14

IRT/Tektronix Investigation of Subjective and Objective Picture Quality for 2-10 Mbit/sec MPEG-2 Video: Phase 1 Results

October 6, 1997

A. Schertz
Sound and Picture Processing Group/
Institut fuer Rundfunktechnik (IRT)
email: schertz@irt.de

N. Franzen, J. Lu, M. Ravel
Information Technologies Group/
Tektronix Measurement Division
email: mihir.ravel@tek.com

1. Overview

The IRT (Institut fuer Rundfunktechnik GmbH, of Muenchen/FRG) and Tektronix Inc. (Beaverton, OR/USA) have recently completed the initial phase of an investigation into the performance of an objective picture quality rating (PQR) method based on a human vision model licensed by Tektronix from Sarnoff Labs. In this note we briefly summarize the results of a blind test comparing the Tektronix/Sarnoff PQR Picture Quality Metric and the subjective Mean Opinion Scores (MOS) of viewers. The data set of 60 video scenes used in the experiment was generated by IRT from 5 different video sequences passed through 2 different MPEG-2 encoders at compressed rates of 2, 3, 4.5, 7, and 10 Mbits/second. The MOS scores were determined by IRT and the objective PQR assessments were determined by Tektronix. The subjective scoring procedure used panels of 25 assessors and followed strict Recommendation ITU-R BT.500-7 (DSCQS method) procedures. The objective PQR scores were computed by Tektronix with the Sarnoff Human Vision Model based on Just-Noticeable Difference principles. No model parameters were adjusted to fit the IRT data set. To avoid possible biases in the experiment, the subjective and objective ratings were exchanged by Tektronix and IRT only after each group had completed their scoring. Given the absence of any adjustments to the model parameters, which are based on human vision science, the agreement between subjective and objective results displays a strong correlation of 0.88. Correlation over typical broadcast quality is 0.91. The results are shown in Figure 3, and are promising for the future use of objective methods in the characterization and monitoring of video picture quality.

2. Video Test Set and Processing

The video test scenes were supplied by IRT to Tektronix in SMPTE 125M 422-625/50 Hz format (i.e. PAL D1 tape format). Each scene is of 9 seconds duration. In the following, HRC stands for "Hypothetical Reference Circuit" (as defined by T1A1.5). Before the video was passed through the HRCs, Tektronix added a bar-code near the top of each video frame. This code is used for determining horizontal and vertical pixel misalignment, frame count and other factors. The stripe was covered for the subjective tests, but the results of a test with a small control group and visible stripe showed that the stripe had little effect on viewer evaluations. After the alignment stripes were added, the sequences were passed through the HRCs by IRT. Two video coders (IRT[1] and Thomson) were employed at bitrates of 2.0, 3.0, 4.5 ,7.0 and 10.0 Mbits/s. Although commercial broadcast systems are unlikely to operate below 3 Mbits/sec, the 2.0 Mbits/sec scenes were included to explore performance beyond normal limits. A final set of HRCs consisted of following a PAL conversion stage with the same two coders running at 3 Mbits/s. It is expected that the PAL conversion in particular would likely introduce some subpixel misalignment. The original sequences and their processing into the test scenes are summarized below.

[1] The "IRT coder" was developed by the IRT and several European partners in the framework of the projects Eureka 625 VADI, Race HD-SAT and Race DISTIMA.

Original Sequences	**(HRCs)**	**Bitrates**
1 Barcelona	1 IRT Coder	2.0 Mbits/s
2 Mobile & Calendar	2	3.0
3 NDR	3	4.5
4 Football (Soccer)	4	7.0
5 Flower Garden	5	10.0
	6 Thomson Coder	2.0 Mbits/s
	7	3.0
	8	4.5
	9	7.0
	10	10.0
	11 PAL+MPEG(Thomson),	3.0 Mbits/s
	12 PAL+MPEG(IRT),	3.0 Mbits/s
	13 Reference - no compression	

Total Test Set of **60 Scenes** = (**5** sequences) **x** [(**2** encoders) **x** (**5** bitrates) + **2** PAL]

Barcelona: colorful patterned extravaganza parade formation on a large playing field (see Figure 4). The camera is slowly zooming out and the motion is low. The background stands contributes fine detail. The sequence is: colorful, low motion, fine detail.

Mobile & Calendar: familiar animation sequence used throughout the video compression community. Involves colorful display of animal cartoon figures, toy train in motion, rolling ball and calendar with text detail. The sequence is: colorful, low motion and fine detail.

NDR: radio announcer standing in front of an aggregate stone wall. The wall forms very fine detail, not much color. The camera slowly zooms out. The main challenge to compression is the detail of the stone wall. The motion content is very low. The sequence is: low motion, fine detail.

Football (Soccer in US): soccer game is being played with the camera angle wide. Not much close in action. The motion is characterized as moderate. The first second of video is quite defocused in the original scene. The sequence is: fast motion, fine detail.

Flower Garden: this sequence is widely used in the video compression research community. The camera, in an open vehicle, is moving at moderate speed passing a colorful flower garden. A windmill in motion and persons are in the background. The garden and bare tree limbs provide fine detail. The apparent motion is characterized as moderate. The sequence is: colorful, low motion, fine detail.

In Figure 4 we have included a typical frame image for each of the above sequences.

3. Subjective Evaluation

The Double Stimulus Continuous Quality Scale (DSCQS) Method (ITU-R BT.500-7) was used for the tests.

The presentation structure consisted of the following phase lengths illustrated in Figure 1.

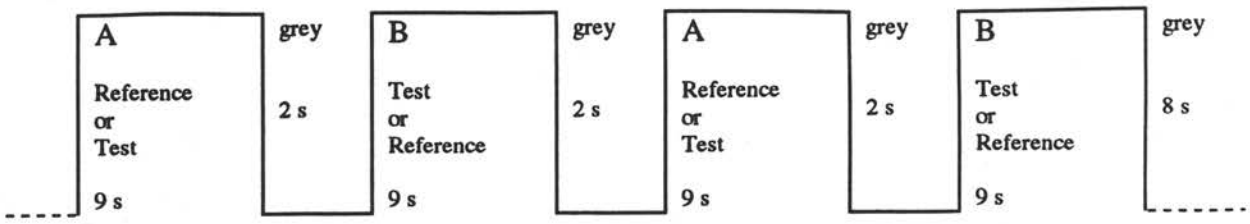


Figure 1. Presentation order for DSCQS method

A was the reference and B the HRC or vice versa, varying from test to test. The order was unknown to the assessors. The overall length of a test was 50 seconds.
For the rating of the test sequences, a test sheet of the following type as shown in Figure 2 was used.

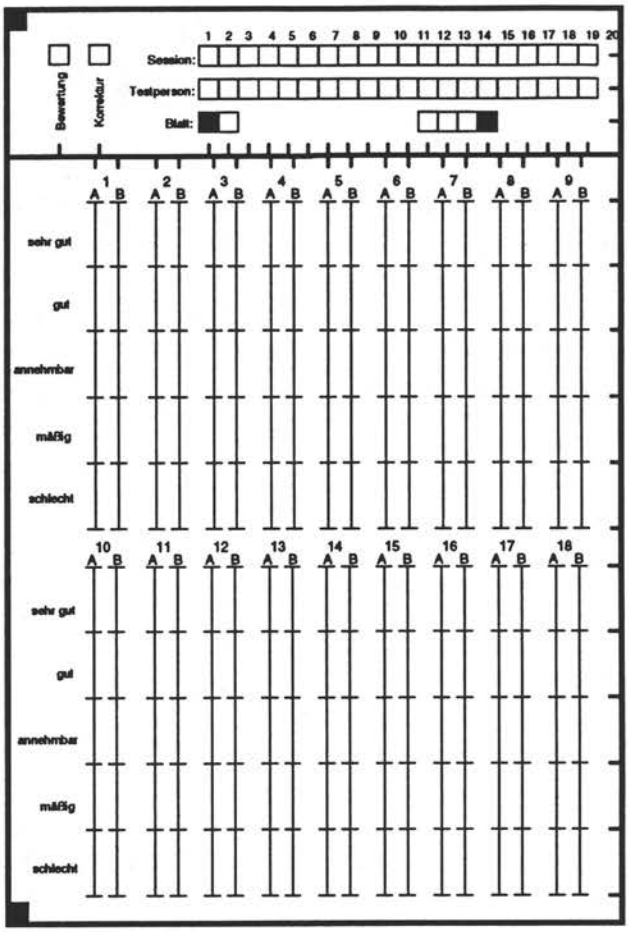


Figure 2. Test sheet used for the assessment of the test sequences

The quality of A and B was indicated by the assessors on a linear scale. The terms of quality on the left side mean: excellent, good, fair, poor, bad. The results were evaluated electronically and the distance between the lower end of the scale and the quality indicator set by the assessor was calculated for each case in millimetres. The difference between the results for reference and HRC was the important result.

In addition to the real test, examples and training sequences were shown. Four examples were shown at the beginning of the first session. They demonstrated the test method and spanned

the quality range to be expected. The viewers were told not assess the sequences because they were only examples. The examples are listed in the following Table 1.

number	test sequence	coder	bitrate / Mbit/s
1	zoom on a street	IRT	3
2	Barcelona 2	Thomson	4
3	zoom on a street	IRT	10
4	Barcelona 2	Thomson	2

Table 1. Example sequences

"Zoom on a street" is a well known BBC production showing a street scene in Edinburgh. Barcelona 2 is a scene from the same production as "Barcelona", but is a close-up of participants.

The training sequences had to be assessed by the subjects who did not know that the results were not evaluated. The training sequences are listed in the following Table 2.

number	test sequence	coder	bitrate / Mbit/s
1	Renata	Thomson	2
2	Table Tennis	IRT	10
3	Renata	Thomson	4
4	Table Tennis	IRT	2
5	Renata	Thomson	10
6	Table Tennis	IRT	4

Table 2. Training Sequences

"Renata" and "Table Tennis" are well known test sequences.

The test sessions were structured in the following way:

Session 1: examples (4) - training sequences (6) - real tests (31)
Session 2: training sequences (6) - real tests (34)

The overall length of session 1 was 34 minutes and 10 seconds, the corresponding time of session 2 was 33 minutes and 20 seconds. 25 assessors took part in the test series, with 15 of them "external" people (housewives, students etc.), and 10 people were members of the IRT staff (non experts). The viewing distance was 6 H (H: picture height). All other conditions were in agreement with ITU-R Rec. BT.500-7. Sony monitors were used.

The bar-code stripes at the top of each picture were covered by dark paper attached to the screen. A test with a small group of five assessors (from IRT staff, non experts) where the stripe was not covered showed that this condition had no significant influence on the results.

The key subjective test results were the mean values (subjective Mean Opinion Scores, MOS) and 95% confidence intervals of the differences between the results for the reference and the HRC. As the whole scale is 100 millimetres long, the worst result is 100, the best one is 0. A result of 20 corresponds to the difference between "excellent" and "good", or between "good" and "fair", etc.

4. Objective Picture Quality Assessment

After the video sequences had been processed by IRT through the HRCs as described above to produce the test set, the PQR objective quality assessments were performed at Tektronix. The process is briefly described as follows:

- video is acquired from D1 tape to computer files for digital processing
- temporal and spatial alignment algorithms are applied to determine misalignments
- the video is then realigned temporally and spatially. For this data set, spatial realignment was performed only to the nearest integral pixel location, hence no interpolation filters were invoked. Temporal alignment is done by frame shifting and does not modify the data in any way.
- the video was then processed with the Tektronix/Sarnoff PQR objective picture quality method. This analysis was carried out by a software version of the quality model running on SUN Sparc workstation. The method generates a frame-by-frame picture quality time history for the full length of video so that continuous quality can be analyzed. For comparison to the subjective assessments, these time histories were condensed into an overall Picture Quality Rating (PQR) for each scene that was a measure of global quality over the duration of the scene.

5. Comparison of Subjective and Objective Assessments

Figure 3 displays the subjective MOS's determined by IRT and the objective PQR's estimated by Tektronix. The vertical error bars display the 95% confidence intervals for the spread in subjective viewer ratings. The relationship between subjective and objective assessments is well behaved and monotonic with a strong correlation of 0.88. From the rightward curvature in the relationship it can be seen that there is a compression in viewer's picture quality assessment as quality degrades towards very poor. This effect is well known in the field of subjective testing, and is consistent with the compression effects found in other areas of human perception such as loudness and brightness. The group of 3 points in the upper right hand corner contains scenes where the encoder either failed catastrophically in regions of the scene or the quality was very poor. If these points are excluded then the correlation coefficient increases to 0.91. Given that the objective quality ratings did not require any fitting or optimization of parameters to the test data set, the results are quite encouraging that objective methods will contribute to reducing the time, expense, and possible biases associated with characterization and monitoring of video.

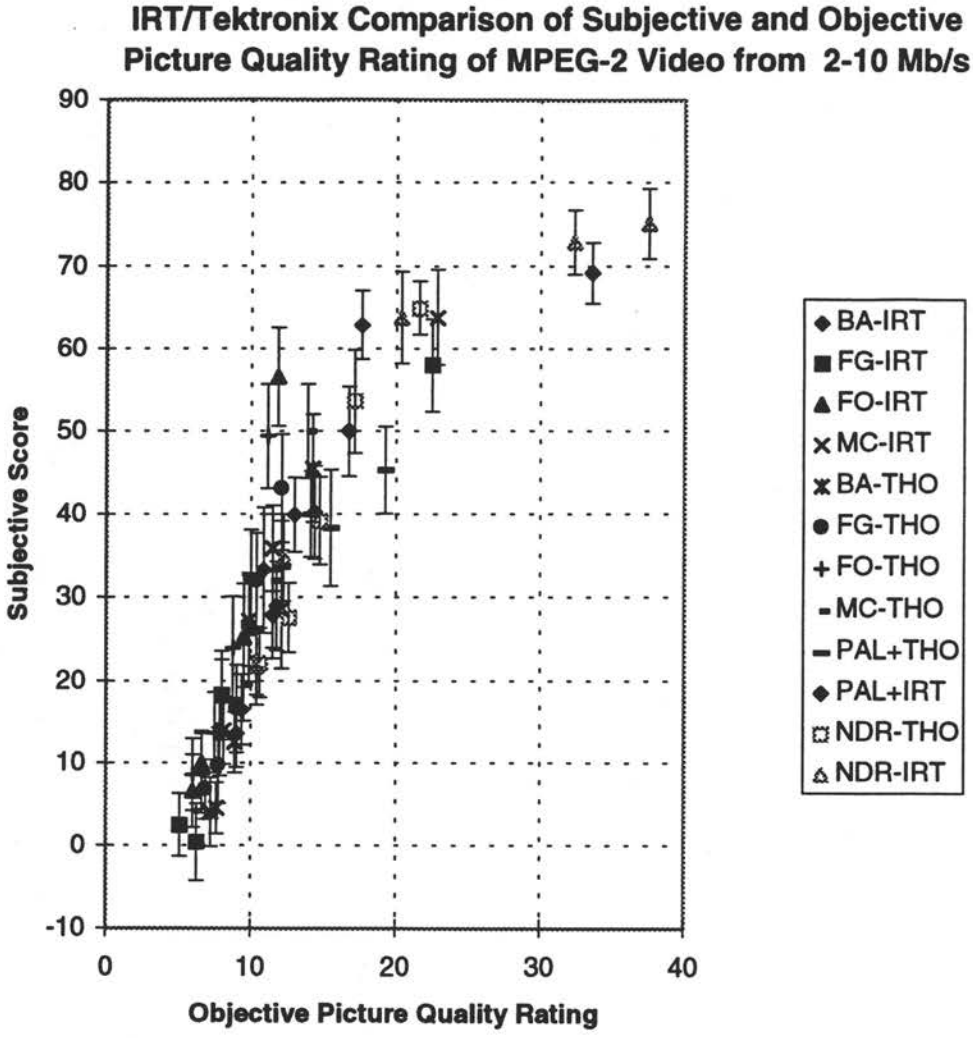
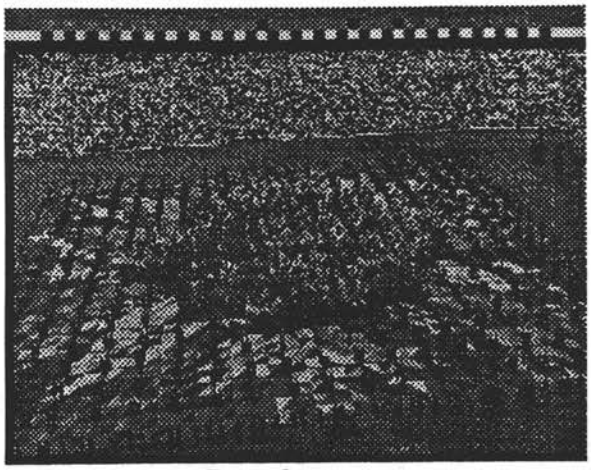
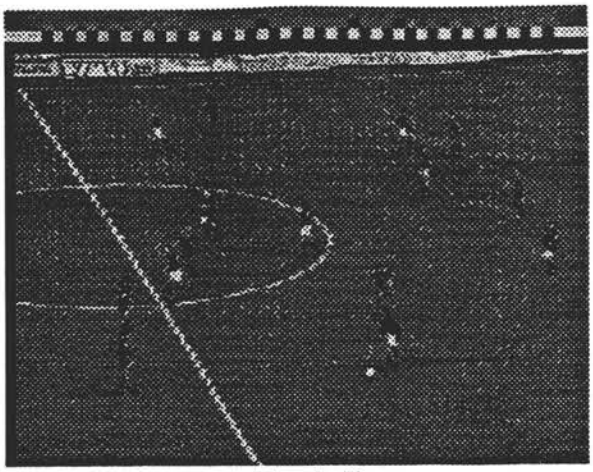


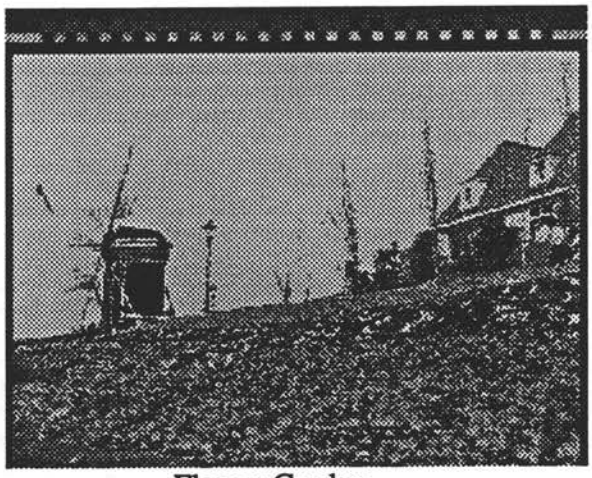
Figure 3. Comparison of IRT Subjective Mean Opinion Scores (MOS's) and Tektronix/Sarnoff Objective Picture Quality Rating (PQR) for 60 2-10 Mbits/sec MPEG-2 and PAL test scenes. The 95% confidence intervals for subjective scores are indicated by vertical bars. Correlation between objective and subjective ratings is 0.88 for the complete data set, and viewer compression in quality rating is apparent for upper right poorest qualty scenes. The correlation is 0.91 if upper rightmost data scenes of poorest quality are excluded.



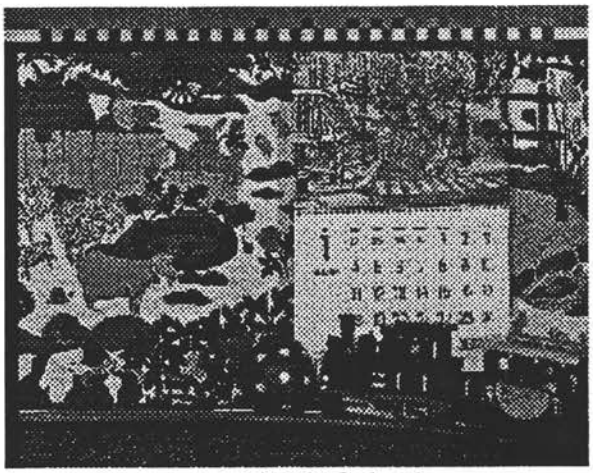
Barcelona



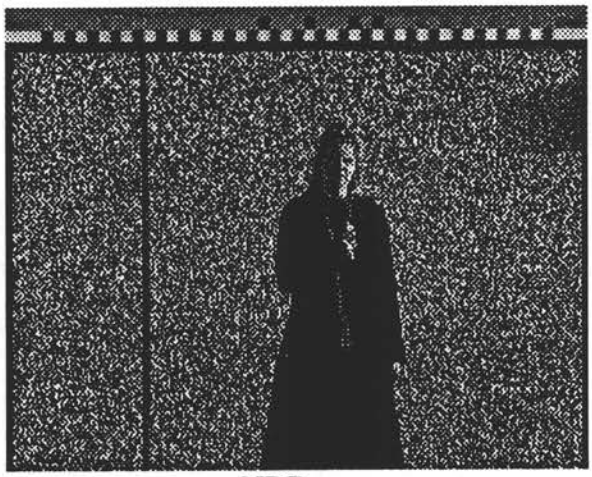
Football



Flower Garden



Mobile & Calendar



NDR

Figure 4. Typical frame images of video test sequences