A novel method for subjective picture quality assessment and further studies of HDTV formats

H Hoffmann, *Member IEEE*, T. Itagaki, *Member IEEE*, D. Wood , T. Hinz, T. Wiegand, *Member IEEE*

Abstract — This paper proposes a novel method for the assessment of picture quality, called Triple Stimulus Continuous Evaluation Scale (TSCES), to allow the direct comparison of different HDTV formats. The method uses an upper picture quality anchor and a lower picture quality anchor with defined impairments. The HDTV format under test is evaluated in a subjective comparison with the upper and lower anchors. The method utilizes three displays in a particular vertical arrangement. In an initial series of tests with the novel method, the HDTV formats 1080p/50, 1080i/25, and 720p/50 were compared at various bit-rates and with seven different content types on three identical 1920 x 1080 pixel displays. It was found that the new method provided stable and consistent results. The method was tested with 1080p/50, 1080i/25, and 720p/50 HDTV images that had been coded with H.264/AVC High profile. The result of the assessment was that the progressive HDTV formats clearly outperformed the interlaced HDTV format. A system chain proposal is given for future media production and delivery to take advantage of this outcome. Recommendations for future research conclude the paper.

***Index Terms*— Compression in broadcasting, High-Definition Television, subjective testing of image quality, flat panel displays.**

I. Introduction

HIGH-DEFINITION TELEVISION (HDTV) is under serious consideration in many countries around the world based on the availability of flat panel displays (FPD) and increasing availability of HDTV content via various media. Many parties are interested in guidance about which HDTV format and compression system to use in the production and distribution environment. In order to answer these kinds of questions, picture quality assessment methods need to be used. This paper deals specifically with the subjective evaluation of HDTV on large flat panel displays.

Manuscript received February 19, 2007. This work was supported by the European Broadcasting Union.

H. Hoffmann is with the European Broadcasting Union, Grand Saconnex, 1218 Switzerland (phone: +41 22 717 2746; fax: +41 22 7474746; e-mail: Hoffmann@ebu.ch).

Dr. T. Itagaki is with School of Engineering & Design, University of Brunel Uxbridge UB8 3PH, UK.

D. Wood is head of the new technologies department with the European Broadcasting Union.

T. Hinz and T. Wiegand are with the Fraunhofer Institute for Telecommunications - Heinrich-Hertz–Institut, Berlin, Germany.

Measuring television picture quality is essential for the development and selection of an HDTV system. Objective methods, such as the measurement of differences between the input and output signals, are only adequate in specific circumstances. Different scene content can be affected in different ways by the same levels of impairments such as noise, and thus objective measurements are often ambiguous. Objective methods cannot fully model the response of the human perceptual systems, or take into account the range of scene content. Hence, the results of objective measurements often do not provide complete information about how an image or video is perceived.

The only accurate and stable methods of evaluating television pictures are psycho-physical evaluation methods, or "subjective evaluations," in defined conditions with defined content that will be critical for the system under test. Such methods are always used for important policy decisions about video systems.

The overall intention of most subjective methods is to establish the average opinion of the population as a whole of the quality associated with an audio-visual system using specific pictures or scenes. This must be done in conditions that are defined and controlled, representative of typical viewing conditions, and from which all biases have been removed or reduced to known levels. The conditions and results must be valid, reproducible, and consistent across laboratories in different parts of the world.

The current methodologies for subjective assessment of the quality of television pictures are given in ITU-R Recommendation BT.500-11[1].

The first method developed by the European Broadcasting Union (EBU), the Double Stimulus Impairment Scale (DSIS), or EBU-I [2], uses the ITU-R BT.500-11 5-grade impairment scale and has been widely used throughout the world. The EBU also refined another method, based on ideas by Allnatt [3], McDiarmid and Derby [4], which it termed the Double Stimulus Quality Scale Method (DSQS). This has also been widely used throughout the world. These methods are based on observer rating test sequences with either discrete or continuous quality or impairment scales.

In this paper we first define a new psycho-physical "Method of Television Picture Quality Evaluation (EBU-II)." We then show how the new method was used in an initial test series. The conditions are described and results analyzed. Finally, the system aspects of HDTV are discussed and suggestions for further research are given.

II. Abbreviations

We abbreviate the various television formats mentioned in this document according to the following nomenclature:

- 1080p/50 is an HDTV format with 1080 horizontal lines and 1920 pixels per line, progressively scanned at 50 frames per second, as specified in SMPTE 274M-2005 [5] and ITU-R BT.709-5 [6].
- 720p/50 is an HDTV format with 720 horizontal lines and 1280 pixels per line, progressively scanned at 50 frames per second, as specified in SMPTE 296M-2001 [7].
- 1080i/25 is an HDTV format with 1080 horizontal lines and 1920 pixels per line, interlace-scanned at 25 frames per second or 50 fields per second, as specified in SMPTE 274M-2005 [5] and ITU-R BT.709-5 [6].
- 576i/25 is a Standard Definition Television Format (SDTV) format with 576 active horizontal lines (625 lines in total) and 720 pixels per line, interlace-scanned at 25 frames per second or 50 fields per second, as specified in ITU-R BT.601-5 [8].

III. Shortcomings of the existing methods:

The measurement scales used have to be translated into the languages in which the tests are being done. However, the adjectives characterizing the image can be interpreted differently by assessors with different mother tongues. There are variable intervals between the meanings of the descriptor adjectives in the scale within the same language, while a given interval varies in perceived size from one language to another. Furthermore, in the existing methods, the reference pictures are displayed on the same screen as the pictures under test, thus relying on the memory of the assessors. We used the DSIS method in our first investigations of the existing HDTV formats 1080i/25 and 720p/50, and a new 1080p/50 HDTV format [9]. We concluded that we could not give a clear answer to the question of which HDTV format would be better and at what bit rate, because we could not include different formats in the same test in an unbiased way. We were only able to report on the failure characteristics of each individual HDTV format.

Our new method addresses these shortcomings and a principal overview on the method was recently published in [10]. It is applied here (but not limited) to HDTV picture quality comparison on large flat panel displays (FPD) and is called the 'Triple Stimulus Continuous Evaluation Scale' method (TSCES) or 'EBU-II'.

IV. Detailed challenges the work should address

The new method should meet the following requirements:

- allow the direct comparison of different HDTV scanning formats with reporting in a single resulting graph;
- be 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);
- provide reliable and reproducible results, with a standard deviation determined only by the natural spread of opinion, and with the stability of the results as constant as possible over the quality range being evaluated;
- provide independence of language in the adjectives describing the perceived image quality, and have scale interval linearity;
- cope with a wide range of picture quality and HDTV formats such as 720p/50 and 1080i/25, with third generation HDTV formats such as 1080p/50, and with standard definition television (SDTV);
- be able to measure accurately a video system's basic quality and failure characteristics (the relationship between quality and the parameters which reduce it);
- be usable with large and medium sized flat panel displays, LCD or PDP, as these will constitute the dominant mode for viewing in the years ahead for both conventional television and the coming generations of high definition television.

V. The Method:

Assessors are presented with three monitors one above the other as shown in **Error! Reference source not found.** Figure 1. For HDTV evaluations, the vertical angles of the three displays are adjusted in such a way that a reference viewer at an eye height of 1.2 m and in a center position relative to the screens maintains a constant viewing distance of 3 times picture height (3h) from all three displays.

This distance matches the design viewing distance for HDTV, which is why it is used here, but the method could be applied to other design viewing distances. Having the monitors mounted above one another, the assessors quickly grasp what is expected of them, and the arrangement is naturally suited to widescreen displays. Using displays with a comfortable viewing angle will also permit more assessors per viewing session. In addition the following settings are applied:

- ITU-R BT.500-11 viewing environment and ambient light conditions
- All three displays show the same scene content at the same time
- All three displays need to be aligned and of the same type and should be reference type displays (unless particular examples of other display categories are being tested)
- The top display serves as an upper reference, providing a high quality anchor
- The middle display shows the pictures under test (preferably including unidentified upper and lower anchor content for verification purposes)

- The bottom display serves as the low anchor with a defined impairment added

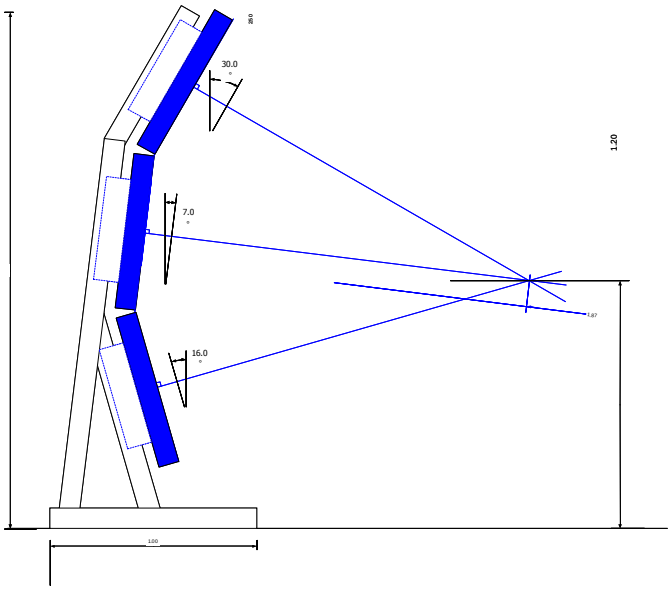

Figure 1 Display rack configuration, with display angles allowing exact 3h viewing distance for a reference viewer with 1.2m eye height (technical drawing by Edgar Wilson, EBU)

The types of impairment used for the bottom anchor must be clearly defined, must be reproducible, and, in order to help the orientation of the assessors, must be of similar characteristics to the impairments expected on the middle display. We have experimented, for example, with adding white noise as a defined impairment factor for the lower anchor bottom display, but found that such an impairment is too different from the impairments caused by H.264/AVC [16,17] coding that we presented on the middle display. A better solution for creating robust lower anchor impairments was found by using the publicly available and defined reference encoders of the same compression system as used for the images under test on the middle display.

For comparison of the HDTV formats 1080p/50, 1080i/25, and 720p/50 in uncompressed and compressed form we propose the following conditions:

A. Content

Top display high image quality anchor:

- uncompressed HDTV signal with 1080p/50

Middle display with images under test:

- 1080p/50, 1080i/25 and 720p/50 HDTV, and 576i/25 SDTV format at various bit-rates. The coding parameters need to be documented. Upper and lower anchors to be included as hidden references.

Bottom display low image quality anchor:

- 576i/25 Standard Definition Television (SDTV) format down-converted from a 1080i/25 HDTV source and then compressed with an algorithm that produces a) a clear lower anchor reference, and b) impairments similar to those of the compressed HDTV image under test on the middle display. This lower anchor also provides a very practical visualization of today's SDTV broadcasts when shown on a large FPD.

B. Presentation:

The scene content on all three monitors must always be identical and in time synchronism. First, a training session and explanation has to be given to the assessors. Following that, each test sequence should have a minimum length of 10 seconds and should be repeated four times before the assessors are asked to vote. We found that the assessors were comfortable assessing the images presented on the middle display compared to the top and bottom displays. The length of each viewing session should be set to a maximum of 30 minutes with two short breaks. The middle display test sequences should be shown in randomized order, and the test sequences should include the upper and lower anchors to verify the consistency of the assessors and the method.

C. Display:

Ideally three reference-quality displays (Grade-1 type) should be used, aligned to each other with identical settings according to the procedures of ITU-R BT.500-11. Unfortunately, no FPD reference-quality displays have been available so far, thus an exact report and characterization (measurements) of the displays' parameters is required.

D. Voting

Assessors should be given clear instructions before the tests begin, and be provided with a computer screen or paper on which is drawn a continuous vertical line 100 mm (4 inches) in length on which to make their assessment. The top end of the line is defined as representing the quality of the top monitor; the bottom end of the line represents 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 bottom limits. The top and bottom are thus upper and lower anchors for the evaluations. In subsequent processing, the results can be mapped onto the 5 impairment categories or quality scales or onto a 100 point continuous quality scale (see ITU-R BT.500-11). An example scale is shown in Figure 2.

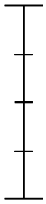

Figure 2 Scale of 100mm length used for voting (here shown in reduced form)

E. How does the method provide robust and repeatable results?

The upper, middle and lower anchor image quality levels are controlled and defined and can be reproduced by other laboratories as long as the coding parameters and algorithms are known and documented. Display settings and room alignments follow the ITU-R BT.500-11 recommendation, and the analysis of the results can utilize well known statistical methods (assessor screening, arithmetic mean, standard deviation, variance etc.).

F. Reporting of the results:

The display parameters, viewing conditions, and voting procedures should be documented. The details of the training session, embedded reference scenes, and sequence order in the actual test should be recorded. The statistical analysis of the results including assessor screening can draw on the guidelines given in ITU-R BT.500-11. The detailed technical parameters of upper anchor signal, the impaired signal, and how the lower anchor signal was generated in particular need to be documented. The type of compression algorithm and settings (i.e. configuration files) for the middle and bottom display content must be recorded.

VI. TESTING THE METHOD

The subjective test sessions were conducted November 13 - 17, 2006, with a total of 173 mainly student assessors (see Figure 3), at the University of Applied Sciences in Wiesbaden, Germany. These assessors had been checked for standard visual acuity and colour perception.

Figure 3 Photo of the viewing session

A. Displays and viewing:

We chose three Pioneer PDP EX5000 consumer displays with 1920 x 1080 pixels resolution. The displays were aligned with a PLUGE signal [11] for brightness and contrast and to a peak luminance of about 100 cd/m^2 with a Photo Research type PR705 spectrophotometer. The display settings and ambient light conditions were identical to those in our previous publication [9]. We therefore exclude a detailed characterization of the displays in this paper. Because the plasma displays used had relatively good viewing angle uniformity, we were able to use two seating rows comprising three to four assessors at 3h and four assessors at 4h viewing distance. Each voting position was exactly documented.

B. System set-up:

Each of the displays was connected via DVI to a DVS Pronto2k workstation that could play out the required uncompressed HDTV and SDTV formats. The three workstations were synchronized via RS422 for start and stop of the sequences. The scene content on all three monitors was identical and in time synchronism.

C. Presentation:

In each viewing session the assessors were given an explanation and a training sequence. The total length of a session was limited to two different content types. Each content type of 10 seconds length was presented four times before the voting was conducted on paper according to the scale shown in Figure 2. Seating position, differentiation between experts and non-experts, vision (acuity and color), gender and age were recorded.

The top display showed 1080p/50 uncompressed pictures (perceived to be "excellent") and the bottom display showed 576i/25 (SDTV) scenes coded with the JM11 reference encoder for H.264/AVC [12] at a bit rate of 3 Mbit/s, with defined encoder settings. This provided an ITU-R BT.500-11 quality category perceived to be "bad" at the given viewing distance, with impairments of a kind similar to those being tested on the middle display.

The following HDTV formats and bit rates were tested on the middle display:

- HDTV formats 1080p/50, 720p/50 and 1080i/25, at 18, 16, 13, 10, 8, 6 Mbit/s, plus uncompressed upper anchor reference.
- SDTV format 576i/25 at 4 Mbit/s, uncompressed, plus the 3 Mbit/s lower anchor reference.

The sequences were presented in randomized order and the assessors were not informed about the formats shown.

D. Content selection:

A limitation of our first tests [9] was that we used only one type of content from the SVT test-set [13]. This content was over-sampled relative to the formats under test (1080p/50, 1080i/25 and 720p/50), because it was generated on 65 mm film at 50 frames per second and scanned to 2160p/50. In the new tests we used a total of seven different sequences. Three of them are from the SVT test set, but in addition we generated four new sequences with a state of the art Sony HDC1500 CCD camera. This camera had a 1920 x 1080 pixel sensor operating at 50 frames per second and provided a 1080p/50 signal on its dual link HD-SDI output for uncompressed recording on a DVS workstation type Pronto2K. The test sequences are described in Table 1.

	Name	**Source Format before downsampling**	**Characterization**

		and origin	
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	Princess-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 Content used in the assessment

E. Processing of the test content:

The SVT content was already available in the various HDTV formats (1080p/50, 1080i/25 and 720p/50) and only had to be converted from an SGI file format with 10-bit and 4:4:4 color resolution to YUV 8-bit 4:2:0 sampling prior to H.264/AVC coding. Details of the SVT content can be found in [13]. The content generated with the HDC1500 camera in the 1080p/50 format (4:2:2, 10 bit) was processed according to the following conditions:

- 1080i/25 from 1080p/50 (CCD): lines of the first 1080i/25 field were generated by box filter/averaging the first frame of the 1080p/50 source. Second field 1080i/25 lines were generated by box filtering/averaging the next 1080p/50 frame. The second field was then multiplexed with the first field leaving one line out. The method is similar to implementations in CCD cameras.
- 720p/50 from 1080p/50 (CCD): the DVS workstation real-time down-sampling function (software version 2.1.1.0) was used to apply a low pass filter followed by a Sinc-filter.
- 576i/25 from 1080i/25 (see above): the DVS workstation down-sampling was used to apply a low pass filter followed by a Sinc filter. The SVT 1080i/25 content was also down-converted to 576i/25 with this method.

F. Encoding method and parameters

One key problem in video compression is operational control of the source encoder. Typical video sequences contain widely varying content and motion. This requires a selection between different coding options with varying rate/distortion efficiency for different parts of the image. The task of coder control is to determine a set of coding parameters, and thereby the bit stream, so that a certain rate/distortion trade-off is achieved for a given decoder.

The coder control used for encoding the HDTV sequences is based on Lagrangian bit-allocation techniques. The popularity of this approach is due to its effectiveness and simplicity. For completeness, we will briefly review the Lagrangian optimization techniques, and explain their application to video coding and temporal decomposition. Finally, this section specifies the settings for the H.264/AVC encoder used for HDTV sequences.

1) Optimization Using Lagrangian Techniques

Consider K source samples that are collected in the K-tuple $\mathbf{S} = (\boldsymbol{S}_1,\ldots,\boldsymbol{S}_K)$. A source sample $\boldsymbol{S}_k$ can be a scalar or vector. Each source sample $\boldsymbol{S}_k$ can be quantized using several possible coding options that are indicated by an index of the set $\boldsymbol{O}_k = (O_{k1},\ldots,O_{kN_k})$. Let $I_k \in \boldsymbol{O}_k$ be the selected index to code $\boldsymbol{S}_k$. Then the coding options assigned to the elements in $\mathbf{S}$ are given by the components in the K-tuple $\mathbf{I} = (I_1,\ldots,I_K)$. The problem of finding the combination of coding options that minimizes the distortion for the given sequence of source samples subject to a given rate constraint R_c can be formulated as

$$\min_{\mathcal{I}} D(\mathcal{S},\mathcal{I}) \quad \text{subject to } R(\mathcal{S},\mathcal{I}) \le R_c \tag{1}$$

Here, $D(\mathbf{S},\mathbf{I})$ and $R(\mathbf{S},\mathbf{I})$ represent the total distortion and rate, respectively, resulting from the quantization of $\mathbf{S}$ with a particular combination of coding options $\mathbf{I}$. In practice, rather than solving the constrained problem in Eq. (1), an unconstrained formulation is employed, that is

$$\mathcal{I}^* = \operatorname*{argmin}_{\mathcal{I}} J(\mathcal{S},\mathcal{I} \mid \lambda) \quad \text{with } J(\mathcal{S},\mathcal{I} \mid \lambda) = D(\mathcal{S},\mathcal{I}) + \lambda \cdot R(\mathcal{S},\mathcal{I}) \tag{2}$$

with $\lambda \ge 0$ being the Lagrange parameter. This unconstrained solution to a discrete optimization problem was introduced by Everett [14]. The solution $\mathcal{I}^*$ to (2) is optimal, if a rate constraint R_c corresponds to λ. In this case the total distortion $D(\mathbf{S},\mathbf{I}^*)$ is minimized for all combinations of coding options with bit rate less or equal to R_c.

We can assume additive distortion and rate measures, and let these two quantities be dependent only on the choice of the

parameter corresponding to each sample. Then, a simplified Lagrangian cost function can be computed using

$$J(\boldsymbol{S}_k,\mathcal{I} \mid \lambda) = J(\boldsymbol{S}_k,I_k \mid \lambda). \tag{4}$$

In this case, the optimization problem in (3) reduces to

$$\min_{\mathcal{I}} \sum_{k=1}^{K} J(\boldsymbol{S}_k,\mathcal{I} \mid \lambda) = \sum_{k=1}^{K} \min_{I_k} J(\boldsymbol{S}_k,I_k \mid \lambda) \tag{5}$$

and can easily be solved by independently selecting the coding option for each $\boldsymbol{S}_k \in \mathbf{S}$. For this particular scenario, the problem formulation is equivalent to the bit-allocation problem for an arbitrary set of quantizers, proposed by Shoham and Gersho [15].

2) Lagrangian Optimization in Hybrid Video Coding

The application of Lagrangian techniques to control a hybrid video coder is not straightforward, because of temporal and spatial dependencies of the rate/distortion costs. Consider a block-based hybrid video codec such as H.264/AVC [16-18]. Let the image sequence s be partitioned into K distinct blocks $\boldsymbol{A}_k$ and the associated pixels be given as $\boldsymbol{S}_k$. The options $\mathbf{O}_k$ to encode each block $\boldsymbol{S}_k$ are categorized into INTRA and INTER, i.e. predictive coding modes with associated parameters. The parameters are transform coefficients and the quantization parameter Q for both modes plus one or more motion vectors for the INTER mode. The parameters for both modes are often predicted using transmitted parameters of preceding modes inside the image. Moreover, the INTER mode introduces a temporal dependency because reference is made to prior decoded pictures via motion-compensated prediction. Hence, the optimization of a hybrid video encoder would require the minimization of the Lagrangian cost function in (2) for all blocks in the entire sequence. This minimization would have to proceed over the product space of the coding mode parameters. This product space is far too large to be evaluated. Therefore, various publications elaborate on reduction of the product space and thus reducing complexity. An overview is given in [19].

A simple and widely accepted method of INTER coding mode selection is to search for a motion vector that minimizes a Lagrangian cost criterion prior to residual coding. The bits and distortion of the following residual coding stage are either ignored or approximated. Then, given the motion vector(s), the parameters for the residual coding stage are encoded. The minimization of a Lagrangian cost function for motion estimation as given in (3) was first proposed by Sullivan and Baker [20].

Therefore, we split the problem of optimum bit allocation for INTER modes in a motion estimation and successive macroblock mode decision process between INTER or INTRA coding modes. The utilized macroblock mode decision is similar to [21] but without consideration of the dependencies of distortion and rate values on coding mode decisions made for past or future macroblocks. Hence, for each macroblock, the coding mode with associated parameters is optimized given the decisions made for prior coded blocks only. Consequently, the coding mode for each block is determined using the Lagrangian cost function in (3). Let the Lagrange parameter λ_{MODE} and the quantization parameter Q be given. The Lagrangian mode decision for a macroblock $\boldsymbol{S}_k$ proceeds by minimizing

$$\begin{aligned} &J_{MODE}(\boldsymbol{S}_k,I_k \mid Q,\lambda_{MODE}) = \\ &D_{REC}(\boldsymbol{S}_k,I_k \mid Q) + \lambda_{MODE} R_{REC}(\boldsymbol{S}_k,I_k \mid Q) \end{aligned} \tag{7}$$

where the macroblock mode I_k is varied over the sets of possible macroblock modes for H.264/AVC. As an example, the following sets of macroblock modes can be used for P pictures (or P slices) when coding progressive-scanned video:

INTRA-4×4, INTRA-16×16, SKIP, INTER-16×16, INTER-16×8, INTER-8×16, INTER-8×8

H.264/AVC additionally provides the following set of sub-macroblock types for each 8×8 sub-macroblock of a P-slice macroblock that is coded in INTER-8×8 mode: INTER-8×8, INTER-8×4, INTER-4×8, and INTER-4×4.

In the case of interlace coding, macroblock pairs, i.e. two vertically arranged macroblocks, are considered and the two macroblocks are coded in either frame mode or field mode. The former treats the samples as in progressive coding, while the latter assigns macroblock rows 0, 2, 4, … 30 to the top macroblock and rows 1, 3, … 31 to the bottom macroblock. The macroblock modes above are then represented when the macroblock pair is in frame and field mode for the coder control.

The distortion $D_{REC}(\boldsymbol{S}_k,I_k|Q)$ and rate $R_{REC}(\boldsymbol{S}_k,I_k|Q)$ for the various modes are computed as follows: For the INTRA modes, the corresponding 8×8 or 4×4 blocks of the macroblock $\boldsymbol{S}_k$ are processed by transformation and subsequent quantization. The distortion $D_{REC}(\boldsymbol{S}_k,\text{INTRA}|Q)$ is measured as the sum of the squared differences (SSD) between the reconstructed (s') and the original (s) macroblock pixels

$$SSD = \sum_{(x,y)\in\boldsymbol{\mathcal{A}}} \left| s[x,y,t] - s'[x,y,t] \right|^2 \tag{8}$$

where the set $\boldsymbol{A}$ represents the samples of the subject macroblock. The rate $R_{REC}(\boldsymbol{S}_k,\text{INTRA}|Q)$ is the rate that results after entropy coding.

For the SKIP mode, the distortion $D_{REC}(\boldsymbol{S}_k,\text{SKIP}|Q)$ and rate $R_{REC}(\boldsymbol{S}_k,\text{SKIP}|Q)$ do not depend on the current quantizer value. The distortion is determined by the SSD between the current picture and the value of the inferred INTER prediction, and the rate is given as approximately one bit per macroblock.

The computation of the Lagrangian costs for the INTER modes is much more demanding than for the INTRA and SKIP modes. This is because of the block motion estimation step. The size of the blocks $\boldsymbol{S}_i$ within a macroblock is $A\times B$ pixels for the INTER-A×B mode. Given the Lagrange parameter λ_{MOTION} and the decoded reference picture s', rate-constrained motion estimation for a block $\boldsymbol{S}_i$ is performed by minimizing the Lagrangian cost function

$$\boldsymbol{m}_i = \arg\min_{\boldsymbol{m}\in\boldsymbol{\mathcal{M}}} \left\{ D_{DFD}(\boldsymbol{S}_i,\boldsymbol{m}) + \lambda_{MOTION} R_{MOTION}(\boldsymbol{S}_i,\boldsymbol{m}) \right\} \tag{9}$$

where M is the set of possible motion vectors and with the distortion term being given by

$$D_{DFD}(\boldsymbol{S}_i,\boldsymbol{m}) = \sum_{(x,y)\in\mathcal{A}} \left| s[x,y,t] - s'[x-m_x, y-m_y, t-m_t] \right|^p \quad (10)$$

with p=1 for the SAD and p=2 for the SSD. $R_{MOTION}(\boldsymbol{S}_i,\boldsymbol{m})$ is the number of bits required to transmit all components of the motion vector (m_x, m_y), and, in the case where multiple reference frames are used, the reference picture index m_t. The search range $\mathcal{M}$ is ±32 integer pixel positions horizontally and vertically and either one or more prior decoded pictures are referenced. Depending on the use of SSD or SAD, the Lagrange parameter λ_{MOTION} has to be adjusted.

The motion search that minimizes (9) proceeds first over integer-pixel locations. Then, the best of those integer-pixel motion vectors is tested to see whether one of the surrounding half-pixel positions provides a cost reduction in (9). This procedure of determination of a sub-pixel position is called half-pixel refinement. Then, the previously determined half-pixel location is used as the center for the corresponding quarter-pixel refinement step. The sub-pixel refinement yields the resulting motion vector $\boldsymbol{m}_i$. The resulting prediction error signal $u[x,y,t,\boldsymbol{m}_i]$ is processed by transformation and subsequent quantization, as in the INTRA mode case. The distortion D_{REC} is also measured as the SSD between the reconstructed and the original macroblock pixels. The rate R_{REC} is given as the sum of the bits for the mode information, the motion vectors as well as the transform coefficients.

A final remark should be made regarding the choice of the Lagrange parameters λ_{MODE} and λ_{MOTION}. In [19, 22] a relationship between the Lagrange parameter and quantization parameter was determined via experimental results for H.263/MPEG-4 Visual. This experiment has also been conducted for H.264/AVC, providing the following equation

$$\lambda_{MODE} = 0.85 \cdot 2^{(Q_{H.264}-12)/3} \quad (11)$$

For the Lagrange parameter for motion estimation, we follow [19, 22] by choosing for SAD in (9)

$$\lambda_{MOTION} = \sqrt{\lambda_{MODE}} \quad (12)$$

Correspondingly for SSD in (9), we would use

$$\lambda_{MOTION} = \lambda_{MODE} \quad (13)$$

Thus, rate control in those codecs is conducted via controlling the quantization parameter and adjusting the Lagrange parameters accordingly using Eqs. (11)-(13).

3) Temporal decomposition for H.264/AVC encoding

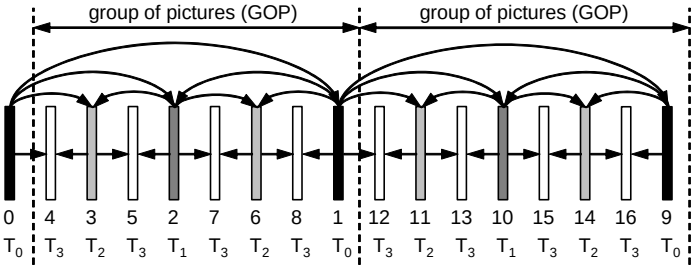

Figure 4 Hierarchical B picture coding structure. The numbers directly below the pictures specify coding order and the symbols T_X specify the temporal layers with X representing the corresponding temporal level.

The temporal structure in our H.264/AVC encoding is called hierarchical B pictures [23], as illustrated in Figure 4. The hierarchy of pictures can be explained by temporal layers. For the base layer pictures (indexed with T_0), P picture coding is often used, as the previous picture is only used for reference. The distance between two P pictures determines the so-called GOP size. Given two surrounding P pictures, the picture half way between them is coded as a B picture (indexed with T_1). Given surrounding T_0 and T_1 pictures, the picture half way between them is also coded as a B picture (but indexed with T_2). This hierarchy of B picture coding can be continued until all pictures are coded. The described hierarchy uses a dyadic partitioning of the temporal axis, although other partitioning is also possible. In this work, we used dyadic partitioning exclusively.

The coding efficiency for hierarchical prediction structures is highly dependent on how the quantization parameters are chosen for pictures of different temporal levels. Intuitively, the base pictures should be coded with highest fidelity, since they are directly or indirectly used as references for motion-compensated prediction of all other pictures. For the next temporal level a larger quantization parameter should be chosen, since the quality of these pictures influences fewer pictures. Following this rule, the quantization parameter should be increased for each subsequent hierarchy level. Based on a given quantization parameter Q_0 for pictures of the temporal base layer, the quantization parameters for enhancement layer pictures of a given temporal level $k > 0$ are determined by $Q_k = Q_0 + 3 + k$. The Lagrange parameters for each picture are adjusted according to Eq. (11). Although this strategy for cascading the quantization parameters over hierarchy levels results in relatively large PSNR fluctuations inside a group of pictures, subjectively the reconstructed video appears to be temporally smooth without any annoying temporal pumping artifacts. We have compared the coding efficiency of dyadic hierarchical prediction structures with P and B pictures with conventional prediction structures as IPPP… and IBBP… (respectively) for a large set of test sequences, of which the results for two earlier example sequences are shown in Figure 5.

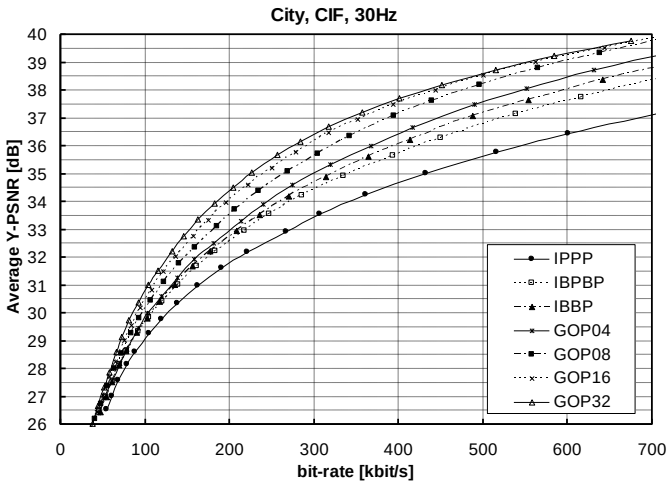

Figure 5 Coding efficiency comparison of hierarchical prediction structures and conventional IPPP, IBPBP, and IBBP coding structures.

4) Exact encoding conditions

Encoding was conducted for the following three picture formats

- 720p/50 H.264/AVC Level 4.0
- 1080p/50 H.264/AVC Level 5.0
- 1080i/25 H.264/AVC Level 4.0

The following H.264/AVC settings were used for all three picture format encodings:

- High Profile used
- 8x8 transform enabled
- Default quantization on
- Default deblocking filter settings on
- Temporal direct mode used
- One slice per picture

The test sequences were compressed using the following matched settings:

720p/50 Level 4.0

- 24-picture hierarchical GOP
- motion vector search range ± 96 pixel

1080p/50 Level 5.0

- 24-picture hierarchical GOP
- motion vector search range ± 128 pixel
- cropping 1080/1088 enabled (padding at lower picture border)

1080i/25 Level 4.0

- 6-picture hierarchical GOP
- RD-optimized MbAFF and Picture-AFF decisions
- motion vector search range ± 128 pixel
- cropping 1080/1088 enabled (padding at lower picture border)

G. Results

Each vote on the 100 mm paper scale was measured and edited in Excel for processing. For example, a mark at the top of 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 as the uncompressed upper anchor on the top display, and a marker at the 0 mm point (bottom) of the scale would have meant that the middle picture was as bad as the lower anchor on the bottom display.

First of all a screening of the votes was performed. From the total of 173 participants (non-experts and experts) four had to be excluded because they mixed up the voting on paper (this was discovered during editing the data), and one participants result was excluded after the statistical screening test of ITU-R BT.500-11.

We first provided the results structured for each test sequence. Assessors that identified themselves as 'expert viewers' were excluded from the following graphs, consequently reducing the overall number of assessors. Figure 6 to Figure 12 show the arithmetic mean with both 3h and 4h viewing distances and the error rate within a 95% confidence interval.

As a general result we can observe:

Hidden references (upper and lower anchor) were clearly detected by the assessors. Even the slight difference between 3 Mbit/s SDTV and 4 Mbit/s SDTV became clearly visible in the votes. With a smaller numbers of assessors (~15) the error increased. In the following descriptions of the sequences used we indicate in parentheses whether the content was generated with the CCD camera (CCD) or scanned in 2160p/50 from 65mm/50fps film (SVT).

Sequence Crowd Run (SVT):

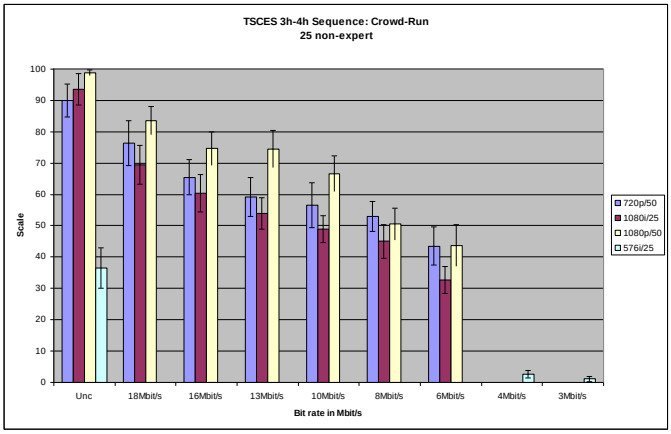

Figure 6 Sequence: Crowd-Run

We used the Crowd Run sequence again in this test in order to compare the results of the new method with our previous DSIS tests [9]. Our assumptions from the earlier test were fully confirmed. With this sequence the 1080p/50 format was in fact rated better than the 720p/50 format and much better rated than the 1080i/25 format.

Parkjoy (SVT):

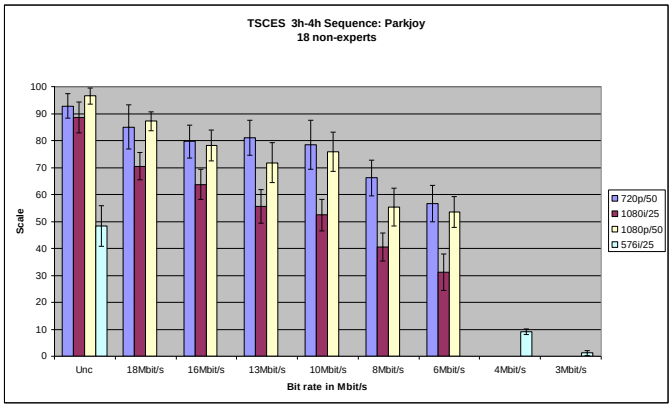

Figure 7 Sequence: Parkjoy

This sequence was relatively critical, thus stressing the encoder. We see that below 16 Mbit/s the 720p/50 format was voted better than 1080p/50. The 1080i/25 format was voted worst.

Princess Run (SVT):

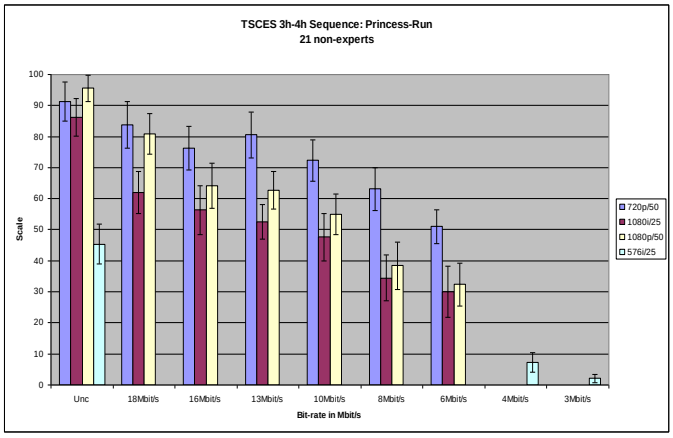

Figure 8 Sequence: Princess-Run

This sequence was very critical, thus stressing the encoder. We see already that at 18 Mbit/s the 720p/50 format was voted better than 1080p/50. The 1080i/25 format was voted worst. An unusual effect can be observed in the 720p/50 voting for 18 Mbit/s and 16 Mbit/s: it seemed that the sequences were presented in the wrong order. However, verification of the playout list did not confirm this. So far we have no explanation for this effect.

Aloha Wave soccer field (CCD):

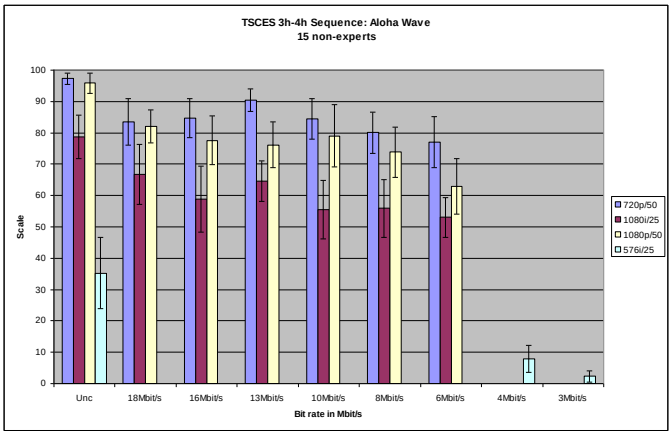

Figure 9 Sequence: Aloha Wave

This graph sequence created some difficulties in interpretation: the content comprised a wide zoom shot of a soccer stadium with considerable texture detail, and contained a camera pan during which the audience was in the process of standing up (so-called "Aloha Wave"). All three formats showed a second peak of maximum quality at a midrange bit rate before dropping off. This did not follow the normal failure characteristics from higher to lower bit rate. This sequence was therefore possibly not suitable for subjective tests from the content point of view. Also, the fact that only 15 assessors participated in this test may have contributed to this result.

Ice Dance (CCD):

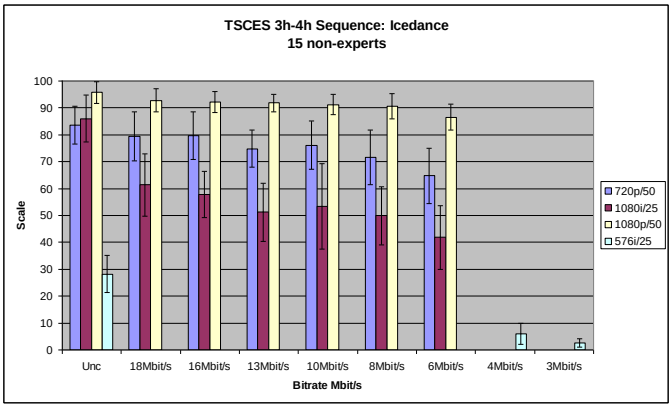

Figure 10 Sequence: Ice-dance

The Ice Dance sequence was a particularly interesting item of content. Two actors were dancing on ice inside a large studio. The criticality was low (an ice surface is easy to encode), but some texture details in the background and the lighting seemed to challenge the interlaced system. Since the sequence was not difficult to encode, the 1080p/50 format was rated best - the assessors appreciated the high spatial temporal resolution that was not significantly masked by compression artifacts; this preference was followed by 720p/50 and 1080i/25.

Dancer (CCD):

As mentioned above, after screening according to ITU-R BT.500-11, one assessor's voting was removed during this sequence.

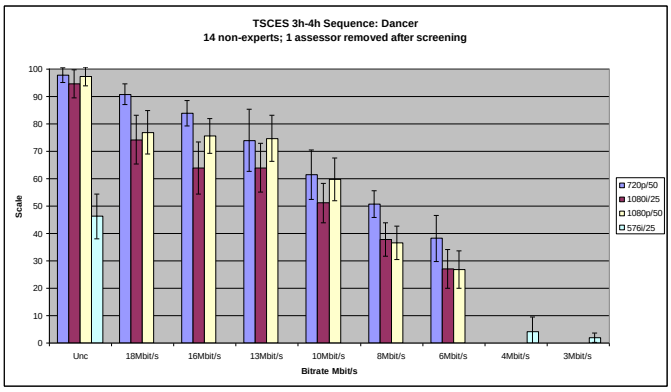

Figure 11 Sequence: Dancer

This sequence can be characterized as difficult. A dancer with a large costume was performing rotational movements on a grass surface, thus stressing the encoder. The failure characteristic was similar for all formats; at lower bit rates 720p/50 was rated best.

Police Boat (CCD):

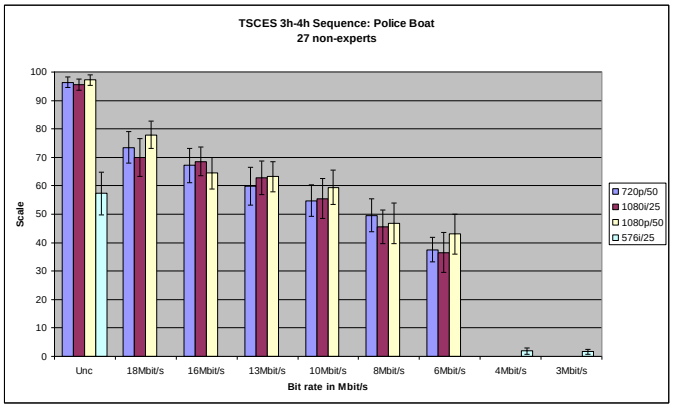

Figure 12 Sequence: Police boat

This was a difficult sequence with a small police boat on water with complex wave motion. Interlaced artifacts were visible on the outline of the boat; the progressive 720p/50 and 1080p50 formats also showed visible coding artifacts in the water. All three formats performed similarly.

The following Figure 13 shows the overall results for the non-experts and Figure 14 for the expert viewers by combining the seven sequences in one graph.

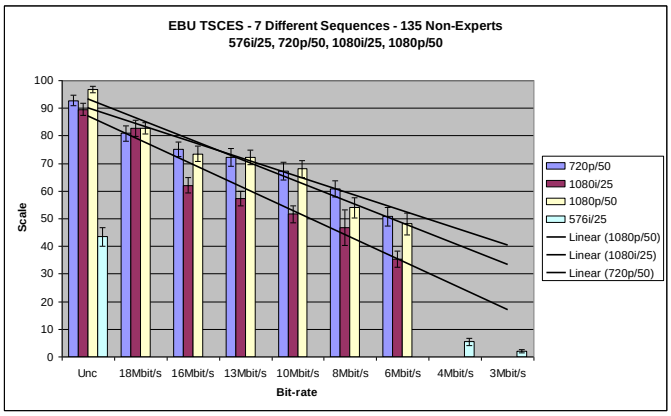

Figure 13 Overall result for non-experts with linear trend line to better visualize the crossover point between 1080p/50 and 720p/50 at about 14 Mbit/s. Note that the linear trend line does not represent a correct interpolation.

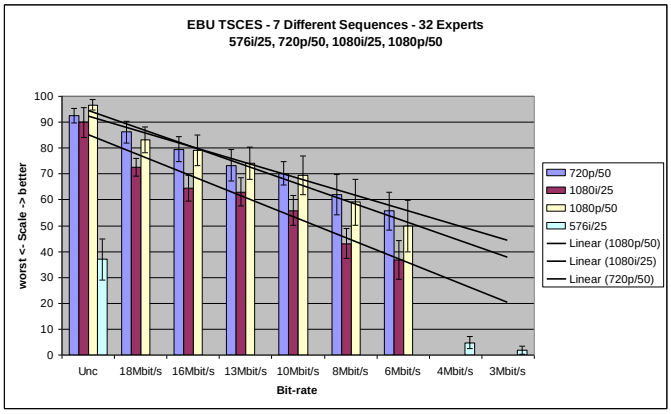

Figure 14 Overall results for the expert viewers amongst the assessors. Linear trend line added, showing a crossover between 1080p/50 and 720p/50 at 16 Mbit/s. Note that the linear trend line does not represent the correct interpolation.

VII. CONCLUSION ON THE NOVEL METHOD

We have shown a new subjective evaluation method which permits the direct comparison of different HDTV formats at different bitrates. The method allows the direct comparison of different HDTV formats and shows the failure characteristics in one common graph, thus permitting a clear comparative analysis. It requires relatively elaborate technical facilities such as three uncompressed HDTV sources, three identical and aligned displays and a display rack, but it provides a robust quality evaluation.

We hope that the method will lead to easier international agreements on video formats and systems. The results appeared reliable and robust and known statistical analysis methods from the ITU-BT.500-11 can be applied. Other laboratories are encouraged to verify this new method and our tests independently.

VIII. Interpretation and discussion of subjective test results and idealized system chain

The practical tests of the method have reinforced the conclusions of earlier investigations on the use of 1080p/50, 720p/50 and 1080i/25, which found that a progressive HDTV format provides better perceived image quality than the 1080i/25 format when compressed with H.264/AVC and viewed on FPDs at typical broadcast bit rates between 6 and 18 Mbit/s. The impact of spatial up-sampling of the 720p/50 format to the 1920 x 1080 pixel resolution of the display did not seem to have any negative impact at 3h and 4h viewing distance for most sequences. In fact, for these sequences the expert viewers commented that the up-scaling artifacts were only visible below 2h viewing distance and that a degree of visible noise in the 720p/50 image at 3h provided a sensation of sharpness. The display size was 50 inches and it is recommended that the findings from [24] be noted, which have shown that displays larger than 50 inches would require a higher spatial resolution than 720p/50 offers. On the other hand, broadcasters are required to identify the target display size for the majority of viewers (i.e. 37 - 42 inch display size).

For the 1080p/50 format, which provides a high spatio-temporal resolution, we found a clear preference by the assessors for uncritical material or for higher bit-rates. One should not forget that the uncompressed video bit rate is twice the bit rate of a 1080i/25 format. With decreasing bit rates the resolution advantage of 1080p/50 was masked by compression artifacts, and the assessors voted in favor of 720p/50. Overall, and with the configurations used in this test, 720p/50 was clearly the most favorable format amongst the three formats under test. However, the authors also believe that the 1080p/50 format has great potential for the future and they encourage further research in this direction.

A further consideration is the impact of the spatial resolution of the original material before encoding and/or down-sampling. Neglecting the low entropy sequence "Ice Dance," the results have given some indication that content derived from 2160p/50 source material provides better results for 720p/50 and 1080i/25 than material generated with the CCD camera. The explanation for this effect can be found in the various areas of CCD camera-related parts (lenses, light, etc.), but also in the principle of spatial over-sampling. From the perspective of the three formats under test (1080p/50, 720p/50 and 1080i/25), the 2160p/50 original material provided an over-sampled source in all cases, while the CCD camera material did so only for 720p/50 (and to a degree also provided temporal over-sampling for 1080i25). We therefore developed an idealized system chain diagram for today's HDTV environments as shown in Figure 15.

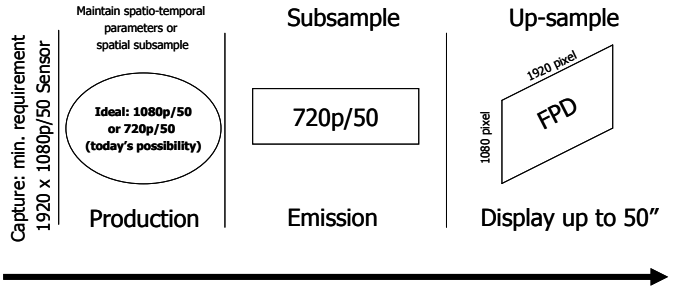

Figure 15 Idealized minimum bit rate system chain to maintain image quality by means of spatial over-sampling at the source format, spatial sub-sampling for transmission purposes, and spatial up-sampling at the display location.

Many of today's HDTV cameras already provide at the sensor point a 1920 x 1080 resolution at 50 frames, but for legacy studio infrastructure reasons (HD-SDI at 1.485 Gbit/s) the output is down-sampled to 1080i/25 or 720p/50. Providing the captured 1080p/50 signal to the wider studio environment would contribute remarkably to the quality of HDTV (i.e. via 3 Gbit/s HD-SDI). Using (in some cases) a 1080p/50 or (in most cases) a 720p/50 format for distribution would then certainly provide a high quality and very economical (bit rate-wise) way to serve displays up to 50 inch diagonal with high quality HDTV signals.

Speculating into the future and maintaining the principles explained above: a 1080p/50 distribution format would require at least a 1080p/50 based production environment but certainly an even higher spatial resolution at the capture point (e.g. 2k sensor). Such an emission format would then be able to serve displays of very high spatial resolution (e.g. 2k) and large size.

A. Impact factors and assistance for further research:

To assist further activities and research in this direction the authors would like to share some feedback factors emerging from the experiments presented in this paper.

On the TSCES method: further experiments may alter the lower anchor to achieve a more equal distribution of votes over the rating scale. It is advised to have at least 20-25 assessors per session.

On content creation and down-sampling methods (HDTV format conversions from 1080p/50 to 1080i/25 and 720p/50, or from 2160p/50 to 1080p/50 in the case of SVT sequences): it would be useful to test different down-sampling filters and to acquire further over-sampled source material (e.g. 2k CCD or CMOS capture). However the filters should be practical for use in cameras.

On content encoding: preferably alternative H.264/AVC encoder implementations and different coding parameters should be used for the encoding of the three formats. In these experiments exclusively the HHI encoder implementation for H.264/AVC was used for encoding the sequences. It would be

beneficial if the sequences could be encoded with other H.264/AVC implementations (perhaps even real-time) and to repeat some TSCES subjective tests.

On displays: to date no large Grade 1 reference matrix display with precise electro-optical transfer characteristics and known deinterlacing performance is available. This may have contributed to the poorer results for the 1080i/25 format. However, the authors believe that the main reason is to be found in the difficulties with the H.264/AVC encoding of 1080i/25 (because it contains only half the vertical-temporal information compared to the progressive formats), and with the interlaced 'footprint' from the content source. Other display technologies should be tested. In fact the authors are currently planning a further test series utilizing large LCD displays.

The authors would welcome cooperation on their research and feedback.

ACKNOWLEDGMENTS

We would like to express our particular thanks to Prof. Hedtke, Dr. Schnoell, Mr. Eichmueller and Mr. Schreiner for their logistical support for the subjective tests at the University of Applied Sciences, Wiesbaden, Germany, and to all participants in the tests. Furthermore, we thank Pioneer for providing the display support and DVS for providing the Pronto2k servers.

References

[1] ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Tech. Rep. ITU-R BT.500-11, 2003, 2003.

[2] K. Bernath, F. Kretz and D. Wood, "The EBU method for organising subjective tests of television picture quality," *EBU Technical Review*, vol. 186, pp. 66-75, 1981.

[3] J. Allnatt, *Transmitted-Picture Assessment.* ,1st ed., vol. 1, Chichester: John Wiley & Sons Ltd., 1983, pp. 301.

[4] I. F. Macdiarmid and P. J. Darby, "Double-Stimulus Assessment of Television Picture Quality," *EBU Technical Review*, vol. 192, pp. 70-79, 1982.

[5] SMPTE 274M-2005, "Television - 1920 x 1080 Image Sample Structure, Digital Representation and Digital Timing Reference Sequences for Multiple Picture Rates," Society of Motion Pictures and Television Engineers, New York.

[6] ITU-R BT.709-5, "Parameter values for the HDTV standards for production and international programme exchange," International Telecommunication Union, Geneva, Tech. Rep. ITU-R BT.709-5, 2002.

[7] SMPTE 296M-2001, " Progressive Image Sample Structure - Analog and Digital Representation and Analog Interface," Society of Motion Pictures and Television Engineers, New York.

[8] ITU-R BT.601-5, "Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect rations," International Telecommunication Union, Geneva, Tech. Rep. ITU-R BT.601-5, 1995.

[9] H. Hoffmann, Dr. T. Itagaki, D. Wood and A. Bock, "Studies on the Bit Rate Requirements for a HDTV Format With 1920 1080 pixel Resolution, Progressive Scanning at 50 Hz Frame Rate Targeting Large Flat Panel Displays," *IEEE Transactions on Broadcasting*, vol. 4, pp. 420-434, Dec. 2006.

[10] H. Hoffmann, T. Itagaki and D. Wood, "New Psycho-physical Method of Television Picture Quality Evaluation (EBU-II)," Electronics Letters, 15 February 2007, Volume 43, Issue 4, p. 212-213

[11] ITU-R BT.814-1. (1994, Specification and alignment procedures for setting of brightness and contrast of displays. International Telecommunication Union, Geneva. [Online]. Available from: http://www.itu.ch

[12] C. Suehring, A. Tourapis and G. Sullivan. (2006, Aug.). H.264/AVC reference software JM11. [Online]. *2006(Sept.)*, Available from: http://iphome.hhi.de/suehring/tml/download/

[13] L. Haglund. (2006, The SVT high definition multi format test set. Swedish Television, Stockholm. [ftp://vqeg.its.bldrdoc.gov].

[14] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, pp. 399-417, 1963.

[15] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Transactions on Accoustics Speech and Signal Proc.*, vol. 36, pp. 1445-1453, Sept. 1988.

[16] ITU-T and ISO/IEC JTC1, "Advanced video coding for generic audiovisual services. ITU-T Rec. H.264 and ISO/IEC 14496-10 (MPEG4-AVC). Version 1: May 2003, Version 2: Jan.2004, Version 3: Sept.2004, Version 5: July 2005,"

[17] T. Wiegand, G. J. Sullivan, G. Bjontegaard and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, pp. 560-576, July. 2003.

[18] G. J. Sullivan and T. Wiegand, "Video compression - from concepts to the H.264/AVC standard," *Proceedings of the IEEE*, vol. 93, pp. 18-31, Jan. 2005.

[19] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Processing*, vol. 15, pp. 74-90, Nov. 1998.

[20] G. J. Sullivan and R. L. Blaker, "Rate-distortion optimized motion compensation for video compression using fixed or variable size blocks," *Proc. GLOBECOM '91*, pp. 85-90, Dec. 1991.

[21] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell and S. K. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 182-190, April. 1996.

[22] T. Wiegand and B. Girod, "Lagrangian multiplier selection in hybrid video coder control," *Proc. ICIP 2001, Thessaloniki*, Oct. 2001.

[23] H. Schwarz, D. Marpe and T. Wiegand, "Hierarchical B pictures," *Joint Video Team Doc JVT-P014,* July. 2005.

[24] R. A. Salmon and J. D. Drewery, "Test of visual acuity to determine the resolution required of a television transmission system," BBC, http://www.bbc.co.uk/rd/pubs/whp/whp-pdf-files/WHP092.pdf, Tech. Rep. WHP 092, 2004.