

"QUEST FOR FINDING THE RIGHT HD FORMAT"

A NEW PSYCHOPHYSICAL METHOD

FOR SUBJECTIVE HDTV ASSESSMENT

H. Hoffmann¹, T.Itagaki², D.Wood³

ABSTRACT

The availability of different HDTV image formats such as 720p/50 or 1080i/25 and 1080p/50 as a potential future format places the question for many users on what HDTV format to use with which compression algorithm and at what bit rate. Subjective test according to the well know standards of the ITU-R BT.500-11 only provide the failure characteristic of each individual HDTV format. The EBU has developed a new subjective testing method called the 'Triple Stimulus Continuous Evaluation Scale' (TSCES) for HDTV formats and has used this method with over 300 assessors to investigate the different HDTV formats on different flat panel displays. H.264/AVC was used as compression system at broadcast bit rates. Thus, comparative graphs between the different HDTV formats were possible and have shown that progressive scanning formats outperform HDTV with 1080i/25. Although the results have also proven that currently 720p/50 is the favourable emission format, there are also indicators for the potential of 1080p/50.

INTRODUCTION

Many broadcasters who are interested in introducing HDTV will be sooner or later faced with the question on what HDTV format and what bit rate to use for their HDTV services. Perhaps, they will conduct subjective test according to the well know standards of the ITU-R BT.500-11, but what they gain from these test is only the failure characteristic of each individual HDTV format - be it 1080i/25, 720p/50 or even 1080p/50 as a future candidate. A comparative graph can not be generated, because the current methods do not allow changing the reference against which the impaired images are evaluated by assessors.

The EBU has been faced with the same problem with its first subjective tests. The testing method was the 'Double Stimulus Impairment Scale' (DSIS) method according to ITU-R BT.500-11 [5] and although some indicative results have pointed towards the advantages of progressive scanning no formal answer or comparative analysis between the different HDTV formats was possible. Some

¹ European Broadcasting Union, Switzerland

² Brunel University, UK

³ European Broadcasting Union, Switzerland

further shortcomings of the current DSIS ITU-R BT.500-11 method was that it utilizes scales and adjectives to describe the impairments, which have to be translated into the languages in which the tests are being done. The adjectives to characterize the image can be understood differently by assessors with different mother tongues. There are variable intervals between the meanings of the descriptive adjectives in the scale in the same language, and variable intervals across different languages. Furthermore, in the existing methods, the reference pictures are displayed on the same screen as the test condition, thus counting on the memory of the assessors. In variants of the ITU-R BT.500-11 methods two screens can be used, one as the reference and one as the test condition, or even a split screen can be created, but this would still not allow the evaluation of three different HDTV formats. The DSIS method with one screen was used in a first series of tests of the three HDTV formats [2]. The conclusion of these tests was that no clear answer could be given to the question of which HDTV format would provide a better image quality at what bit-rate. The only safe result of these tests was to report the failure characteristics of each individual HDTV format (i.e. 1080i/25 uncompressed reference against 1080i/25 compressed images).

The new method addresses these shortcomings and details on the new method are given in [3, 4]. It is dedicated here (but not limited) to the HDTV picture quality comparison on large Flat Panel Displays (FPD) and is called the 'Triple Stimulus Continuous Evaluation Scale' method (TSCES).

Requirements on the new subjective HDTV testing method

- to allow the direct comparison of different HDTV scanning formats with reporting in one resulting graph,
- easy to use by non-expert assessors. Non experts are used as an average opinion of the public at large is sought rather than that of experts,
- to provide reliable and reproducible results, with a standard deviation only determined by the natural spread of opinion, and the stability of the results as constant as possible across the range of qualities being evaluated,
- independence of language adjectives describing the perceived image quality, and should have independence of scale interval linearity,
- can cope with a wide range of picture qualities and HDTV formats such as 720p/50, 1080i/25, 1080p/50, and SDTV,
- can be used to accurately measure a video system's basic quality and failure characteristics (the relationship between quality and the parameters which reduce it),

- can be used with large and medium sized flat panel displays, such as LCD or PDP, as these will be the dominant mode for viewing in the years ahead for conventional television and the coming generations of high definition television.

Design, set-up and testing method

As shown in [3, 4], the new method uses three displays to allow assessors a direct comparison between the different formats. Wooden mock up display frames were built and different placements of displays side-by-side or vertical were tested and discussed with the expert staff of the technical department in the EBU. A further critical point was the viewing environment. The method should accommodate six to eight viewers, depending on the display (viewing angle issue), at a viewing distance of 3h and 4h. A horizontal configuration of three displays, side by side, would have had the disadvantage that none of the viewers could get an optimum view of all three displays in terms of distance, viewing angle and so forth. For that reason it was decided to use a vertical setup of the displays, where each of the three displays was slightly angled to provide optimum viewing to the centre assessor in front of the screen (see Figure 1).

For HDTV evaluations, the three displays were adjusted at an angle in such a way that a reference viewer at an eye-height of 1.2 m and in a centre position to the screens always had a constant eye-distance of 3 times picture height (3h) to all three displays (Figure 1).

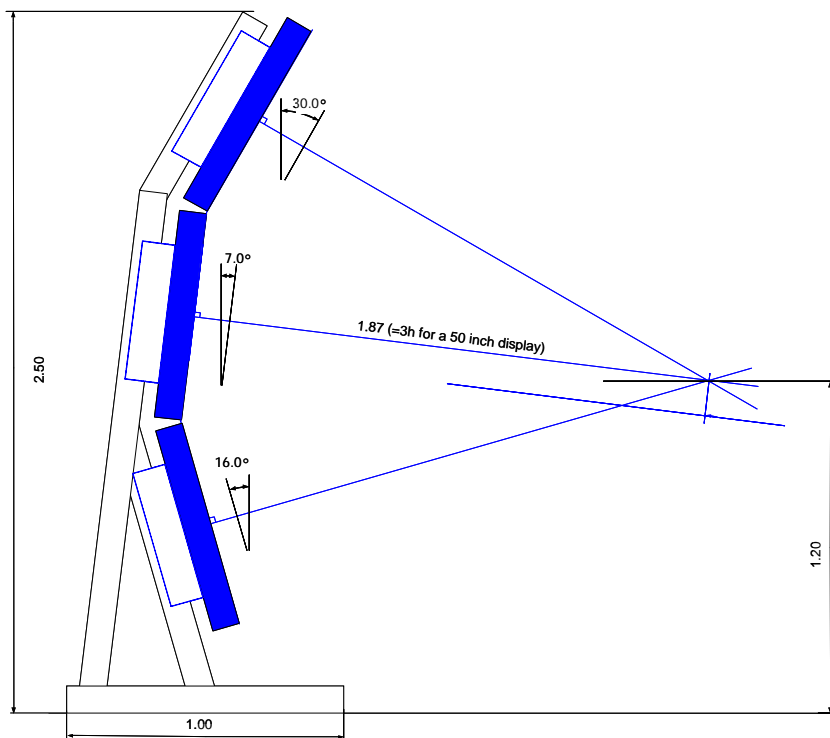


Figure 1 Display rack construction diagram. All units in meters, angles in degrees. Technical drawing and supervision of mechanical construction by Edgar Wilson EBU, Technical Department. The dimensions were designed for max. 52-inch displays.

The design viewing distance for HDTV is three times (3h) the picture height of the display and this is used here for a max. 52-inch display diagonal. Having the monitors mounted above one another, the assessors quickly grasp what needs to be done, and the arrangement suits widescreen display well. Using displays with a comfortable viewing angle will permit more assessors per viewing session.

A further problem to solve was how the content should be presented and evaluated by non expert assessors. Showing the individual HDTV formats on each screen (i.e. 1080p/50 top, 1080i/25 middle, 720p/50 bottom) would have made the voting very difficult and probably would only be sensible for expert viewers. Such a configuration was used at demonstrations at the International Broadcasting Convention 2006 in Amsterdam [1] and it became clear from comments received that formal subjective testing would require a different form of presentation.

The solution for providing simple understanding and easy voting was found in the following: the top display serves as an upper reference, providing a high quality anchor. Normally, humans perceive 'top' to be 'excellent or very good', and the uncompressed 1080p/50 currently represents the best image format. The bottom display serves as the low anchor with a defined impairment, so that the images are perceived as 'bad' images. The middle display shows the pictures under test in different formats and bit rates, preferably including hidden upper and lower anchor content.

The type of impairment used for the bottom anchor must be clearly defined, must be reproducible, and must have similar characteristics to the impairments expected in the middle display to help the orientation of the assessors. Experiments with, for example, adding white noise as a defined impairment factor for the lower anchor (bottom display), found that such impairments are too different to the impairments caused by H.264/AVC coding presented in the middle display. A better solution for creating robust lower anchor impairments was found by using the publicly available and defined H.264/AVC reference encoders, i.e. the same compression system as used for the images under test on the middle display. In addition, to reflect practical situations for consumers today when SDTV is viewed on large FPD, it was decided to apply this H.264/AVC compression to the SDTV image format derived from the 1080i/25 content.

In addition, the following settings need to be applied:

- ITU-R BT.500-11 viewing environment and ambient light conditions,
- All three displays have to show the same scene content synchronised at the same time,
- All three displays need to be of the same type and aligned and should be grade 1 reference type displays (unless particular types of display are being tested). For the experiment reported below high-end consumer displays were used.

In order to meet the objectives of showing all three formats in synchronized form, it was necessary to find technologies that could provide the three signal sources, 1080p/50, 1080i/25 and 720p/50 as well as SDTV synchronized via time-code and in uncompressed form to the displays. The displays had only DVI interfaces thus the signal source also had to provide DVI outputs. As layout servers, three

DVS pronto2k were used and synchronized via RS-422 time code. The servers provided uncompressed DVI and HD-SDI and dual link HD-SDI outputs.

Assessors should be given clear instructions before the tests begin, and a stabilizing sequence of 4 sequences that is to be later ignored in the analysis. Paper sheets with continuous 100 mm lines to make the assessment can be used. A useful future development would be to allow voting on wireless PDAs. The top of the line is explained as representing the quality of the top monitor. The bottom of the line is explained as representing the quality of the bottom monitor. The assessors are asked to mark on the line where the overall quality of the central monitor falls between the top and the bottom. The top and the bottom are thus upper and lower anchors for the evaluations. When processing, results can be mapped onto the 5 grade impairment or quality scales or a 100 point continuous quality scale. An example scale is shown in Figure 2.

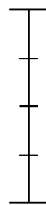


Figure 2 Scale of 100mm length used for voting (here shown in reduced form). The 25%, 50%, 75% lines are only helpers for the assessors. The full scale was available for voting

DSIS allows two repetitions of the sequence before voting on one display. The new method requires more time since assessors have to scan over three displays. It is therefore advised to allow four repetitions of the 10 second test sequences and to include after each sequence a counter that informs assessors of the number of the sequence they are observing (i.e. number embedded in mid grey, see Figure 3).

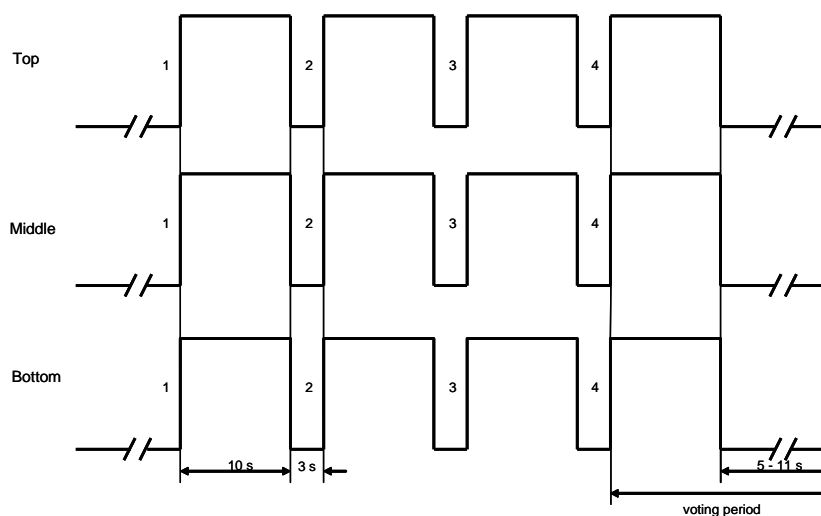


Figure 3 Voting procedure

The maximum length per session should be limited to 20-30 minutes then a break is required for the assessors.

SUBJECTIVE HDTV TEST WITH THE NEW METHOD

The method was tested first in November 2006 with 178 students at the University of Applied Sciences in Wiesbaden-Germany (see Figure 4) and using three 50-inch 1920x1080 pixels PDP. Seven different sequences in the HDTV formats 720p50, 1080i25 and 1080p/50 software coded in MPEG-4 AVC (detailed parameters see [3]) with bitrates from 6 to 18 Mbit/s were used in testing the new method. A second round of test was conducted at the EBU headquarters with about 150 assessors and using three 52 inch LCD with 1920x1080 pixels. This second tests used 5 sequences and were conducted in March 2007.



Figure 4 Photo of the viewing session in Wiesbaden

Contents under test for each display

Upper display high image quality anchor:

- uncompressed HDTV signal with 1080p/50

Middle display with images under test:

- 1080p/50, 1080i/25 and 720p50 HDTV at 6, 8, 10, 13, 16, 18 Mbit/s and 576i/25 SDTV format at 4 Mbit/s and the upper and lower anchor was included as hidden references. For each format the bit-rates per presentation were randomized.

Bottom display low image quality anchor:

- 576i/25 Standard Definition Television (SDTV) format down-converted from the 1080i/25 Content and compressed with H.264/AVC at 3 Mbit/s representing a practical SDTV broadcast condition.

Table 1 shows the 7 different test sequences used in the tests.



	Name	Source Format before down- sampling and origin	Characterization	Image type
1	Crowd Run	2160p/50 SVT Test Set	- Medium critical: No camera movement, but trees and grass and running crowd	
2	Park- Joy	2160p/50 SVT Test Set	- Critical: Camera pan, water, trees and running people	
3	Princes- s-Run	2160p/50 SVT Test Set	- Critical Camera pan, trees, grass and running person	
4	Aloha- Wave	1080p/50 Sony HDC1500	- Medium critical: Soccer stadium, "aloha-wave" in audience	
5	Ice- Dance	1080p/50 Sony HDC1500	- Non-critical: In house shot, white ice-ground with two moving actors plus camera pan; some background with detail structures	
6	Dancer	1080p/50 Sony HDC1500	- Critical: Soccer stadium. Dancing person on grass with lots of reflection in the costume of the person	
7	Police- boat	1080p/50 Sony HDC1500	- Critical: Police boat drifting on water	

Table 1 Summary of test sequences for TSCES with PDP display

Results:

Each vote given on a 100 mm paper scale was measured and edited in Excel for processing. For example a marker at the 100 mm line would have meant that the assessor had the impression that

the picture under test in the middle display had the same quality like the uncompressed upper anchor on the top display, and a marker on the 0 mm point of the scale that the middle picture is as bad as the lower anchor bottom display.

First of all a screening of the votes was performed and those assessors votes who did not identify the upper and lower anchor within 20 % were removed. As a general result we observed that the hidden references (upper and lower anchor) were clearly detected by most of the assessors. Even the slight difference of 3 Mbit/s SDTV to 4 Mbit/s SDTV became clearly visible in the votes. With smaller numbers of assessors (<15) the statistical error increases.

RESULTS OF THE SUBJECTIVE TESTS

In the following, due to length limitations of this paper, only some graphs of the subjective tests are shown. The PowerPoint presentation will include more detailed analysis.

The sequence Crowd Run (Figure 5, Figure 6) has shown that 1080p/50 was preferred until about 10 Mbit/s. Then 720p/50 was more preferred for 8 Mbit/s and 6 Mbit/s. This is due to the fact that compression artefacts masked any resolution advantage of the 1080p/50 format and 720p/50 with less compression artefacts was more preferred. The 1080i/25 format was not appreciated. In addition one can see that the statistical error with fewer assessors as used for the LCD is higher than with the PDP where 29 assessors were used. This suggest that the new subjective test method requires like other subjective testing methods (see ITU-R BT.500-11) a sufficient number of assessors.

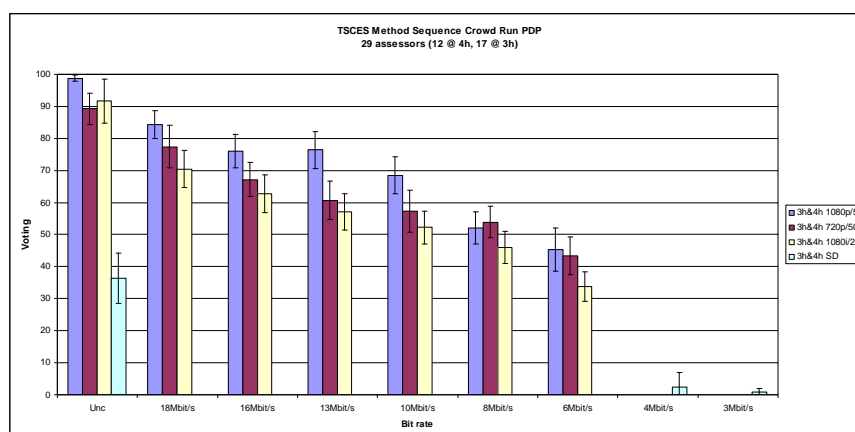


Figure 5 Crowd Run on PDP

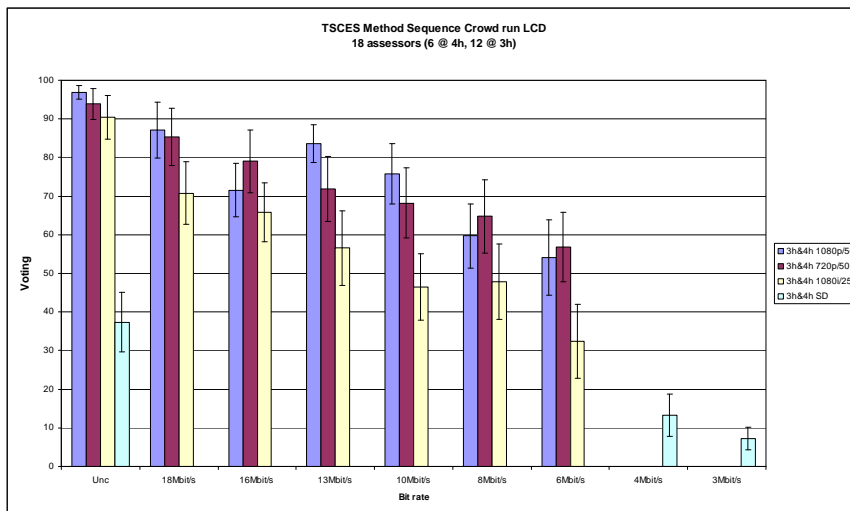


Figure 6 Crowd Run on LCD. Note that the limited number of assessors produced a higher statistical error. See also the issue for 1080p/50 and 16 Mbit/s coding which seemed to be less good rated than the 13 and 10 Mbit/s.

The sequence Park-Joy (Figure 7, Figure 8) has shown that 720p/50 was already the preferred format at bit rates around 18 to 16 Mbit/s. This is due to the image criticality which produces compression artefacts for the 1080p/50 format that are not so apparent for the 720p/50 format. Again 1080i/25 was not appreciated.

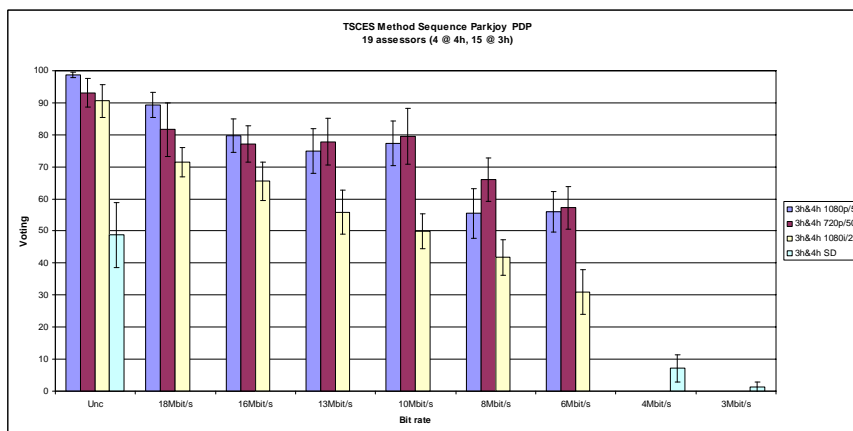


Figure 7 Parkjoy on PDP

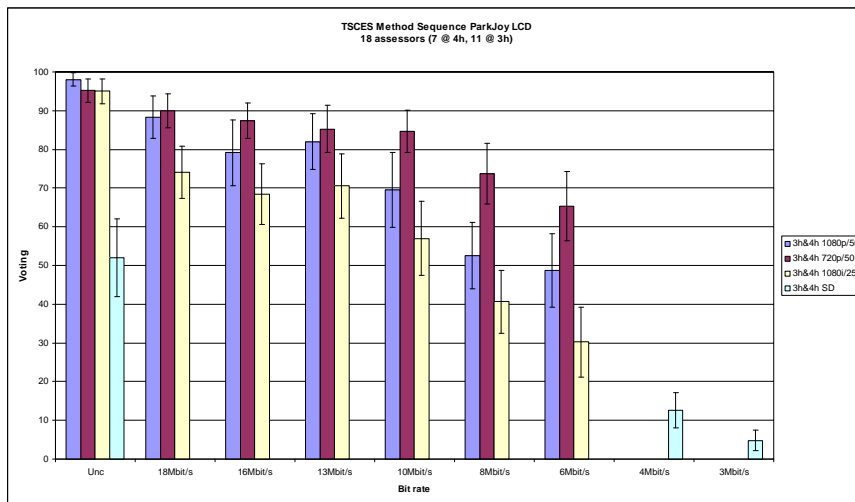


Figure 8 Parkjoy on LCD

The Ice-Dance sequence (Figure 9, Figure 10) has low image criticality. Therefore less compression artefacts are produced and assessors clearly appreciated the higher resolution of the 1080p/50 format.

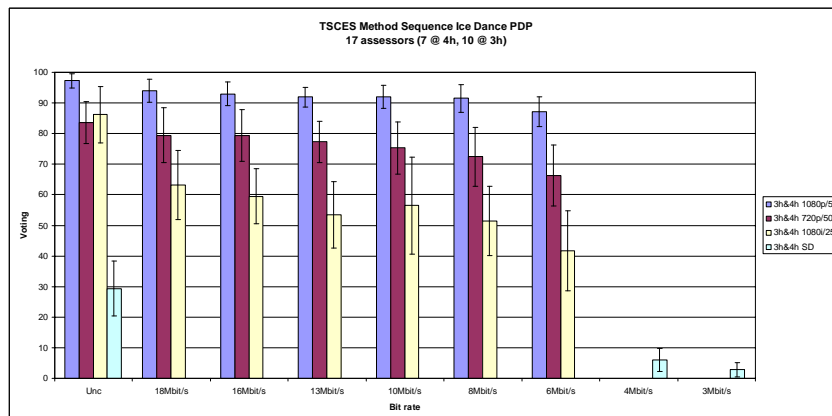


Figure 9 Ice Dance on PDP

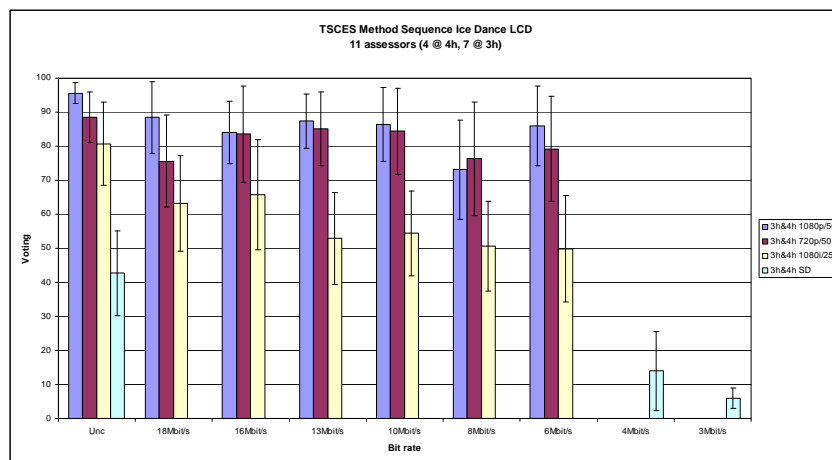


Figure 10 Ice Dance on LCD

In Figure 11 the overall results for both tests and for both display types are shown.

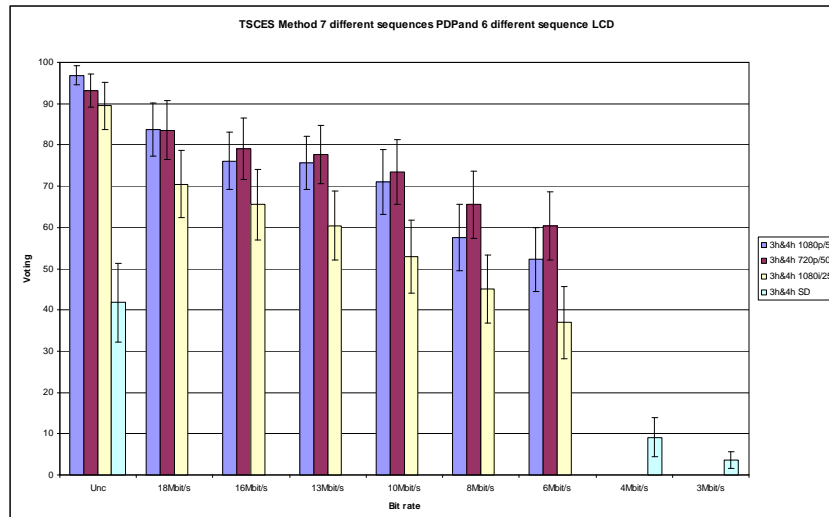


Figure 11 Overall results for all sequences on PDP and LCD

When looking at the results for the **uncompressed** domain the following conclusions can be drawn:

- The 1080i/25 and 720p/50 image format were very similarly rated, but for most of the sequences the 720p/50 format was rated better,
- The de-interlacer in the displays had a greater impact on the native resolution of the display at 3h viewing distance than the spatial up-conversion of 720p/50.

For the images in the **compressed** domain the following conclusion can be drawn:

- The 1080p/50 and 720p/50 image format are very closely rated, but at bit-rates < 17 Mbit/s the 720p/50 format was rated better,
- The 1080i/25 image format was always rated the worst HDTV format in image quality terms.

When investigating the **display dependency** of the votes, Figure 12, Figure 13, Figure 14 show that the assessors rated the image quality relative close for the 1080p/50 format and with some minor differences for the 720p/50 (spatial up-conversion process) and the 1080i/25 (de-interlacing process) formats.

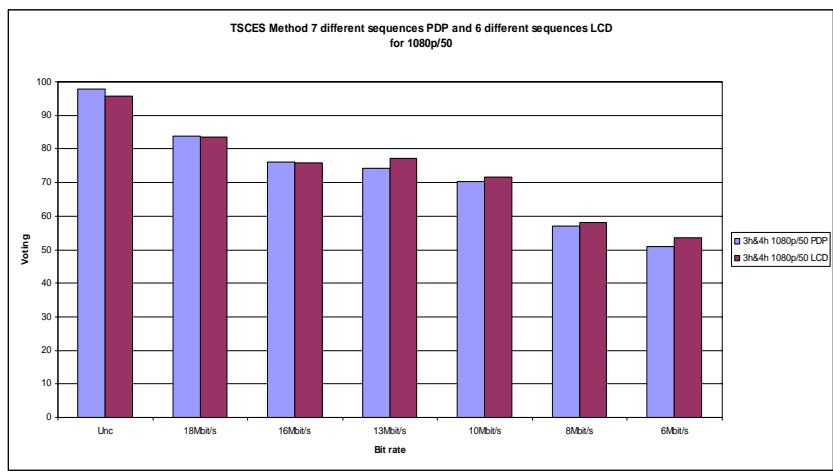


Figure 12 Comparison of assessors votes on 1080p/50 for the LCD and the PDP over all sequences

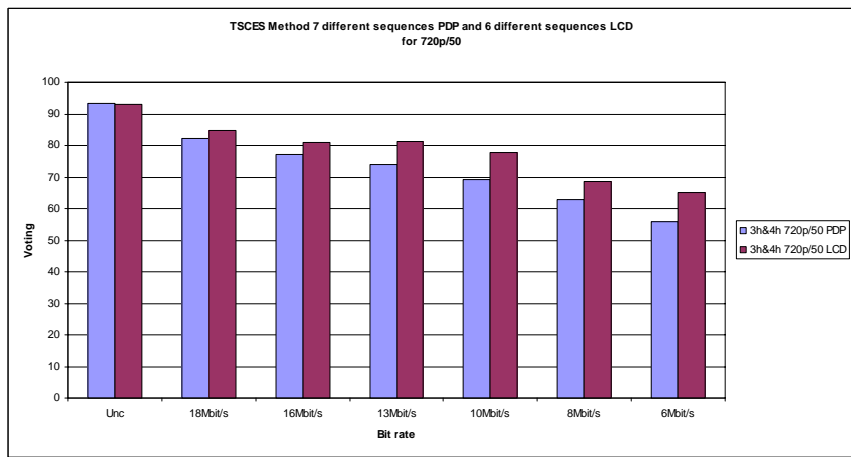


Figure 13 Comparison of assessors votes on 720p/50 for the LCD and the PDP over all sequences

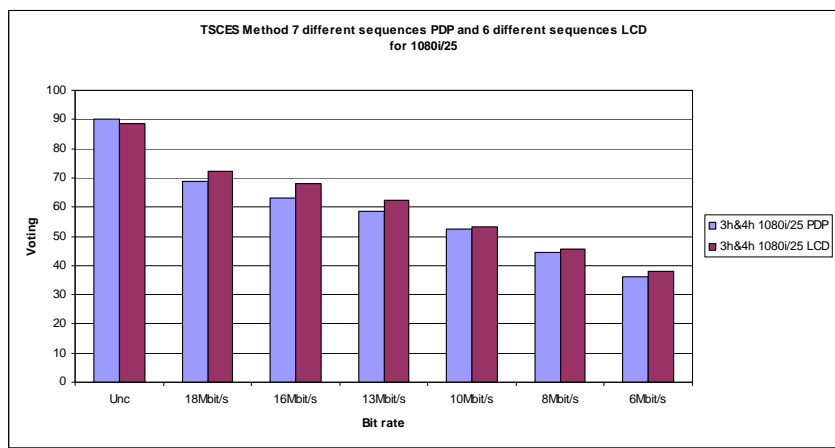


Figure 14 Comparison of assessors votes on 1080i/25 for the LCD and the PDP over all sequences

CONCLUSIONS

In general, based on the test results, it can be stated that the progressive scanning HDTV formats outperformed the 1080i/25 format in terms of image quality at all compressed bit-rates tested. Consequently, and as an overall recommendation, it can be concluded that the 1080i/25 HDTV image format should no longer be applied in production and broadcasting. This would make de-interlaced processing for HDTV (not for legacy SDTV) in FPD obsolete, thus an unknown factor of image quality impairment in the digital chain could be avoided.

Broadcasters which would apply HDTV with a progressive chain at 50 frames per second would be able to provide the similar image quality to the consumers at a lower bit-rate in emission than broadcasters using 1080i/25 (see example in Figure 15). This is also an economical advantage.

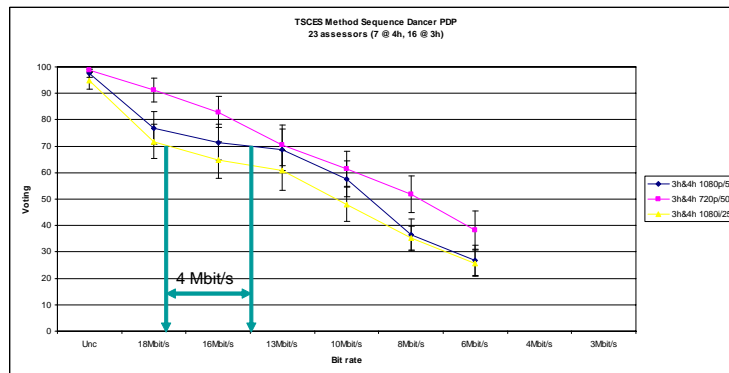


Figure 15 Example for economical bit rate saving by using 720p/50 instead of 1080i/25. The same image quality is achieved with a lower bit-rate in the 720p/50 case. Note: a software codec.

Major contributors to the poor performance of 1080i/25 are the image artefacts created by the compression algorithm, i.e. in the H.264/AVC software encoder. When comparing the 1080p/50 and 1080i/25 format in the uncompressed domain with the new TSCES method, a slightly degraded image quality due to the de-interlacing was determined. In the compressed domain the impact of the de-interlacing might have additionally contributed to the image impairment; however, this was not directly measured.

The 720p/50 signal has a lower spatial resolution than the 1080i/25 format, but double the temporal information. The encoder created less annoying artefacts and the up-scaling process in the display did not create dominant image artefacts when viewed at 3h and 4h. In the uncompressed domain, when comparing 1080p/50 and 720p/50, a slightly degraded image quality based on the up-scaling process of the 720p/50 to the 1920 x 1080 pixel resolution of the display was determined.

A further result was that a 1080p/50 HDTV format would not require more bit-rates in broadcasting than the existing 1080i/25 format when (software coded) H.264/AVC compression is applied. This offers a migration path for all broadcasters which apply 1080i/25 in broadcasting today. With higher critical content and decreasing bitrates however, the 720p/50 format has shown a better image quality than the 1080p/50 format, even when viewed on 1920 x 1080 pixels FPD at the given size. This was due to the effect that, with increasing criticality of the test images, the compression

artefacts of the 1080p/50 became significantly more apparent than with the 720p/50 format and masked any advances in spatial resolution. 1080p/50 with its significantly higher spatial resolution than 720p/50 (thus more information to compress), 'overloaded' the encoder and image artefacts became apparent. On the other hand, and this was shown with the less critical sequence Ice-Dance, the assessors clearly appreciated the higher spatial resolution of the 1080p/50 format against the 720p/50 format when compression artefacts did not mask the spatial resolution advantages of 1080p/50.

This leads further to the conclusion that a 720p/50 HDTV is the right choice for broadcasters today and a 1080p/50 broadcasting format would only provide significant image quality advances over the 720p/50 format with the currently available FPD, if the bit-rates are kept at the same bit-rate or a little bit less than with today's HDTV 1080i/25 broadcasts (which means in the area of $\cong 18$ Mbit/s for the video data rate). On the other hand one has to recognise that 3G-HDTV could be required with new display types of even higher resolution than 1920 x 1080 pixels or for non-consumer home application where the display sizes are greater than about 60 inches.

The subjective tests were conducted with a novel method. The new method utilized an upper and lower reference thus providing comparative data about the image quality of the individual HDTV formats. It was shown that non-expert assessors understood the method very well. The statistical errors produced with the method can be kept low enough with about 25 to 30 assessors, but even as tested here with a fewer number of assessors good results could be derived.

The presentation at IBC will include also a demonstration of the different image format, if a 1080p/50 source and project can be provided.

ACKNOWLEDGEMENTS

We would like to express our particular thanks Prof. Hedtke, Dr. Schnoell, Mr. Eichmueller and Mr. Schreiner for the logistical support for the subjective tests at the University of Applied Sciences Wiesbaden-Germany and all participants in the tests. Further to Mr. Tobias Hinz from the Heinrich Hertz Institute in Berlin for the coding of the sequences.

REFERENCES

[1] HOFFMANN, H., 2006. HDTV - EBU format comparisons at the IBC-2006. *EBU Technical Review*, (308), pp. 1-8.

[2] HOFFMANN, H., ITAGAKI, D., T, WOOD, D. and BOCK, A., 2006. Studies on the bit rate requirements for a HDTV format with 1920 x 1080 pixel resolution, progressive scanning at 50 Hz frame rate targeting large flat panel displays. *IEEE Transactions on Broadcasting*, **4**(52), pp. 420-434.

[3] HOFFMANN, H., ITAGAKI, D., T, WOOD, D., HINZ, T. and WIEGAND, T., 2007. A novel method for subjective picture quality assessment and further studies HDTV formats [under review]. *IEEE Transactions on Broadcasting* .

[4] HOFFMANN, H., ITAGAKI, T. and WOOD, D., 2007. New Psycho-physical Method of Television Picture Quality Evaluation (EBU-II). *IEE Electronics Letters*, **43**(4), pp. 212-213.

[5] ITU-R BT.500-11, 2003. *Methodology for the subjective assessment of the quality of television pictures*. ITU-R BT.500-11. Geneva: International Telecommunication Union.