e.g., DSIS or DSCQS) or multi-stimulus presentation (SAMVIQ).

The 3D display itself has a large impact on the stability and reproducibility of the subjective experiment. 3D display technology is still improving; different technologies exist and, at this point, none can be recommended as a reference display. The viewing angle, the field of view, the amount of crosstalk and the brightness are often limiting factors. In particular, crosstalk can produce ghosting artifacts and in the worst case can prevent proper binocular fusion of the stereo-images. Special attention is required on the way the display itself processes internally the 3D content. Recent 3DTV displays apply internally some post-processing for crosstalk reduction or a format conversion takes place, and rendering artifacts may easily outweigh the added value of depth [3].

Two studies have reported a slight adaptation of an existing 2D methodology. In [1], an experiment was conducted to study the subjective quality of 3DTV in presence of transmission errors and examine the impact of related error concealment strategies. The ACR method was used for the experiment. However, the training session before the actual test used the DSCQS method to allow participants to get accustomed to the types of stimuli, their characteristics, and the range of distortion that could be expected in the experiment. During the experiment, participants were asked to provide an opinion of quality of experience on the 5-point discrete scale and were also provided the possibility to indicate any experienced visual discomfort using a check box. However, providing this information was not mandatory. In [2], an experiment was conducted in an IPTV scenario using test set-up was the same than in [1]. Both training and test sessions used the ACR method for presentation and voting. Additionally, viewers were asked to provide a judgment of quality of experience on the 5-point rating scale, as well as a judgment of visual comfort on a 5-point scale making a comparison to visual comfort experienced in 2DTV (much more comfortable, more comfortable, as comfortable, less comfortable, much less comfortable than watching 2DTV). Both attributes had to be rated at once (for the same stimulus presentation).

A lack of studies on the reliability of the subjective data clearly exists in the literature. Studies that can provide insight on the effects of the testing protocol/methodology are needed and will help to identify a suitable protocol to assess various aspects of 3D QoE.

2 Points of investigation

The points of investigation that are specific to subjective quality assessment of stereoscopic 3D video include:

- Repeatability of a given test methodology:
    - This is a very crucial point for any subjective testing methodology. Empirical data are needed to prove that a given methodology can produce repeatable and reproducible data.
    - Repetition of the same experiment (same test set with same methodology) can provide such empirical evidence
- Ability to separately assess the different basic perceptual attributes related to 3D quality (picture quality, viewing comfort and depth quality). An analogy can be made to audio-visual quality where cross-modal interaction between audio and video has been documented. In the same way, the question is whether subjects are able to assess independently visual

quality, depth quality and visual comfort. If not, then is it relevant to ask them to judge these separate attributes? See also the point on "role of instructions".

- Necessity to use anchors (2D and 3D anchors) in the test stimuli:
    - The potential of 3D lies in the increased quality of experience compared to 2D. Viewers will only embrace 3D if it provides a better viewing experience than 2D. The underlying question is whether or not subjects are more able to judge 3D quality if they are asked to compare it to 2D, instead of simply judging a 3D stimulus on its own (or even in comparison to some 3D reference).
    - With the hypothesis that subjects know more easily how to judge a 2D video stimulus, one adaptation of the 2D methodologies could make explicit reference to a 2D version of the stimulus. Explicit comparison can be made in the stimulus presentation and/or in the rating scale
- Viewing conditions (e.g., viewing angle):
    - Currently 3 simultaneous viewers are allowed in front of a 2D HDTV screen in a subjective test. Because of the increase of crosstalk with viewing angle (angular position), this number may need modification (e.g., is a maximum of 1 or 2 viewers a more appropriate number for 3D tests?)
- Display characteristics:
    - What is the influence of stereoscopic display characteristics (mainly crosstalk level/characteristics) on quality judgment
    - Method to characterize and select a stereoscopic display for conducting subjective experiments (e.g., maximum crosstalk =< crosstalk threshold)
- Sequence duration:
    - Short (10-sec) videos have been traditionally used in 2D video subjective testing with overall rating to avoid problems of recency effects. Literature has shown that subjects can confidently provide a judgment of image quality for this range of duration.
    - The underlying question is whether or not such a short video duration is suitable to assess visual comfort and depth quality. Some works, without providing empirical data but only survey, have suggested that longer duration may be needed.
    - Alternatively to the use of longer duration, test designs using stimulus repetition may provide a different path of investigation.
- Role of instructions and more elaborated practice session: These two points may need more emphasis in case of 3D than in 2D.
    - Most subjects are not well experienced with viewing of 3D content. Most of them have viewed maybe a few 3D movies but experience is far from comparable to exposure to 2DTV. As a consequence, subjects may not well understand how they should judge the 3 basic perceptual attributes for two reasons:
        - Firstly, they may not well understand the meaning of the attribute to judge.
        - Secondly, they may not know if they need to consider this attribute alone or not. For example, in judging visual quality, should the perception of depth (depth quality) be taken into account? Should visual comfort be taken into account?
    - Clear definition of depth quality and visual comfort:
        - Depth quality: from experience, this is usually the most difficult attribute to be judged. As viewers are not so experienced with viewing of 3D content, they usually find it difficult to know how to provide a judgment.
        - Visual comfort: although there is a natural sense in knowing what is and is not comfortable viewing, precise description of symptoms may be necessary.
- Use of additional questionnaires (besides the quality rating):

- Use of ad-hoc additional questionnaires (similar to simulator sickness questionnaire) should also be investigated to gain more understanding in how people judge 3D and react to it.
- Which questions are relevant in which context? When should these questions be asked?

3 Proposal

Priority needs to be identified and collaborative work initiated between 2 or more labs for each of the points of investigation listed above. Although it is recognized that there may be interaction between the points investigated, the only practical approach so far is to address each of them separately. Experimental results will provide empirical evidence for the need to modify/adapt existing methodologies.

References

[1] 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," in Proc. Packet Video Workshop, Hong Kong, December 2010.

[2] K. Wang, M. Barkowsky, R. Cousseau, K. Brunnström, R. Olsson, P. L. Callet, and M.Sjöström, "Subjective evaluation of HTDV stereoscopic videos in IPTV scenarios using absolute category rating," in Proc. SPIE Conf. Stereoscopic Displays and Applications XXII, vol. 7863, San Francisco, January 2011.

[3] M. Barkowsky, R. Cousseau, and P. Le Callet, "Influence of depth rendering on the quality of experience for an autostereoscopic display," in Proc. Int. Workshop on Quality of Multimedia Experience (QoMEx), San Diego, July 2009