VIDEO QUALITY EXPERTS GROUP

Report
on the Validation of Video Quality Models
for High Definition Video Content

Version 2.0, June 30, 2010

For more information contact:
Arthur Webster webster@its.bldrdoc.gov Co-Chair VQEG
Filippo Speranza Filippo.Speranza@crc.gc.ca Co-Chair VQEG

Regarding the Use of HDTV data:

Subjective data, objective model validation data, and model analyses are published in this report.

The source and processed video sequences for experiments vqeghd1, vqeghd2, vqeghd3, vqeghd4, and vqeghd5 have been approved for redistribution and use in research experiments. Proper approval must be obtained from the copyright holders of the source video sequences. To obtain approval for access to the source video sequences, the Content User Agreement form available from the Consumer Digital Video Library (www.cdvl.org) must be completed. The source and processed video sequences for experiment vqeghd6 is not available for redistribution and may only be obtained with permission from FUB.

Appropriate uses for VQEG HDTV Phase I subjective data, objective data, video material, and analyses include:

- Subjective data and video material may be used to train new objective video quality models
- The VQEG HDTV statistics and analyses may be included in another paper
- Objective data and video material may be used to confirm the performance of a model mentioned in this report.
- Additional experiments may be performed using this video material and subjective data

Inappropriate uses for VQEG HDTV Phase I subjective data, objective data, video material, and analyses include the following:

- Proposing a model for standardization, based upon use of the VQEG HDTV Phase I datasets, of any model not mentioned in this report is not permitted.
- Use of the video material in a commercial application is not permitted (e.g., product brochure, customer demonstration).
- It is not allowed to claim that a model not mentioned in this report has superior performance to the models mentioned in this report, based upon the use of this dataset.
- Models that are trained on these datasets must not be compared to the models submitted to VQEG for independent validation in 2009. Such a comparison is misleading, because the experiments contain mainly source scenes and HRCs that were unknown to the model developers. Additionally, this comparison is misleading because the sixth dataset has been kept private.

Publications resulting from any use of the VQEG HDTV data, analyses, or video material must:

- Mention the VQEG Final Report
- Respect the copyright holders' usage limitations on appropriate uses of the source video
- State clearly that the model was trained on this video material, where appropriate.

Acknowledgments
This report is the product of efforts made by many people over the past two years. It will be impossible to acknowledge all of them here but the efforts made by individuals listed below at dozens of laboratories worldwide contributed to the report.

Editing Committee:
Margaret Pinson, NTIA (USA)
Filippo Speranza, CRC (Canada)

List of Authors:
Akira Takahashi, NTT (Japan)
Christian Schmidmer, Opticom (Germany)
Chulhee Lee, Yonsei University (Korea)
Filippo Speranza, CRC (Canada)
Jun Okamoto, NTT (Japan)
Kjell Brunnström, Acreo (Sweden)
Lucjan Janowski, AGH University (Poland)
Marcus Barkowsky, IRCCyN (France)
Margaret Pinson, NTIA (USA)
Nicolas Staelens, Ghent University – IBBT (Belgium)
Quan Huynh Thu, Psytechnics (UK)
Rima Green, Tektronix (USA)
Roland Bitto, Opticom (Germany)
Ron Renaud, CRC (Canada)
Silvio Borer, Swissqual (Switzerland)
Taichi Kawano, NTT (Japan)
Vittorio Baroncini, FUB (Italy)
Yves Dhondt, Ghent University – IBBT (Belgium)

Important Technical Contributors:
Alexander Raake, Deutsche Telekom AG, Laboratories (Germany)
Audry Younkin, Intel (USA)
Arthur Webster, NTIA (USA)
Fabrice Mougin, Witbe (France)
Greg Cermak, Verizon (USA)
Jean-Michel Planche, Witbe (France)
Jörgen Gustafsson, Ericsson (Sweden)
Marie-Neige Garcia, Deutsche Telekom AG, Laboratories (Germany)
Martin Pettersson, Ericsson (Sweden)
Mathiew Carnec, AccepTV (France)
Mikolaj Leszczuk, AGH University (Poland)
Osamu Sugimoto, KDDI (Japan)
Patrick Le Callet, IRCCyN (France)
Paul Rolland, Witbe (France)
Phil Corriveau, Intel (USA)
Piotr Romaniak, AGH University (Poland)
Romuald Pépion, IRCCyN (France)
Stefan Winkler, Cheetah Technologies (USA)
Stephen Wolf, NTIA (USA)

Table of Contents

Contents

EXECUTIVE SUMMARY

FINAL REPORT FROM THE VIDEO QUALITY EXPERTS GROUP ON THE VALIDATION OF OBJECTIVE MODELS OF HDTV QUALITY ASSESSMENT, PHASE I

This document presents results from the Video Quality Experts Group (VQEG) HDTV validation testing of objective video quality models. This document provides input to the relevant standardization bodies responsible for producing international Recommendations and regional Standards.

The High Definition Television (HDTV) Test contains two parallel evaluations of test video material. One evaluation is by panels of human observers (i.e., subjective testing). The other is by objective computational models of video quality (i.e., proponent models). The objective models are meant to predict the subjective judgments. Each subjective test will be referred to as an "experiment" throughout this document.

This HDTV Test addresses four video formats directly (1080p at 25 and 29.97 frames-per-second, and 1080i at 50 and 59.94 fields-per second) and two video formats indirectly (720p at 50 and 59.94 frames-per-second). This HDTV Test addressed three types of models: full reference (FR), reduced reference (RR), and no reference (NR). FR models have full access to the source video; RR models have limited bandwidth access to the source video; and NR models do not have access to the source video[1].

Six subjective experiments provided data against which model validation was performed. The experiments were divided between the four 1080 video formats. 720p was inserted into experiments as a test condition, for example by converting 1080i 59.94 fields-per-second video to 720p 59.94 frames-per-second, compressing the video, and then converting back to 1080i. A common set of carefully chosen video sequences were inserted identically into each experiment, to anchor the video experiments to one another and assist in comparisons between the subjective experiments. These common sequences were used to map the six experiments onto a single scale (called the "aggregated superset" in this report). The subjective experiments included processed video sequences with a wide range of quality. The impairments examined were restricted to MPEG-2 and H.264, both coding only and coding plus transmission errors.

A total of 12 independent testing laboratories coordinated to perform subjective testing (AGH University, Psytechnics, NTIA/ITS, Ghent University – IBBT, Verizon, Intel, FUB, CRC, Acreo, Ericsson, IRCCyN, and Deutsch Telekom AG Laboratories). Objective models were submitted after the six secret experiments were near completion (e.g., after scene selection, PVS generation, and most of the subjective testing) to allow proponents the best opportunity to improve their model. 14 models were submitted, 6 were withdrawn, and 8 are presented in this report.

Results for models submitted by the following five proponent organizations are included in this HDTV Final Report:

- NTT (Japan)
  - FR model – NTT_QE_HD
- OPTICOM (Germany)
  - FR model – PEVQ-HD Special Build 3.4
- SwissQual (Switzerland)
  - FR model – VQuad-HD
- Tektronix (USA)
  - FR model – VQEG.bat. Version 2.5.93
- Yonsei University (Korea)
  - FR model – Yonsei HDFR
  - RR models – Yonsei_HDRR56k, Yonsei_HDRR128k & Yonsei_HDRR256k

[1] All NR models were withdrawn.

The HDTV data may not be used as evidence to standardize any other objective video quality model that was not tested within this phase. This comparison would not be fair, because another model could have been trained on the HDTV data.

The intention of VQEG is to make five of the six HDTV subjective datasets available to other researchers.

MODEL PERFORMANCE EVALUATION TECHNIQUES

The models were evaluated using two statistics that provide insights into model performance: Root-Mean Squared Error (RMSE) and Pearson Correlation. Each model was fitted to each subjective experiment and the aggregated superset, by optimizing Pearson Correlation with subjective data first, and minimizing RMSE second. RMSE is considered the primary metric for analysis in this report. Thus, RMSE is used to determine whether a model is in the group of top performing models for one video format/resolution (i.e. a group of models that include the top performing model and models that are statistically equivalent to the top performing model).

When examining the total number of times a model is statistically equivalent to the top performing model, comparisons between models should be performed carefully. Determining which differences in totals are statistically significant requires additional analysis that is not available. As a general guideline, small differences in these totals do not indicate an overall difference in performance. This refers to the tables below.

PSNR was computed as a reference measure, and compared to all models. PSNR was computed using an exhaustive search for calibration and one constant delay for each video sequence. PSNR was calculated according to ITU-T Draft Rec. J.340, which included temporal and spatial calibration. However, to save computation time, the luminance gain & offset calculation for PSNR were calculated separately and input to the PSNR algorithm as constants, and an appropriate search range was chosen for each dataset. Models were required to perform their own calibration, where needed.

FR MODEL PERFORMANCE

FR model results from NTT, OPTICOM, Swissqual, Tektronix, and Yonsei are included in this report.

Primary Analysis of FR Models

The performance of each FR model is summarized in the table below. “Superset RMSE” identifies the primary metric (RMSE) computed on the aggregated superset (i.e., all six experiments mapped onto a single scale). “Top Performing Group Total” identifies the number of experiments (0 to 6) for which this model was either the top performing model or statistically equivalent to the top performing model. “Better Than PSNR Total” identifies the number of experiments (0 to 6) for which the model was statistically better than PSNR. “Better Than Superset PSNR” lists whether each model is statistically better than PSNR on the aggregated superset. “Superset Correlation” identifies the Pearson Correlation computed on the aggregated superset.

Metric	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
Superset RMSE	0.71	0.74	0.88	0.56	0.65	0.74
Top Performing Group Total	1	0	0	5	3	1
Better Than PSNR Total	0	0	0	4	4	1
Better Than Superset PSNR	No	No	No	Yes	No	No
Superset Correlation	0.78	0.76	0.63	0.87	0.82	0.76

The body of this report includes other metrics including Pearson Correlation & RMSE calculated on individual experiments, confidence intervals, statistical significance testing on individual experiments, analysis on subsets of the data that include specific impairments (e.g., H.264 coding-only), scatter plots, and the fit coefficients for each model.

FR Model Conclusions

- VQEG believes that at least one FR model performed well enough to be included in normative sections of Recommendations.
- The scope of these Recommendations should be written carefully to ensure that the use of the models is defined appropriately.
- If the scope of these Recommendations includes video system comparisons (e.g., comparing two codecs), then the Recommendation should include instructions indicating how to perform an accurate comparison.
- None of the evaluated models reached the accuracy of the normative subjective testing.

RR MODEL PERFORMANCE

RR models were submitted by Yonsei for the following bit-rates: 56 kbits/s, 128 kbits/s, and 256 kbits/s. When comparing these RR models to PSNR, it must be noted that PSNR is an FR model (i.e., PSNR needs full access to the source video). Thus, equivalence to PSNR indicates that the RR model showed good performance while using a lower bandwidth.

Primary Analysis of RR Models

The performance of each RR model is summarized in the table below. "Superset RMSE" identifies the primary metric (RMSE) computed on the aggregated superset (i.e., all six experiments mapped onto a single scale). "Top Performing Group Total" identifies the number of experiments (0 to 6) for which this model was either the top performing model or statistically equivalent to the top performing model. "Equivalent To or Better Than PSNR Total" identifies the number of experiments (0 to 6) for which the model was statistically equivalent to or better than PSNR. "Equivalent To Superset PSNR" lists whether each model is statistically equivalent to PSNR on the aggregated superset. "Superset Correlation" identifies the Pearson Correlation computed on the aggregated superset.

Metric	PSNR	Yonsei56k	Yonsei128k	Yonsei256k
Superset RMSE	0.71	0.73	0.73	0.73
Top Performing Group Total	6	4	4	4
Equivalent To or Better Than PSNR Total	6	4	4	4
Equivalent To Superset PSNR	Yes	Yes	Yes	Yes
Superset Correlation	0.78	0.77	0.77	0.77

The body of this report includes other metrics including Pearson Correlation & RMSE calculated on individual experiments, confidence intervals, statistical significance testing on individual experiments, analysis on subsets of the data that include specific impairments (e.g., H.264 coding-only), scatter plots, and the fit coefficients for each model.

RR Model Conclusions

- VQEG believes that some of the RR models may be considered for standardization making sure that the scopes of these Recommendations are written carefully to ensure that the use of the models is defined appropriately.
- If the scope of these Recommendations includes video system comparisons (e.g., comparing two codecs), then the Recommendation should include instructions indicating how to perform an accurate comparison.
- None of the evaluated models reached the accuracy of the normative subjective testing.
- All of the RR models performed statistically equivalent to or better than PSNR. It must be noted that PSNR is a FR model requiring full access to the source video.

NR MODEL PERFORMANCE

All NR models submitted to VQEG for validation were withdrawn.

1. LIST OF ACRONYMS

ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating Hidden Reference
AVC	Advanced Video Coding
AVI	Audio Video Interleave
Cb	Chroma blue
CBR	Constant Bit Rate
Cr	Chroma red
CI	Confidence Interval
CODEC	COder-DECoder
CRC	Communications Research Center (Canada)
DMOS	Difference Mean Opinion Score (as defined by ITU-R)
DMOSp	Difference Mean Opinion Score, predicted
DSCQS	Double Stimulus Continuous Quality Scale
DSIS	Double Stimulus Impairment Scale
FPS	Frames per second
FR	Full Reference
HD	High Definition (television)
HDTV	High Definition Television
HRC	Hypothetical Reference Circuit
HVS	Human Visual System
ILG	Independent Laboratory Group
IPTV	Internet Protocol Television
ITU	International Telecommunications Union
ITU-R	ITU Radiocommunications Standardization Sector
ITU-T	ITU Telecommunications Standardization Sector
KDDI	Combined company formed from KDD and IDO Corporation
LCD	Liquid Crystal Display
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
MPEG	Moving Picture Experts Group
NR	No (or Zero) Reference
NTT	Nippon Telegraph and Telephone
OS	Opinion Score – single subject answer
PLR	Packet Loss Ratio
PSNR	Peak Signal to Noise Ratio
PVS	Processed Video Sequence
RMSE	Root Mean Square Error
RR	Reduced Reference
SRC	Source Reference Channel or Circuit
SROI	Spatial Region of Interest
VBR	Variable Bit Rate
VQEG	Video Quality Experts Group
VQR	Video Quality Rating (as predicted by an objective model)

2. INTRODUCTION

The main purpose of the Video Quality Experts Group (VQEG) is to provide input to the relevant standardization bodies responsible for producing international Recommendations regarding the definition of an objective Video Quality Metric in the digital domain. To this end, VQEG initiated a program of work to validate objective quality models that may be applied to measure the perceptual quality of High Definition services.

For the purposes of this document, HDTV is defined as being of or relating to an application that creates or consumes High Definition television video format that is digitally transmitted over a communication channel. Common applications of HDTV that are appropriate to this study include television broadcasting, video-on-demand and satellite and cable transmissions. The measurement tools recommended by the HDTV group will be used to measure quality both in laboratory conditions using a full reference (FR) method and in operational conditions using reduced reference (RR) or no-reference (NR) methods.

This document describes the evaluation tests of the performance of objective perceptual quality models conducted by the Video Quality Experts Group (VQEG). It describes the roles and responsibilities of the model proponents and of the Independent Lab Group (ILG). The text is based on discussions and decisions from meetings of the VQEG HDTV working group at the periodic face-to-face meetings as well as on conference calls and in email discussion. All aspects of the subjective testing were performed by the independent lab group (ILG) in secret. The HDTV ILG included the following organizations: Acreo, AGH University, CRC, Deutsche Telekom AG Laboratories, Ericsson, FUB, Ghent University – IBBT, IRCCyN, Intel, NTIA, Psytechnics, and Verizon.

The goal of the HDTV project was to analyze the performance of models suitable for application to digital video quality measurement in HDTV applications. A secondary goal of the HDTV project was to develop HDTV subjective datasets that may be used to improve HDTV objective models. The performance of objective models with HD signals was determined from a comparison of viewer ratings of a range of video sample quality obtained in controlled subjective tests and the quality predictions from the submitted models. The method selected for the subjective testing was the Absolute Category Rating with Hidden Reference (ACR-HR).

14 models were submitted, 6 were withdrawn, and 8 are reported on in this report. This report analyzes the following models:

Proponent	**Video Resolution & Bit-rate**	**Model Name**
NTT (Japan)	FR	NTT_QE_HD
Opticom (Germany)	FR	PEVQ-HD Special Build 3.4
SwissQual (Switzerland)	FR	VQuad-HD
Tektronix (USA)	FR	VQEG.bat. Version 2.5.93
Yonsei University (Korea)	FR	Yonsei HDFR
	RR 56k	Yonsei_HDRR56k
	RR 128k	Yonsei_HDRR128k
	RR 256k	Yonsei_HDRR256k

To fully characterize the performance of the models, a full range of representative transmission and display conditions were examined. To this end, the test cases (hypothetical reference circuits or HRCs) simulated the range of potential behavior of cable, satellite, and terrestrial transmission networks and broadband communications services. Video-only test conditions were limited to secondary distribution of MPEG-2 and H.264 coding, both coding-only and with transmission errors. Both digital and analog impairments were considered. The recommendation(s) resulting from this work should be appropriate for services delivered on high definition displays computer desktop monitors, and high definition display television technologies.

Display formats that addressed in these tests were: 1080i at 50 and 60 Hz; and 1080p at 25 and 30 fps That is, all sources were 1080p or 1080i, including upscaled 720p or 1366x768 as well as 1080p 24fps content that had been rate-converted.

Each subjective experiment contained a common set of 24 video sequences. These common sequences spanned the range of quality desired, and served to provide consistency between experiments. The subjective data for the common set sequences were used to map the six tests onto a single scale.

Once all subjective testing was near completion, proponents submitted objective models. Proponents were able to submit for evaluation Full Reference (FR), Reduced Reference (RR), and No Reference (NR) models. The side-channels allowable for the RR models were: 56 kbps, 128 kbps, and 256 kbps. Proponents could submit one model of each type. PSNR results are presented for comparison purposes.

Of the six experiments conducted by the ILG, five will be made publically available, and the sixth will be kept private. **The intention of VQEG is that the HDTV Phase I data may not be used as evidence to standardize any objective video quality model which was not tested within this phase. This comparison would not be fair, because another model could have been trained on the HDTV Phase I data.**

This final report summarizes the results and conclusions of the analysis along with recommendations for the use of objective perceptual quality models for each HDTV format.

3. TEST LABORATORIES

3.1. ILG Independent Laboratory Group (ILG)

The independent lab group (ILG) had the role of independent arbitrator for the HDTV test. The ILG performed six subjective tests. For these tests, the ILG was the sole responsible for all aspects related to scene choice, HRC choice, and the design of each subjective test. The ILG's subjective datasets were held secret prior to model & subjective dataset submission. The ILG also validated proponent models and performed the official data analysis. The members of the ILG were:

Table 1. Independent Laboratories Group (ILG)

Acreo, Sweden	www.acreo.se
AGH University, Poland	www.agh.edu.pl/ www.kt.agh.edu.pl qoe.kt.agh.edu.pl
Communications Research Centre (CRC), Canada	www.crc.ca
Deutsche Telekom AG Laboratories, Germany	www.laboratories.telekom.com www.aipa.tu-berlin.de
Ericsson, Sweden	www.ericsson.com
Fondazione Ugo Bordoni (FUB), Italy	www.fub.it/it/areas/audiovideosignalprocessing/attivitaeprogetti2010
Ghent University – IBBT	http://www.ibbt.be, http://multimedialab.elis.ugent.be, http://ibcn.intec.ugent.be
Intel, USA	www.intel.com
IRCCyN, University of Nantes, France,	www2.irccyn.ec-nantes.fr/ivcdb
National Telecommunications and Information Administration (NTIA), USA,	www.its.bldrdoc.gov/n3/video
Psytechnics, UK	www.psytechnics.com
Verizon Laboratories, USA	www.verizon.com

3.2. Proponent Laboratories

The proponents submitted one or more models to the ILG for validation. However, proponents were limited to only one model for each type: FR, RR, and NR. Proponents were responsible for running their own model on all video sequences, and submitting the resulting objective data for validation. Proponents paid a fee to the ILG laboratories performing the subjective experiments to cover basic costs of those experiments. The list of proponents whose models are included in this report are:

Table 2. Proponents

NTT (Japan)	www.ntt.co.jp/qos/eng/index.html
Opticom (Germany)	www.opticiom.de
SwissQual (Switzerland)	http://www.swissqual.com/
Tektronix (USA)	www.tektronix.com
Yonsei University (Korea)	http://web.yonsei.ac.kr/hdsp/en

NOTE: After model submission, proponents were allowed to run alternate sets of viewers using settings different from those used by the ILG laboratories.

NTT ran subjects using a professional quality monitor to compare data with two experiments where the ILG had used a high-end consumer grade monitor. Analyses of the NTT subjective data are included in this HDTV Final Report in Appendix VI, “Expansion of Scope to Include CRT Monitors.” NTT received the video material and conducted this experiment after submitting their model.

4. DESIGN OVERVIEW: SUBJECTIVE EVALUATION PROCEDURE

This section provides an overview of the methodology used in the VQEG HDTV test to perform subjective testing. For full details of the test procedure used in the VQEG HDTV test, the interested reader is referred to the official test plan, which can be obtained available from: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/. Six subjective tests were performed by the ILG. The tests assessed the subjective quality of video material presented on a variety of display technologies in a simulated viewing environment. The display resolution, however, was 1920 X 1080 in all tests.

4.1. Viewers

Each subjective experiment collected valid data from 24 participants. A statistical criterion was used to verify that the data from a viewer were correlated to the average of the other viewers' data (see Appendix V: Method for post-experiment screening of subjects). Viewers whose data did not satisfy that criterion were rejected and substituted with new viewers.

All viewers were screened prior to participation for normal (20/30) visual acuity with or without corrective glasses (per Snellen test or equivalent) and normal color vision (per Ishihara test or equivalent).

4.2. Video Material: Specific and Common set

In each of the six experiments, the video material consisted of 168 video samples, which included the processed video sequences (PVS) and their corresponding unprocessed ("reference") video sequence (SRC). The duration of each video sample (either PVS or SRC) was of 10 seconds.

For each experiment, the video samples were functionally divided into two subsets: a specific set and a common set. The first set consisted of PVS and SRC specific to that experiment. This specific set always included 9 original SRC plus all the PVS obtained by processing the nine SRC with 15 different HRC. Thus the total of PVS included in the specific set was 144 (9 + 9*15). The second set consisted of PVS and SRC that common to other experiments. This common set included 4 original SRC plus all the PVS obtained by processing the four SRC with 5 different HRC. Thus the total of PVS included in the 'common set' was 24 (4 + 4*5).

The Common Video Sequences were all filmed and processed in 1080p 25fps. The original plan to use 1080p 24fps SRC was dropped because insufficient 1080p 24fps content existed. The 1080p 25fps was converted to 1080i 50fps by copying without any changes (i.e., playing 1080p 25fps on a 1080i 50fps channel). The 1080p 25fps was converted to 1080p 29.97fps by repeating every 5th frame. The 1080p 25fps was converted to 1080i 59.94fps by first slowing down to 1080p 24fps (i.e., a speed reduction of 1/24) then using a hardware frame-rate converter to perform 3-2 pull down.

4.3. Subjective Test Methodology

The ACR Method with Hidden Reference. The subjective picture quality of the video samples was assessed, in all six subjective tests, using the absolute category rating scale (ACR) method [ITU-T Rec. P.910]. The ACR method is a single stimulus method in which the video samples are presented one at a time, as shown in Figure 1, and rated independently using the five-grade video quality scale shown in Figure 2. Note that the numerical values attached to each category in Figure 2 were used for data analysis only and they were not shown to the viewers. During the data analysis the ACR scores given to the processed versions were subtracted from the ACR scores given to the corresponding reference to obtain a DMOS. This procedure is known as "hidden reference" (henceforth referred to as ACR-HR). This choice was made due to the fact that ACR provides a reliable and standardized method that allows a large number of test conditions to be assessed in any single test session.

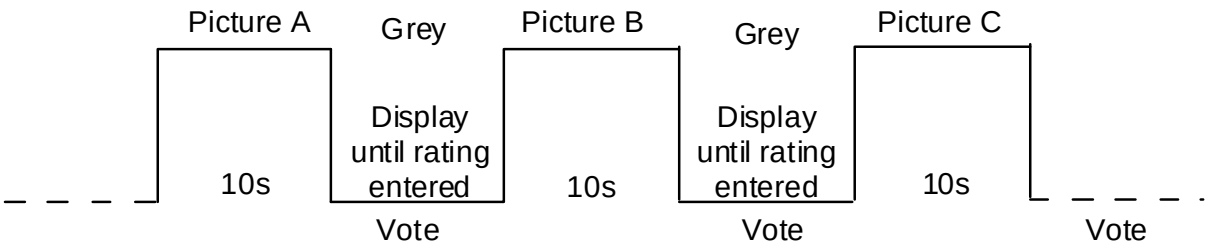


Figure 1 – ACR basic presentation structure.

5 Excellent
4 Good
3 Fair
2 Poor
1 Bad

Figure 2 – The ACR rating scale.

Test Control Software. Some experiments were implemented and controlled by computer software. For these, the viewers performed the experiment using custom made software provided by Acreo. The software controlled both the timing and order of presentation of the stimuli. The custom software also collected and de-randomized the viewers' ratings. Other experiments used alternative methods for displaying the video to viewers, such as Blu Ray Disc playback.

Randomization. The order of presentation of the video samples was changed randomly for different groups of viewers. Laboratories were instructed to use a minimum of two randomized viewer orderings per experiment.

Instructions to viewers. All laboratories were asked to provide instructions which followed agreed upon guidelines to ensure consistency across subjective experiments.
To familiarize the viewers with the assessment tasks and with the levels of video qualities used in the experiment, a small number of practice trials were administered at the beginning of the experimental session. To control the effects of fatigues, a short break was given after about half of the video sample had been assessed. Accordingly, in this scenario, each experiment included the following steps:

1. Introduction and instructions to viewer.
2. Practice clips: these test clips allow the viewer to familiarize with the assessment procedure and software. They represented the range of distortions found in the experiment. Ratings given to practice clips were not used for data analysis.
3. Assessment of first half of the video samples.
4. Short break.
5. Practice clips (this step was optional but advised to regain viewer's concentration after the break).
6. Assessment of second half of the video samples.

4.4. Experimental Settings

The test room conformed to Recommendation ITU-R BT.500-11 for all laboratories. In general, a test session involved only one viewer per display assessing the test material. Viewers were seated directly in line with the center of the video display at a viewing distance equal to three times the height of the picture (i.e., 3H) in all six experiments.

4.5. Display Specifications

All subjective experiments used either high-end consumer TV (Full HD) or professional grade LCD monitors. The basic details of the monitors used for the six experiments are reported in Table 3.

Table 3 – Monitor Characteristics

Model	vqeghd1	vqeghd2	vqeghd3	vqeghd4	vqeghd5	vqeghd6
Diagonal Size	47”	40”	40”	42"	24”	30”
Dot Pitch	0.5419 mm	0.461 mm	0.46 mm	0,484 mm (calculated)	0.270 mm	0.2505 x 0.2505 mm
Resolution (Native)	1920x1080	1920x1080	1920x1080	1920 x 1080	1920x1200	Signal 1: 2560 × 1600 (16:10 aspect ratio) Signal 2: 1920 x 1200 (16:10 aspect ratio)
Gray to Gray Response Time	Black-White response time typically 6.5 ms, max 12ms	–	Blur Edge Time, average 28ms	3 ms	6 ms	6 ms
Color Temperature	6500°K	6500°K	6313°K	Cool/Medium/Warm	6500°K	6500°K
Calibration	GretagMacbeth Eye One Display 2	Ok	Spyder 3 Elite 3.0.4 Gamma ≈ 2.2	Yes	Spyder Pro2	EYE ONE – Display 2
Calibration Method	Calibration with EyeOne	I1 pro (xrite)	Luminance & color measurement	Following ITU-R BT.500-11	Restore the monitor to factory settings, and run the calibration unity of the SpyderPro2 device	Color Navigator
Bit Depth	8 bits/color	24	8 bits/color	10 (R,G,B)	8 bits	Look-up table: 12 bits per color Internal processing: 16 bits per color
Refresh Rate	60 Hz	60	60 Hz	100Hz TruMotion	50 Hz	60 Hz
Label	TCO’03	–	Samsung LE40A796R2MXZF	LG 42LH7000 Television	TCO03	TCO’03

5. Limitations on SRCs, HRCs and calibration:

5.1. Limitation on SRCs

5.1.1. Selection of Source Sequences (SRC)

The following video formats were tested:

- 1080i 60 Hz (30 fps) Japan, Korea, US
- 1080p (25 fps) Europe
- 1080i 50 Hz (25 fps) Europe
- 1080p (30 fps) Japan, US

5.1.2. Requirements for Camera and SRC Quality

The source video was used in the testing if an expert in the field considers the quality to be good or excellent on an ACR-scale. The source video should have no visible coding artifacts. 1080i footage may be de-interlaced and then used as SRC in a 1080p experiment. 1080p enlarged from 720p or 1080i enlarged from 1366x768 or similar are valid HDTV source. 1080p 24fps film footage can be converted and used in any 1080i or 1080p experiment. The frame rate of the unconverted source must be at least as high as the target SRC (e.g., 720p 50fps can be converted and used in a 1080i 50fps experiment, but 720p 29.97fps cannot be converted and used in a 1080i 59.94fps experiment).
At least ½ of the SRC in each experiment must have been shot originally at that experiment's target resolution (e.g., not de-interlaced, not enlarged).

5.1.3. Content

The source sequences are representative of a range of content and applications. The list below identifies the types of test material that form the basis for selection of sequences.

1) movies, movie trailers
2) sports
3) music video
4) advertisement
5) animation
6) broadcasting news (business and current events)
7) home video
8) general TV material (e.g., documentary, sitcom, serial television shows)

5.1.4. Scene Cuts

Scene cuts shall occur at a frequency that is typical for each content category.

5.1.5. Scene Duration

Final source sequences will 10 seconds. Source scenes used for HRC creation will typically use extra 2s content at the beginning and end.

5.1.6. Source Scene Selection Criteria

Source video sequences selected for each test used the following criteria:

1. All source must have the same frame rates (25fps or 30fps).
2. Either all source must be interlaced; or all source must be progressive.
3. At least one scene must be very difficult to code.
4. At least one scene must be very easy to code.
5. At least one scene must contain high spatial detail.
6. At least one scene must contain high motion and/or rapid scene cuts (e.g., an object or the background moves 50+ pixels from one frame to the next).
7. If possible, one scene should have multiple objects moving in a random, unpredictable manner.

8. At least one scene must be very colorful.
9. If possible, one scene should contain some animation or animation overlay (e.g., cartoon, scrolling text).
10. If possible, at least one scene should contain low contrast (e.g., soft or blurred edges).
11. If possible, at least one scene should contain high contrast (e.g., hard or clearly focused edges, such as the SMPTE birches scene).
12. If possible, at least one scene should contain low brightness (e.g., dim lighting, mostly dark).
13. If possible, at least one scene should contain high brightness (e.g., predominantly white or nearly white).

5.1.7. Scene Pools Overview

Initial scene pools were developed by one lab and then reviewed by a panel of ILG. The ILG panel ensured that the quality of each sequence was good or excellent on an ACR-scale, by replacing low quality SRC. Each pool was designed to have a sufficient range of source material and all six scene pools were balanced so that individual SRCs are not over-used. Scene pool 1 had mostly known source material. The other five scene pools contained a high percentage of secret source material.
All original SRC were 14 seconds' duration. After each original 14s SRC was processed by the relevant HRC, the 14s output was then edited to produce a 10s PVS. For the original SRC, this was achieved by removing the first 2s and final 2s. For a PVS, the 10s edit was achieved by removing the first (2 + N) seconds and final (2 – N) seconds, where N is the temporal registration shift needed to meet the temporal registration limits. Only the middle 10s sequence was stored for use in subjective testing. Objective models were given the 10s PVS, and their choice of either the 10s or 14s SRC.
Sample frames for most of the SRC used are shown below. All SRC associated with Tests 1-5 can be redistributed royalty free for research and development purposes. Test 6 must be held private, and some of Test 6's SRC cannot be displayed in this report. SRC marked "Known SRC" were publically available prior to the HDTV Test (e.g., redistributed by VQEG to proponents, freely available for download on the internet, used in the VQEG Multimedia experiment at a different resolution). SRC marked as "Secret SRC" were not available to any proponent in any form prior to model submission (e.g., filmed for the HDTV test). SRC marked "(edited)" were edited differently than the raw footage (e.g., to add visual interest).

VQEGHD 1 – 1080p 29.97 fps


VQEGHD 1, SRC 1 – Known SRC

VQEGHD 1, SRC 2 – Known SRC


VQEGHD 1, SRC 3 – Known SRC

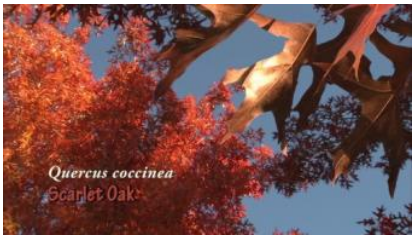

VQEGHD 1, SRC 4 – Known SRC


VQEGHD 1, SRC 5 – Secret SRC


VQEGHD 1, SRC 6 – Known SRC


VQEGHD 1, SRC 7 – Known SRC


VQEGHD 1, SRC 8 – Secret SRC



VQEGHD 1, SRC 9 – Known SRC

VQEGHD 2 – 1080i 59.94fps


VQEGHD 2, SRC 1 – Secret SRC


VQEGHD 2, SRC 2 – Known SRC


VQEGHD 2, SRC 3 – Known SRC

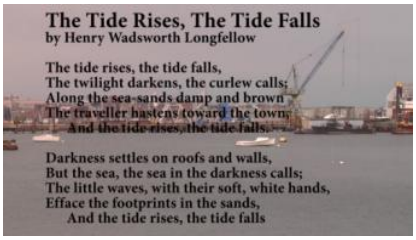

VQEGHD 2, SRC 4 – Secret SRC


VQEGHD 2, SRC 5 – Known SRC


VQEGHD 2, SRC 6 – Secret SRC


VQEGHD 2, SRC 7 – Secret SRC


VQEGHD 2, SRC 8 – Secret SRC



VQEGHD 2, SRC 9 – Secret SRC

VQEGHD 3 – 1080p 29.97fps


VQEGHD 3, SRC 1 – Known SRC


VQEGHD 3, SRC 2 – Secret SRC


VQEGHD 3, SRC 3 – Known SRC


VQEGHD 3, SRC 4 – Secret SRC


VQEGHD 3, SRC 5 – Secret SRC


VQEGHD 3, SRC 6 – Secret SRC


VQEGHD 3, SRC 7 – Known SRC

VQEGHD 3, SRC 8 – Known SRC


VQEGHD 3, SRC 9 – Known SRC

VQEGHD 4 – 1080i 50fps


VQEGHD 4, SRC 1 – Secret SRC


VQEGHD 4, SRC 2 – Secret SRC


VQEGHD 4, SRC 3 – Secret SRC

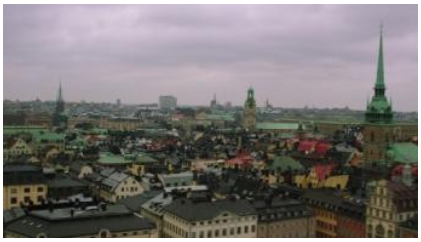

VQEGHD 4, SRC 4 – Secret SRC


VQEGHD 4, SRC 5 – Secret SRC


VQEGHD 4, SRC 6 – Known SRC (edited)


VQEGHD 4, SRC 7 – Secret SRC


VQEGHD 4, SRC 8 – Known SRC


VQEGHD 4, SRC 9 – Secret SRC

VQEGHD 5 – 1080p 25fps


VQEGHD 5, SRC 1 – Known SRC


VQEGHD 5, SRC 2 – Secret SRC


VQEGHD 5, SRC 3 – Known SRC (edited)


VQEGHD 5, SRC 4 – Secret SRC


VQEGHD 5, SRC 5 – Secret SRC

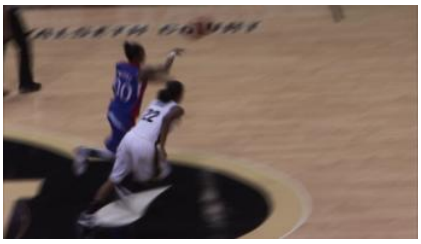

VQEGHD 5, SRC 6 – Secret SRC


VQEGHD 5, SRC 7 – Known SRC

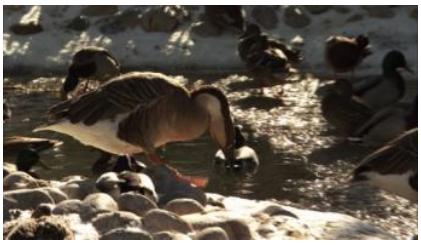

VQEGHD 5, SRC 8 – Known SRC


VQEGHD 5, SRC 9 – Secret SRC

VQEGHD 6 – 1080i 50fps


VQEGHD 6, SRC 1 – Secret SRC


VQEGHD 6, SRC 2 – Secret SRC


VQEGHD 6, SRC 3 – Secret SRC


VQEGHD 6, SRC 4 – Secret SRC


VQEGHD 6, SRC 5 – Known SRC (edited)
VQEGHD 6, SRC 6 – Secret SRC
VQEGHD 6, SRC 7 – Secret SRC
VQEGHD 6, SRC 8 – Secret SRC
VQEGHD 6, SRC 9 – Secret SRC

Note: sample frames from VQEG HD 6, SRC 6, 7, 8, and 9 cannot be included in this report due to usage restrictions on the footage

Common Set SRC


Common Set SRC 11 – Secret SRC


Common Set SRC 12 – Secret SRC


Common Set SRC 13 – Known SRC (edited)


Common Set SRC 14 – Secret SRC

The Common Video Sequences were all filmed in 1080p 25fps.

5.2. HRC Constraints and Sequence Processing

5.2.1. Sequence Processing Overview

The HRCs were selected separately by the ILG. In some cases where IP was involved in the HRC, the transport streams were saved and Ethereal dumps should be captured (e.g., vqeghd1).

5.2.2. Format Conversions

A PVS is the same scale, resolution, and format as the original. An HRC can include transformations such as 1080i to 720p 1080i as long as one pixel of video is displayed as one pixel native display. No up-sampling or down-sampling of the video image is allowed in the final PVS.
Thus, it is not allowable to show 720p footage that is "windowed" in a 1280 x 720 region of a 1080 video.

5.2.3. PVS Duration

All SRCs and PVSs used in testing were 10 seconds long.

5.2.4. Evaluation of 720p

Note that 720p was part of this test plan as included as HRCs.

5.2.5. Constraints on Hypothetical Reference Circuits (HRCs)

The subjective tests were performed to investigate a range of HRC error conditions including both mild and severe errors. These error conditions are limited to the following:

- Compression artifacts (such as those introduced by varying bit-rate, codec type, frame rate and so on)
- Pre- and post-processing effects were allowed but were seldom used in the actual tests other than de-interlacing.
- Transmission errors

HRCs in one experiment may be the same or different from HRCs in other experiments.

The overall selection of the HRCs was done such that most, but not necessarily all, of the codecs, bit rates, encoding modes and impairments set out in the following sections are represented.

5.2.6. Coding Schemes

Only the following coding schemes are allowed:

- MPEG-2
- H.264 (AVC high profile and main profile).

5.2.7. Video Bit-Rates:

Bit rates were chosen to accommodate the coding schemes above and to span a wide range of video quality:

- 1080p SRC: 1–30 Mbps
- 1080i SRC: 1–30 Mbps

5.2.8. Video Encoding Modes

The encoding modes may include, but are not limited to:

- Constant-bit-rate encoding (CBR)
- Variable-bit-rate encoding (VBR)

5.2.9. Frame rates

For those codecs that only offer automatically-set frame rate, this rate was decided by the codec. Some codecs have options to set the frame rate either automatically or manually. For those codecs that have options for manually setting the frame rate, and should an HRC require a manually set frame rate, the minimum frame rate used was 25 fps.
Manually set frame rates (new-frame refresh rate) may include:

- 1080p SRC: 25, 29.97, 30 fps
- 1080i SRC: 25, 29.97, 30 fps

5.2.10. Transmission Errors

Transmission error conditions were allowed. The types of errors include packet errors (both IP and Transport Stream) such as packet loss, packet delay variation, jitter, overflow and underflow, bit errors, and over the air transmission errors. Error concealment was included in at least some of the HRCs. Transmission errors were produced by random packet loss, bursty packet loss, and line conditions specified in G.1050 (e.g., 131 through 133, and A to H).

5.3. Processing and Editing of Sequences

5.3.1. Pre-Processing

The HRC processing may include, typically prior to the encoding, one or more of the following:

- Filtering
- De-interlacing
- Color space conversion (e.g. from 4:2:2 to 4:2:0)
- Down and up sampling is allowed.
- Downscaling to 720p (i.e., paired with post-processing that up-scales back to 1080) is of particular interest.

This processing was considered part of the HRC. Pre-processing was supposed to be realistic and not artificial.

5.3.2. Post-Processing

Post-processing effects may be included in the preparation of test material, such as:

- Down and up sampling is allowed

- De-blocking
- Up-scaling from 720p to 1080i or 1080p (i.e., paired with pre-processing that down-scales to 720p).

Pre-processing was supposed to be realistic and not artificial.

5.3.3. Chain of Coder/Decoder

An HRC can consist of a chain of coder/decoder steps. For example, MPEG-2 encoder followed by MPEG-2 decoder, then H.264 encoder, followed by the H.264 decoder. These HRCs should represent realistic conditions.

These chains may include transmission errors in any transmission. If transmission errors are present in the first leg, then the bandwidth of the first leg should be sufficiently high (e.g., as used in real world scenarios).

5.4. Calibration

5.4.1. Artificial Changes to PVSs

No artificial changes were allowed to the PVSs.

The following impairments were allowed:

- Any impairments produced by agreed codecs.
- Any impairments produced by transmission errors. Transmission errors can be simulated by valid network simulators.
- Manual introduction of freeze frames and manual dropping frames are allowed only to correct temporal alignment violations. If manual introduction of freeze frames and manual dropping frames are made, the ILG should report the correction with detailed explanations.
- Manual shift of the entire video sequence to bring horizontal and vertical shift to be within +/- 1 pixels.
- Manual re-scaling of the entire video sequence to eliminate spatial scaling, if and only if this allows the use of a transmission error HRC that would otherwise be eliminated. Any remaining spatial scaling (if any) must be less than one pixel horizontally and less than one line vertically, such that it is difficult or impossible to tell that any scaling problem previously existed.

The disallowed impairments include, but are not limited to:

- Any changes of pixel values of PVSs.
- Any changes of pixel positions of PVSs.

5.4.2. Recommended HRC Calibration Constraints

All of the calibration constraints were recommended levels. There were no compulsory calibration limits.

The choice of HRCs and Processing by the ILG should remain within the following calibration limits (i.e., when comparing Original Source and Processed sequences).

- maximum allowable deviation in *luminance gain* is +/- 10%
- maximum allowable deviation in *luminance offset* is +/- 20
- maximum allowable *Horizontal Shift* is +/- 1 pixel
- maximum allowable *Vertical Shift* is +/- 1 line
- maximum allowable *Horizontal Cropping* is 30 pixels
- maximum allowable *Vertical Cropping* is 20 lines
- no *Vertical or Horizontal Re-scaling* is allowed
- *Temporal Alignment* The first and the last 1 second may only have +/- quarter second temporal shift and will not contain anomalous freeze frames longer than 0.1 second. The maximum of the total freeze is 25% of the total length of the sequence.

- No portion of the PVS can be included that do not have an associated portion in the SRC.
- In addition, the entire PVS should be contained in the associated 10-second SRC
- A maximum of 2 seconds might be cut off from the PVS.
- *Dropped or Repeated Frames* are excluded from above temporal alignment limit
- no visible *Chroma Differential Timing* is allowed
- no visible *Picture Jitter* is allowed
- A *frame freeze* is defined as any event where the video pauses for some period of time then restarts. Frame freezes are allowed in the current testing. *Frame freezing* or pure black frames (e.g., from over-the-air broadcast lack of delivery) should not be longer than 2 seconds duration.
- *Frame skipping* is defined as events where some loss of video frames occurs. Frame skipping is allowed in the current testing.
- Note that where frame freezing or frame skipping is included in a test then source material containing still / nearly still sections are recommended to form part of the testing.
- *Rewinding* is not allowed. Where it is difficult or impossible by a visual inspection to tell if a PVS has rewinding the PVS will be allowed in the test.

For HRCs that include simulated transmission errors, the freeze-frame restriction and the temporal alignment restrictions are to be relaxed because they are difficult to enforce. However, ILG reserved the right to reject PVSs that seem to violate freeze-frame and temporal alignment restrictions in an extreme or artificial way that should not be encountered in real delivery of HD. The intent of this rule is to allow PVSs created by transmission error HRCs operating in a "reasonable" mode, while excluding (a) PVSs that may have been artificially constructed to disadvantage other models and (b) PVSs created by "excessive" transmission errors. ILG judgments of "reasonable," "extreme," "artificial," and "excessive" are to be treated in the same spirit as the calls of football/soccer referees.

Laboratories should verify adherence of HRCs to these limits by using software packages (NTIA software suggested) in addition to human checking.

5.4.3. Required HRC Calibration Constraints

The following constraints must be met by every PVS. These constraints were chosen to be easily checked by the ILG, and to provide proponents with feedback on their model's calibration intended search range.

- maximum allowable deviation in *luminance gain* is +/- 20% (Recommended is +/- 10%)
- maximum allowable deviation in *luminance offset* is +/- 50 (Recommended is +/- 20)
- maximum allowable *Horizontal Shift* is +/- 5 pixels (Recommended is +/- 1)
- maximum allowable *Vertical Shift* is +/- 5 lines (Recommended is +/- 1)
- No PVS may have visibly obvious scaling.
- The color space must appear to be correct (e.g., a red apple should not mistakenly rendered be rendered "blue" due to a swap of the Cb and Cr color planes).
- No more than 1/2 of a PVS may consist of frozen frames or pure black frames (e.g., from over-the-air broadcast lack of delivery).
- Pure black frames (e.g., from over-the-air broadcast lack of delivery) must not occur in the first 2-seconds or the last 4-seconds of any PVS. The reason for this constraint, is that the viewers may be confused and mistake the black for the end of sequence.
- When creating PVSs, a 14-second SRC should be used, with +2 second of extra content before and after. All of the content visible in the PVS should correspond to SRC content from either the edited 10-second SRC or the longer 14-second SRC.
- The first frame of each 10-second PVS should closely match the first frame of the 10-second SRC (unless the video sequence begins with a freeze-frame). Note that in section 7.2 it is recommended that the first half second and the last half second might not contain any noticeable freezing so that the evaluators might not be confused whether the freezing comes from impairments or the player.
- The field order must not be swapped (e.g., field one moved forward in time into field two, field two moved back in time into field one).

The intent was that all PVSs contain realistic impairments that could be encountered in real delivery of HDTV (e.g., over-the-air broadcast, satellite, cable, IPTV).

6. Model Evaluation Criteria:

This chapter describes the evaluation metrics and procedure used to assess the performance of an objective video quality model as an estimator of video picture quality in a variety of applications.

6.1. Evaluation Procedure

The performance of each objective quality model was characterized by three prediction attributes: accuracy, monotonicity and consistency.

The statistical metrics root mean square error (RMSE), Pearson correlation, and outlier ratio together characterize the accuracy, monotonicity and consistency of a model's performance. These statistical metrics are named evaluation metrics in the following. The calculation of each evaluation metric is performed along with its 95% confidence intervals. To test for statistically significant differences among the performance of various models, a test based on the F-test used the RMSE; tests based on approximations to the Gaussian distribution were constructed for the Pearson correlation coefficient and the Outlier Ratio.

The evaluation metrics were calculated using the objective model outputs and the results from viewer subjective rating of the test video clips. The objective model provides a single number (figure of merit) for every tested video clip. The same tested video clips get also a single subjective figure of merit. The subjective figure of merit for a video clip represents the average value of the scores provided by all subjects viewing the video clip.

The evaluation analysis is based on DMOS scores for the FR and RR models, and on MOS scores for the NR model. Discussion below regarding the DMOS scores was applied identically to MOS scores. For simplicity, only DMOS scores are mentioned for the rest of the chapter.

The objective quality model evaluation was performed in three steps. The first step is a mapping of the objective data to the subjective scale. The second calculates the evaluation metrics for the models and their confidence intervals. The third tests for statistical differences between the evaluation metrics value of different models..

6.2. Data Processing

6.2.1. PVS Discarded Prior to Model Analysis

The following video sequences were not used during model analysis. These sequences were discarded by the ILG prior after model submission, when the following problems were identified.

1. The 1080p 25fps to 1080i 59.94fps conversion worked quite well perceptually, but had the unforeseen consequence. The 3-2 pull down pattern was different for the original and processed video sequences. This created an artificial situation that would never be encountered in actual practice. Thus, the common set for the 1080i 59.94fps test (vqeghd2) was used to map all tests to one scale (i.e., using the MOS). However, the objective models were not compared to the vqeghd2 common sequences.
2. Experiment vqeghd1, SRC 2, HRC 15.
3. Experiment vqeghd3, SRC 6, HRC 7.

6.2.2. Calculating DMOS Values

The data analysis was performed using the difference mean opinion score (DMOS) for FR and RR methods and using the MOS for NR models. DMOS values were calculated on a per subject per PVS basis. The appropriate hidden reference (SRC) was used to calculate the DMOS value for each PVS. DMOS values were calculated using the following formula:

DMOS = MOS (PVS) – MOS (SRC) + 5

In using this formula, higher DMOS values indicate better quality. Lower bound is 1 as MOS value but higher bound could be more than 5. Any DMOS values greater than 5 (i.e. where the processed sequence is rated better quality than its associated hidden reference sequence) was considered valid and included in the data analysis.

6.2.3. Mapping to the Subjective Scale

Subjective rating data often are compressed at the ends of the rating scales. It is not reasonable for objective models of video quality to mimic this weakness of subjective data. Therefore, a non-linear mapping step was applied before computing any of the performance metrics. A non-linear mapping function that has been found to perform well empirically is the cubic polynomial:

$$DMOSp = ax^3 + bx^2 + cx + d \quad (1)$$

where DMOSp is the predicted DMOS, and the VQR is the model's computed value for a clip-HRC combination. The weightings *a*, *b* and *c* and the constant *d* are obtained by fitting the function to the data [DMOS, VQR].

The mapping function maximizes the correlation between DMOSp and DMOS :

$$DMOSp = k(a'x^3 + b'x^2 + c'x) + d$$

with constant k = 1, d = 0

This function must be constrained to be monotonic within the range of possible values for our purposes. Then the root mean squared error is minimized over k and d.

a = k*a'

b = k*b'

c = k*c'

This non-linear mapping procedure has been applied to each model's outputs before the evaluation metrics are computed.

ILG used the coefficients of the fitting function that produce the best correlation coefficient provided that it is a monotonic fit.

6.2.4. Analysis, Averaging Process and Aggregation Procedure

Primary analysis of model performance was calculated both per processed video sequence per experiment, and per processed video sequence on a superset of all six experiments aggregated into one data set. The average correlation and RMSE over all six experiments is also reported.

6.3. Evaluation Metrics

Once the mapping was applied to objective data, two evaluation metrics: root mean square error and Pearson correlation coefficient. The calculation of each evaluation metric was performed along with its 95% confidence interval. RMSE is considered the primary metric.

6.3.1. Pearson Correlation Coefficient

The Pearson correlation coefficient R (see equation 2) measures the linear relationship between a model's performance and the subjective data. Its great virtue is that it is on a standard, comprehensible scale of -1 to 1 and it has been used frequently in similar testing.

$$R = \frac{\sum_{i=1}^{N} (Xi - \overline{X}) * (Yi - \overline{Y})}{\sqrt{\sum (Xi - \overline{X})^2} * \sqrt{\sum (Yi - \overline{Y})^2}} \quad (2)$$

Xi denotes the subjective score (DMOS(i) for FR/RR models and MOS(i) for NR models) and Yi the objective score (DMOSp(i) for FR/RR models and MOSp(i) for NR models).. N in equation (2) represents the total number of video clips considered in the analysis.

Therefore, in the context of this test, the value of N in equation (2) is:

- N=152 for FR/RR models (=166-14 since the evaluation for FR/RR discards the reference videos and there are 14 reference videos in each experiment).

- N=166 for NR models.
- Note, if any PVS in the experiment is discarded for data analysis, then the value of N changes accordingly.

The sampling distribution of Pearson's R is not normally distributed. "Fisher's z transformation" converts Pearson's R to the normally distributed variable z. This transformation is given by the following equation :

$$z = 0.5 \cdot \ln\left(\frac{1+R}{1-R}\right) \quad (3)$$

The statistic of z is approximately normally distributed and its standard deviation is defined by:

$$\sigma_z = \sqrt{\frac{1}{N-3}} \quad (4)$$

The 95% confidence interval (CI) for the correlation coefficient is determined using the Gaussian distribution, which characterizes the variable z and it is given by (5)

$$CI = \pm K1 * \sigma_z \quad (5)$$

NOTE1: For a Gaussian distribution, K1 = 1.96 for the 95% confidence interval. If N<30 samples are used then the Gaussian distribution must be replaced by the appropriate Student's t distribution, depending on the specific number of samples used.

Therefore, in the context of this test, K1 = 1.96.

The lower and upper bound associated to the 95% confidence interval (CI) for the correlation coefficient is computed for the Fisher's z value:

$$LowerBound = z - K1 * \sigma_z$$

$$UpperBound = z + K1 * \sigma_z$$

NOTE2**:** The values of Fisher's z of lower and upper bounds are then converted back to Pearson's R to get the CI of correlation R.

6.3.2. Root Mean Square Error

The accuracy of the objective metric is evaluated using the root mean square error (RMSE) evaluation metric.

The difference between measured and predicted DMOS is defined as the absolute prediction error Perror:

$$Perror(i) = DMOS(i) - DMOS_p(i) \quad (6)$$

where the index i denotes the video sample.

NOTE: DMOS(i) and DMOSp(i) are used for FR/RR models. MOS(i) and MOSp(i) are used for NR models.

The root-mean-square error of the absolute prediction error Perror is calculated with the formula:

$$rmse = \sqrt{\left(\frac{1}{N-d}\sum_N Perror[i]^2\right)} \quad (7)$$

where N denotes the total number of video clips considered in the analysis, and d is the number of degrees of freedom of the mapping function (1).

In the case of a mapping using a 3rd-order monotonic polynomial function, d=4 (since there are 4 coefficients in the fitting function).

In the context of this test plan, the value of N in equation (7) is:

- N=152 for FR/RR models (since the evaluation discards the reference videos and there are 14 reference videos in each experiment)
- N=166 for NR models
- NOTE: if any PVS in the experiment is discarded for data analysis, then the value of N changes accordingly.

The root mean square error is approximately characterized by a χ^2 (n) [2], where n represents the degrees of freedom and it is defined by (8):

$$n = N - d \tag{8}$$

where N represents the total number of samples.

Using the χ^2 (n) distribution, the 95% confidence interval for the rmse is given by (9) [2]:

$$\frac{rmse * \sqrt{N-d}}{\sqrt{\chi^2_{0.025}(N-d)}} < rmse < \frac{rmse * \sqrt{N-d}}{\sqrt{\chi^2_{0.975}(N-d)}} \tag{9}$$

6.4. Statistical Significance of the Results

6.4.1. Significance of the Difference between the Correlation Coefficients

The test is based on the assumption that the normal distribution is a good fit for the video quality scores' populations. The statistical significance test for the difference between the correlation coefficients uses the H_0 hypothesis that assumes that there is no significant difference between correlation coefficients. The H_1 hypothesis considers that the difference is significant, although not specifying better or worse.

The test uses the Fisher-z transformation (3) [2]. The normally distributed statistic Z_N (10) is determined for each comparison and evaluated against the 95% t-Student value for the two–tail test, which is the tabulated value t(0.05) =1.96.

$$Z_N = \frac{z1 - z2 - \mu_{(z1-z2)}}{\sigma_{(z1-z2)}} \tag{10}$$

$$\text{where } \mu_{(z1-z2)} = 0 \tag{11}$$

and

$$\sigma_{(z1-z2)} = \sqrt{\sigma_{z1}^2 + \sigma_{z2}^2} \tag{12}$$

σ_{z1} and σ_{z2} represent the standard deviation of the Fisher-z statistic for each of the compared correlation coefficients. The mean (11) is set to zero due to the H_0 hypothesis and the standard deviation of the difference metric z1-z2 is defined by (12).

The standard deviation of the Fisher-z statistic is given by (13):

$$\sigma_z = \sqrt{1/(N-3)} \tag{13}$$

where N represents the total number of samples used for the calculation of each of the two correlation coefficients.

Using (12) and (13), the standard deviation of the difference metric z1-z2 therefore becomes:

$$\sigma_{z1-z2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$$

where N1=N2=N

6.4.2. Significance of the Difference between the Root Mean Square Errors

Considering the same assumption that the two populations are normally distributed, the comparison procedure is similar to the one used for the correlation coefficients. The H_0 hypothesis considers that there is no difference between rmse values. The alternative H_1 hypothesis is assuming that the lower prediction error value is statistically significantly lower. The statistic defined by (14) has a F-distribution with n1 and n2 degrees of freedom [2].

$$\zeta = \frac{(rmse_{\max})^2}{(rmse_{\min})^2} \tag{14}$$

$rmse_{max}$is the highest rmse and $rmse_{min}$is the lowest rmse involved in the comparison. The ζ statistic is evaluated against the tabulated value F(0.05, n1, n2) that ensures 95% significance level. The n1 and n2 degrees of freedom are given by N1-d, respectively and N2-d, with N1 and N2 representing the total number of samples for the compared average rmse (prediction errors) and d being the number of parameters in the fitting equation (1).

If ζ is higher than the tabulated value F(0.05, n1, n2) then there is a significant difference between the values of RMSE.

7. Common Video Clip Analysis and Interpretation:

According to the test plan common set was used in order to create superset

"*Second, if the data appears consistent from lab to lab, then the common set of video sequences will be used to map all video sequences onto a single scale, forming a "superset".*"

In this section we present consistency analysis. In the test plan consistent was defined as

„*The criteria used will be established during audio calls, before model submission (e.g., proposals include (1) average lab-to-lab correlation for all experiments must be at least 0.94, and also for every individual experiment, the average lab-to-lab correlation to all other experiments must be at least 0.91; and (2) a Chi-Squared Pearson Test or F-Test).*"

Since the details of Chi-Squared Pearson Test and F-Test were not described in details we based our analysis on the lab-to-lab correlation. Chi-Squared Pearson Test analysis is shown in Appendix IV.

Tables 4 and 5 show correlations between different ILGs and between an ILG and all other ILGs.

Table 4. Lab-to-Lab correlation for MOSes

	ILG2	**ILG3**	**ILG4**	**ILG5**	**ILG6**	**All other ILGs**
ILG1	0.965	0.950	0.983	0.972	0.950	0.975
ILG2		0.980	0.969	0.981	0.982	0.989
ILG3			0.943	0.963	0.990	0.978
ILG4				0.986	0.952	0.978
ILG5					0.968	0.987
ILG6						0.980

Table 5. Lab-to-Lab correlation for DMOSes

	ILG2	**ILG3**	**ILG4**	**ILG5**	**ILG6**	**All other ILGs**
ILG1	0.967	0.924	0.981	0.962	0.950	0.969
ILG2		0.974	0.971	0.978	0.981	0.990
ILG3			0.948	0.967	0.985	0.973
ILG4				0.986	0.951	0.980
ILG5					0.962	0.985
ILG6						0.979

For both MOS and DMOS we obtained overall mean correlation 0.97, **which is higher** than accepted in the test plan 0.94. The same **all lab-to-lab correlations are higher** than accepted in the test plan 0.91. The most different are ILG1 and ILG3 (DMOS correlation 0.924). Nevertheless, it does not indicate that one of them should be excluded since for both of them we can find other similar ILGs (ILG4 and ILG6 for ILG1 and ILG3 respectively). Therefore, we can be almost sure that the observed differences are just caused by rand differences between different subjects' sets.

Based on the lab-to-other labs analysis (last column in Tables 1 and 2) we should use ILG2 as a reference common set representation in the superset. ILG2 correlation with other ILGs is the highest for both MOS and DMOS values.

8. Official ILG Data Analysis:

The official ILG data analysis presented in this section was computed using MATLAB code. RMSE is the primary metric for analysis.

Superset analysis was performed by joining all six individual experiments into one experiment using the subjective data associated with the common set, using the techniques described in NTIA's Technical Report on the VQEG MultiMedia Phase I data (NTIA Technical Report TR-09-457, "Techniques for Evaluating Overlapping Video Quality Models Using Overlapping Subjective Data Sets).

The DMOS aggregated superset weights for each experiment used a first order linear predictor ($y = \boldsymbol{A} * x + \boldsymbol{B}$), where $\boldsymbol{x}$ was DMOS and constants $\boldsymbol{A}$ and $\boldsymbol{B}$ are given in the table below:

Experiment	*A*	*B*
vqeghd1	0.937207235171285350	-0.001371722758369663
vqeghd2	0.893175129456469150	0.526312467144340550
vqeghd3	1.143655394947402900	-0.459729436472336660
vqeghd4	0.984623393773172760	-0.222091332723857500
vqeghd5	0.980826932875394460	-0.034482873401984152
vqeghd6	0.900285576261259160	0.698591285363933000

The common set for vqeghd5 was retained for DMOS. The common set sequences from the other datasets were discarded, to avoid overly emphasizing the common set clips within the superset analyses.

The MOS aggregated superset was also calculated. The MOS superset is not used in the VQEG ILG Official Data Analysis, because all NR models were withdrawn. The MOS aggregated superset weights for each experiment used a first order linear predictor ($y = \boldsymbol{A} * x + \boldsymbol{B}$), where $\boldsymbol{x}$ was MOS and constants $\boldsymbol{A}$ and $\boldsymbol{B}$ are given in the table below:

Experiment	*A*	*B*
vqeghd1	0.95418609952275657000	-0.13564600904313739000
vqeghd2	0.89605340995520955000	0.44532944589988510000
vqeghd3	1.15361055689733670000	-0.07324503832825869600
vqeghd4	0.98166603271205799000	-0.14287989729935313000
vqeghd5	0.98358230565365468000	-0.17980480905826920000
vqeghd6	0.90857597526894707000	0.49253493565600009000

The common set for vqeghd5 was retained for MOS. The common set sequences from the other datasets were discarded, to avoid overly emphasizing the common set clips within the superset analyses.

8.1. PSNR

PSNR was calculated according to ITU-T Draft Rec. J.340, which included temporal and spatial calibration. However, to save computation time, the luminance gain & offset calculation for PSNR were calculated separately and input to the PSNR algorithm as constants, and an appropriate search range was chosen for each dataset. The spatial region of interest (SROI) was in all cases set to (top=11, left=15, bottom=1076, right=1902) where the top-left corner of the image is coordinate (1,1) and the SROI coordinates are inclusive. The luminance gain & offset values and search ranges are as follows:

Dataset	Spatial Uncertainty	Temporal Uncertainty	Luminance Gain & Offset
vqeghd1	Horizontal +/- 5	+/- 10 frames	Gain 1.0

	Vertical +/- 0	except for HRC 15, which used +/- 65 frames.[2]	Offset 0.0
vqeghd2	Horizontal +/- 0 Vertical +/- 0	+/- 15 frames	Gain 1.0 Offset 0.0
vqeghd3	Horizontal +/- 0 Vertical +/- 0	+/- 10 frames	Gain 1.0 Offset 0.0
vqeghd4	Horizontal +/- 0 Vertical +/- 0	+/- 10 frames	HRC 1-8 gain 1.159873 and offset -18.91867 HRC 9-10 gain 1.0 and offset 0.65545 HRC 11 gain 1.0 and offset 1.8997 HRC 12-15 gain 1.0 and offset 1.53968
vqeghd5	Horizontal +/- 0 Vertical +/- 0	+/- 10 frames	Gain 1.0 Offset 0.0
vqeghd6	Horizontal +/- 0 Vertical +/- 0	+/- 10 frames	Gain 1.0 Offset 0.0
common set	Horizontal +/- 0 Vertical +/- 0	+/- 10 frames	Gain 1.0 Offset 0.0

These values were calculated using the calibration routines associated with the NTIA General Model, calculated according to ITU-T J.244, and confirmed by Swissqual using a different technique.

8.2. FR Models

Eight FR models were submitted to VQEG. Of these, three were withdrawn and the remaining five models's official data analysis are presented in this section.

The official ILG data analysis of the FR models is presented in the following tables. Table 6 lists the Pearson Correlation for each FR model, for the six datasets (1 through 6), as well as the superset. The final line of this table indicates the average correlation value for that model over all six experiments (i.e., excluding the superset). Table 7 lists the RMSE for each FR model, for the six datasets (1 through 6), as well as the superset. The final line of this table likewise indicates the average RMSE value for that model over all six experiments (i.e., excluding the superset). Table 8 identifies (for each dataset and the superset) all models that were statistically equivalent to the top performing model. This calculation is performed for each dataset and the superset separately, and such top performing models are indentified with a "1". The final line indicates the total number of datasets where that model was in the top performing group. Table 9 identifies (for each dataset and the superset) all models that were statistically better than PSNR. This calculation is performed for each dataset and the superset separately, and such models are indentified with a "1". The final line indicates the total number of datasets where that model performed significantly better than PSNR.

Table 6. Pearson Correlation of HDTV FR Models for Individual Datasets and Aggregated Superset

Dataset	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
vqeghd1	0.85	0.82	0.48	0.88	0.77	0.75
vqeghd2	0.58	0.62	0.59	0.83	0.73	0.66
vqeghd3	0.85	0.84	0.73	0.92	0.82	0.88
vqeghd4	0.81	0.79	0.76	0.82	0.88	0.74
vqeghd5	0.72	0.64	0.53	0.85	0.83	0.62
vqeghd6	0.79	0.72	0.81	0.90	0.87	0.90
Superset	0.78	0.76	0.63	0.87	0.82	0.76
Average [1..6]	0.77	0.74	0.65	0.86	0.82	0.76

[2] The +/- 65 frame temporal uncertainty for vqeghd1 HRC 15 was to accommodate a video sequence that was later discarded as being noncompliant with the HDTV Test Plan calibration constraints.

Table 7. RMSE of HDTV FR Models for Individual Datasets and Aggregated Superset

Dataset	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
vqeghd1	0.65	0.71	1.09	0.59	0.79	0.82
vqeghd2	0.84	0.81	0.83	0.58	0.70	0.78
vqeghd3	0.59	0.61	0.78	0.45	0.65	0.53
vqeghd4	0.66	0.68	0.73	0.65	0.53	0.75
vqeghd5	0.77	0.85	0.93	0.59	0.61	0.87
vqeghd6	0.65	0.74	0.63	0.46	0.53	0.47
superset	0.71	0.74	0.88	0.56	0.65	0.74
Average [1..6]	0.69	0.73	0.83	0.55	0.64	0.70

Table 8. HDTV FR Models in the Top Performing Group (1) versus Models with Less Accurate Performance (0); Computed from RMSE

Dataset	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
vqeghd1	1	0	0	1	0	0
vqeghd2	0	0	0	1	0	0
vqeghd3	0	0	0	1	0	0
vqeghd4	0	0	0	0	1	0
vqeghd5	0	0	0	1	1	0
vqeghd6	0	0	0	1	1	1
superset	0	0	0	1	0	0
Total [1..6]	1	0	0	5	3	1

Table 9. HDTV FR Models that Performed Significantly Better than PSNR (1) versus Models with Performance Statistically Equivalent to or Worse than PSNR (0); Computed from RMSE

	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
vqeghd1	0	0	0	0	0	0
vqeghd2	0	0	0	1	1	0
vqeghd3	0	0	0	1	0	0
vqeghd4	0	0	0	0	1	0
vqeghd5	0	0	0	1	1	0
vqeghd6	0	0	0	1	1	1
Superset	0	0	0	1	0	0
Total [1..6]	0	0	0	4	4	1

Table 10 lists the RMSE for each FR model, for subdivisions of the superset. These subdivisions divide the data by coding type (H.264 or MPEG-2) as well as by the presence of transmission errors (Errors) or whether the HRC contained coding artifacts only (Coding). Because the experiments were not designed to have these variables evenly span the full range of quality, only RMSE are presented for these subdivisions. Objective models were not re-fitted prior to this analysis (i.e., the polynomial fit for the entire superset was used for each of these RMSE calculations). Table 11 identifies (for each type of HRC) all models that were statistically equivalent to the top performing model. Table 12 identifies (for each type of HRC) all models that were statistically better than PSNR.

Table 10. RMSE of HDTV FR Models for Aggregated Superset, Divided by HRC Type

Dataset	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
H.264 Coding	0.75	0.72	0.80	0.55	0.61	0.63
H.264 Error	0.67	0.75	0.89	0.56	0.65	0.84
mpeg-2 Coding	0.78	0.80	0.84	0.60	0.66	0.80
mpeg-2 Error	0.66	0.73	1.10	0.59	0.74	0.77
Coding	0.75	0.74	0.80	0.56	0.62	0.67
Error	0.67	0.74	0.97	0.57	0.68	0.81

Table 11. HDTV FR Models in the Top Performing Group (1) by HRC Type Versus Models with Less Accurate Performance (0); Computed from RMSE

Dataset	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
H.264 Coding	0	0	0	1	0	0
H.264 Error	0	0	0	1	0	0
mpeg-2 Coding	0	0	0	1	1	0
mpeg-2 Error	1	0	0	1	0	0
Coding	0	0	0	1	0	0
Error	0	0	0	1	0	0

Table 12. HDTV FR Models that Performed Significantly Better than PSNR (1) by HRC Type Versus Models with Performance Statistically Equivalent to or Worse than PSNR (0); Computed from RMSE

Dataset	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
H.264 Coding	0	0	0	1	1	1
H.264 Error	0	0	0	1	0	0
mpeg-2 Coding	0	0	0	1	1	0
mpeg-2 Error	0	0	0	0	0	0
Coding	0	0	0	1	1	1
Error	0	0	0	1	0	0

Table 13 identifies the lower and upper confidence interval for each correlation and RMSE mentioned in Table 6 and 7. In this table, "lower bound" is defined to be less accurate than the indicated measurement, and "upper bound" is defined to be more accurate than the indicated measurement.

Table 13. Confidence Intervals for FR Models' Pearson Correlation and RMSE

PSNR						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.85	0.80	0.89	0.65	0.73	0.59
vqeghd2	0.58	0.45	0.68	0.84	0.96	0.75
vqeghd3	0.85	0.80	0.89	0.59	0.67	0.53
vqeghd4	0.81	0.75	0.86	0.66	0.74	0.59
vqeghd5	0.72	0.63	0.79	0.77	0.87	0.69
vqeghd6	0.79	0.73	0.84	0.65	0.73	0.58
superset	0.78	0.75	0.80	0.71	0.75	0.68
NTT						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.82	0.76	0.87	0.71	0.80	0.63
vqeghd2	0.62	0.50	0.71	0.81	0.93	0.73
vqeghd3	0.84	0.79	0.88	0.61	0.69	0.55
vqeghd4	0.79	0.72	0.84	0.68	0.77	0.61

vqeghd5	0.64	0.53	0.72	0.85	0.96	0.77
vqeghd6	0.72	0.63	0.79	0.74	0.83	0.66
superset	0.76	0.73	0.79	0.74	0.77	0.70
Opticom						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.48	0.35	0.59	1.09	1.23	0.98
vqeghd2	0.59	0.47	0.69	0.83	0.95	0.74
vqeghd3	0.73	0.64	0.79	0.78	0.88	0.70
vqeghd4	0.76	0.68	0.82	0.73	0.82	0.65
vqeghd5	0.53	0.41	0.64	0.93	1.05	0.84
vqeghd6	0.81	0.74	0.86	0.63	0.71	0.57
superset	0.63	0.59	0.67	0.88	0.93	0.84
Swissqual						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.88	0.83	0.91	0.59	0.67	0.53
vqeghd2	0.83	0.77	0.87	0.58	0.66	0.52
vqeghd3	0.92	0.89	0.94	0.45	0.51	0.41
vqeghd4	0.82	0.75	0.86	0.65	0.73	0.58
vqeghd5	0.85	0.80	0.89	0.59	0.66	0.53
vqeghd6	0.90	0.87	0.93	0.46	0.52	0.42
superset	0.87	0.85	0.89	0.56	0.59	0.54
Tektronix						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.77	0.70	0.83	0.79	0.89	0.71
vqeghd2	0.73	0.64	0.80	0.70	0.80	0.63
vqeghd3	0.82	0.76	0.87	0.65	0.73	0.58
vqeghd4	0.88	0.84	0.91	0.53	0.60	0.48
vqeghd5	0.83	0.78	0.88	0.61	0.69	0.55
vqeghd6	0.87	0.82	0.90	0.53	0.59	0.47
superset	0.82	0.80	0.84	0.65	0.68	0.62
Yonsei FR						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.75	0.67	0.81	0.82	0.92	0.73
vqeghd2	0.66	0.55	0.75	0.78	0.88	0.69
vqeghd3	0.88	0.84	0.91	0.53	0.60	0.48
vqeghd4	0.74	0.66	0.80	0.75	0.85	0.67
vqeghd5	0.62	0.51	0.71	0.87	0.98	0.78
vqeghd6	0.90	0.86	0.92	0.47	0.53	0.42
superset	0.76	0.73	0.79	0.74	0.77	0.70

Table 14 identifies the coefficients for the monotonic, polynomial fitting between each model and each dataset. All analysis on the superset for one model uses the same mapping (i.e., superset analysis and the division by HRC type). These coefficients were computed using the following equation:

$$DMOSp = A3 \cdot x^3 + A2 \cdot x^2 + A1 \cdot x + A0$$

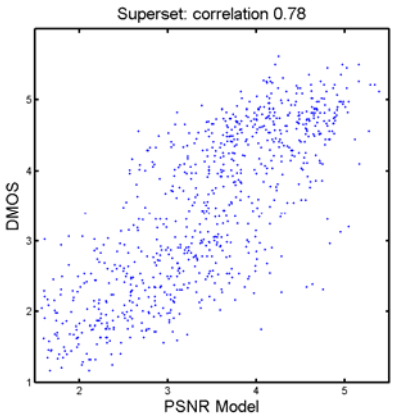
Table 14. Monotonic Polynomial Coefficients for FR Models

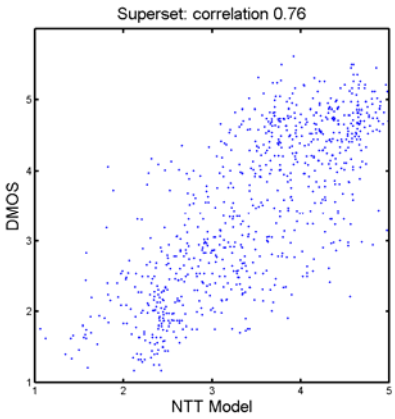
PSNR				
Dataset	**A3**	**A2**	**A1**	**A0**

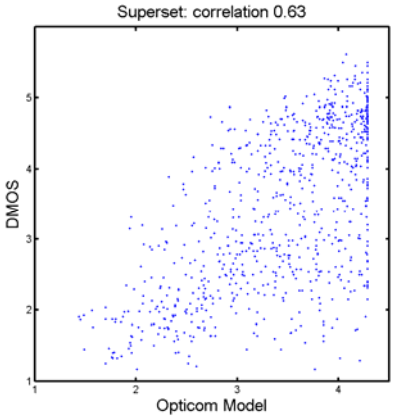
vqeghd1	-0.0000306550279250	-0.0019879070241439	0.3921369753140520	-5.7280907926448500
vqeghd2	-0.0001334181504514	0.0111764931035134	-0.1877690908303150	1.9953778604294600
vqeghd3	-0.0001932109488014	0.0197073219079854	-0.5006502641704290	5.3895615601423200
vqeghd4	-0.0001935191412653	0.0131947404548850	-0.1153986433564110	0.6574532383968700
vqeghd5	-0.0000148995264923	0.0061596181685647	-0.1807405959360720	3.6252029746604300
vqeghd6	-0.0001813435735504	0.0123033171141708	-0.0165560675513631	-3.0727162004210500
superset	-0.0001427599450031	0.0137910530512573	-0.2912028187262140	3.3051795730244100
NTT				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	-0.0991847963872846	0.8110641291428480	-0.7820652077618230	0.9127067754168750
vqeghd2	-0.1181878070429630	1.1448852868251800	-2.6901492618038500	3.8341722796662400
vqeghd3	-0.1014668182639410	0.9614253343962790	-1.7214306356691700	2.5074890045038100
vqeghd4	-0.0695208549063990	0.5392273229835210	-0.2515976773441100	1.4423651722973500
vqeghd5	-0.0005320215143262	-0.1711839802741540	1.9028806475084500	-0.5863634288922570
vqeghd6	-0.0689840989697419	0.5378916958102050	-0.2581679996571180	1.0002632216087900
superset	-0.0473566957759718	0.3596087865076620	0.2761694318804200	0.4705357356597510
Opticom				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	0.0194253864400859	-0.3481537083941570	2.2807565093721200	-1.3279378844098300
vqeghd2	-0.0506872080826940	0.5167756660599310	-0.9654449156833460	1.7397071025822300
vqeghd3	-0.0642972135133286	0.6752295046979770	-1.4270612814454200	2.5892925649976100
vqeghd4	-0.0284814302183273	0.1535563668003300	0.8955697575095650	-0.1668046892408590
vqeghd5	-0.0367502750968249	0.2972269360151380	0.0655911688703438	1.1156512729041800
vqeghd6	-0.0765800660211873	0.8819708335946610	-2.2495516644164200	2.8628671940500600
superset	-0.0467883007483602	0.4625594864549470	-0.6497394120129090	1.6503224064470900
Swissqual				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	-0.0390529532768338	0.3000819489929240	0.2389686214309000	1.1064136671327600
vqeghd2	0.0955240432470277	-0.7053502910429440	2.2785828004402200	-0.3501285074880110
vqeghd3	0.0060017052377394	0.0632292577562197	0.2004616292762410	1.6954879852495900
vqeghd4	-0.0638326373511991	0.4936587120783710	-0.3036608590242600	1.9997878787334900
vqeghd5	-0.0292507606786930	0.2811401948144040	-0.0673892478253106	1.9736766380073900
vqeghd6	-0.0720128488190934	0.6959409672343660	-1.1590469381490800	2.1185604442417500
superset	-0.0069626740478583	0.1136773990155330	0.3547703598000850	1.3056870202460500
Tektronix				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	-0.0000015345209802	0.0000227114200458	-0.0186260577757991	4.9810578730260700
vqeghd2	0.0000004430898298	0.0002001178417874	-0.0533162632509381	5.0637213854754100
vqeghd3	0.0000051413666900	-0.0007733026811158	0.0004195355219074	4.9244640327914700
vqeghd4	0.0000026087551342	-0.0005785073798566	0.0045575394585733	4.8727127283246800
vqeghd5	0.0000054539395279	-0.0008338882650292	0.0031594671685185	4.9485628510620900
vqeghd6	0.0000055462589169	-0.0009883911960620	0.0068822065690831	4.6236914841515000
superset	0.0000049732211974	-0.0007607962321771	0.0004127693339106	4.8178282402146500
Yonsei FR				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	-0.0002595420882522	0.0194552471264414	-0.2477296238903770	0.5057043071630600

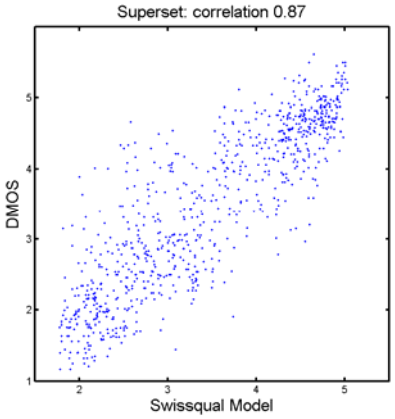
vqeghd2	-0.0001888304682161	0.0171908255886766	-0.3627966350289180	3.0477256606126300
vqeghd3	-0.0001662901649060	0.0159756121531555	-0.3503849785206330	3.4931808523544200
vqeghd4	0.0000709285567507	-0.0146036589546189	0.8771301011775830	-11.8076355926550000
vqeghd5	0.0000856842509876	-0.0148754898352676	0.8225419919304200	-10.3928080379510000
vqeghd6	-0.0002724313571128	0.0283462868127231	-0.7913935029260610	7.5303046941061700
superset	-0.0000228650420972	-0.0004104270038837	0.2450061426536330	-3.4134081646700500

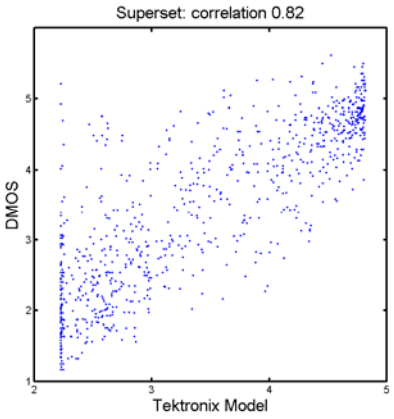
Following are plots for each FR model, showing the model's fitted value on the X-axis and the superset DMOS on the Y-axis. PSNR is plotted for comparison purposes. The superset correlation is shown above the plot. Please note that due to the transformation that was applied to the individual datasets while forming the aggregated superset, some DMOS values are greater than 5.0.

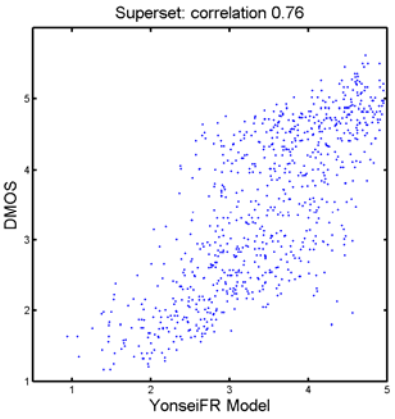












8.3. RR Models

Three RR models were submitted to VQEG. None of the models were withdrawn and so all three models' official data analysis are presented in this section.

The official ILG data analysis of the RR models is presented in the following tables. Table 15 lists the Pearson Correlation for each RR model, for the six datasets (1 through 6), as well as the superset. The final line of this table indicates the average correlation value for that model over all six experiments (i.e., excluding the superset). Table 16 lists the RMSE for each RR model, for the six datasets (1 through 6), as well as the superset. The final line of this table likewise indicates the average RMSE value for that model over all six experiments (i.e., excluding the superset). Table 17 identifies (for each dataset and the superset) all models that were statistically equivalent to the top performing model. This calculation is performed for each dataset and the superset separately, and such top performing models are indentified with a "1". The final line indicates the total number of datasets where that model was in the top performing group. Table 18 identifies (for each dataset and the superset) all models that were statistically better than or equivalent to PSNR. Equivalent performance to PSNR is considered appropriate in this case because PSNR is an FR metric. This calculation is performed for each dataset and the superset separately, and such models are indentified with a "1". The final line indicates the total number of datasets where that model performed significantly better than PSNR.

Table 15. Pearson Correlation of HDTV RR Models for Individual Datasets and Aggregated Superset

Dataset	PSNR	Yonsei56k	Yonsei128k	Yonsei256k
vqeghd1	0.85	0.77	0.77	0.77
vqeghd2	0.58	0.61	0.61	0.61
vqeghd3	0.85	0.86	0.86	0.86
vqeghd4	0.81	0.75	0.75	0.75
vqeghd5	0.72	0.55	0.56	0.56
vqeghd6	0.79	0.90	0.90	0.90
superset	0.78	0.77	0.77	0.77
average	0.77	0.74	0.74	0.74

Table 16. RMSE of HDTV RR Models for Individual Datasets and Aggregated Superset

Dataset	PSNR	Yonsei56k	Yonsei128k	Yonsei256k
vqeghd1	0.65	0.79	0.79	0.79
vqeghd2	0.84	0.82	0.82	0.82
vqeghd3	0.59	0.57	0.58	0.57
vqeghd4	0.66	0.74	0.74	0.74
vqeghd5	0.77	0.92	0.92	0.92
vqeghd6	0.65	0.47	0.47	0.47
superset	0.71	0.73	0.73	0.73
average	0.69	0.72	0.72	0.72

Table 17. HDTV RR Models in the Top Performing Group (1) versus Models with Less Accurate Performance (0); Computed from RMSE

Dataset	PSNR	Yonsei56k	Yonsei128k	Yonsei256k
vqeghd1	1	0	0	0
vqeghd2	1	1	1	1
vqeghd3	1	1	1	1
vqeghd4	1	1	1	1
vqeghd5	1	0	0	0
vqeghd6	1	1	1	1
superset	1	1	1	1
Total	6	4	4	4

Table 18. HDTV RR Models that Performed Significantly Equivalent To or Better than PSNR (1) versus Models with Performance Statistically Worse than PSNR (0); Computed from RMSE

Dataset	PSNR	Yonsei56k	Yonsei128k	Yonsei256k
vqeghd1	1	0	0	0
vqeghd2	1	1	1	1
vqeghd3	1	1	1	1
vqeghd4	1	1	1	1
vqeghd5	1	0	0	0
vqeghd6	1	1	1	1
superset	1	1	1	1
Total	6	4	4	4

Table 19 lists the RMSE for each RR model, for subdivisions of the superset. These subdivisions divide the data by coding type (H.264 or MPEG-2) as well as by the presence of transmission errors (Errors) or whether the HRC contained coding artifacts only (Coding). Because the experiments were not designed to have these variables evenly span the full range of quality, only RMSE are presented for these subdivisions. Objective models were not re-fitted prior to this analysis (i.e., the polynomial fit for the entire superset was used for each of these RMSE calculations). Table 20 identifies (for each type of HRC) all models that were statistically equivalent to the top performing model. Table 21 identifies (for each type of HRC) all models that were statistically better than or equivalent to PSNR.

Table 19. RMSE of HDTV RR Models for Aggregated Superset, Divided by HRC Type

HRC Type	PSNR	Yonsei56k	Yonsei128k	Yonsei256k
H.264 Coding	0.75	0.65	0.65	0.65
H.264 Error	0.67	0.86	0.85	0.86
mpeg-2 Coding	0.78	0.81	0.81	0.80
mpeg-2 Error	0.66	0.68	0.68	0.68
coding	0.75	0.69	0.69	0.69
error	0.67	0.79	0.78	0.79

Table 20. HDTV RR Models in the Top Performing Group (1) by HRC Type Versus Models with Less Accurate Performance (0); Computed from RMSE

HRC Type	PSNR	YonseiRR56k	YonseiRR128k	YonseiRR256k
H.264 Coding	1	1	1	1
H.264 Error	1	0	0	0
mpeg-2 Coding	1	1	1	1
mpeg-2 Error	1	1	1	1
Coding	1	1	1	1
Error	1	0	0	0

Table 21. HDTV RR Models Equivalent To or Better Than PSNR (1) by HRC Type Versus Models with Significantly Worse Performance than PSNR (0); Computed from RMSE

HRC Type	PSNR	YonseiRR56k	YonseiRR128k	YonseiRR256k
H.264 Coding	1	1	1	1
H.264 Error	1	0	0	0
mpeg-2 Coding	1	1	1	1
mpeg-2 Error	1	1	1	1
Coding	1	1	1	1
Error	1	0	0	0

Table 22 identifies the lower and upper confidence interval for each correlation and RMSE mentioned in Table 15 and 16. In this table, "lower bound" is defined to be less accurate than the indicated measurement, and "upper bound" is defined to be more accurate than the indicated measurement.

Table 22. Confidence Intervals for RR Models' Pearson Correlation and RMSE

PSNR						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.85	0.80	0.89	0.65	0.73	0.59
vqeghd2	0.58	0.45	0.68	0.84	0.96	0.75
vqeghd3	0.85	0.80	0.89	0.59	0.67	0.53
vqeghd4	0.81	0.75	0.86	0.66	0.74	0.59
vqeghd5	0.72	0.63	0.79	0.77	0.87	0.69
vqeghd6	0.79	0.73	0.84	0.65	0.73	0.58
superset	0.78	0.75	0.80	0.71	0.75	0.68
Yonsei56k						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.77	0.69	0.83	0.79	0.89	0.71
vqeghd2	0.61	0.49	0.70	0.82	0.93	0.73
vqeghd3	0.86	0.82	0.90	0.57	0.65	0.51
vqeghd4	0.75	0.67	0.81	0.74	0.84	0.67
vqeghd5	0.55	0.43	0.65	0.92	1.04	0.83
vqeghd6	0.90	0.86	0.92	0.47	0.53	0.42
superset	0.77	0.74	0.79	0.73	0.77	0.70
Yonsei128k						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.77	0.69	0.83	0.79	0.89	0.71
vqeghd2	0.61	0.49	0.71	0.82	0.93	0.73
vqeghd3	0.86	0.81	0.90	0.58	0.65	0.52
vqeghd4	0.75	0.67	0.81	0.74	0.84	0.67
vqeghd5	0.56	0.44	0.66	0.92	1.03	0.82
vqeghd6	0.90	0.86	0.92	0.47	0.53	0.42
superset	0.77	0.74	0.79	0.73	0.77	0.70
Yonse256k						
Dataset	**Correlation**	**Lower**	**Upper**	**RMSE**	**Lower**	**Upper**
vqeghd1	0.77	0.70	0.83	0.79	0.89	0.71
vqeghd2	0.61	0.49	0.71	0.82	0.93	0.73
vqeghd3	0.86	0.82	0.90	0.57	0.64	0.51
vqeghd4	0.75	0.67	0.81	0.74	0.84	0.67
vqeghd5	0.56	0.44	0.66	0.92	1.03	0.83
vqeghd6	0.90	0.86	0.92	0.47	0.53	0.42
superset	0.77	0.74	0.79	0.73	0.77	0.70

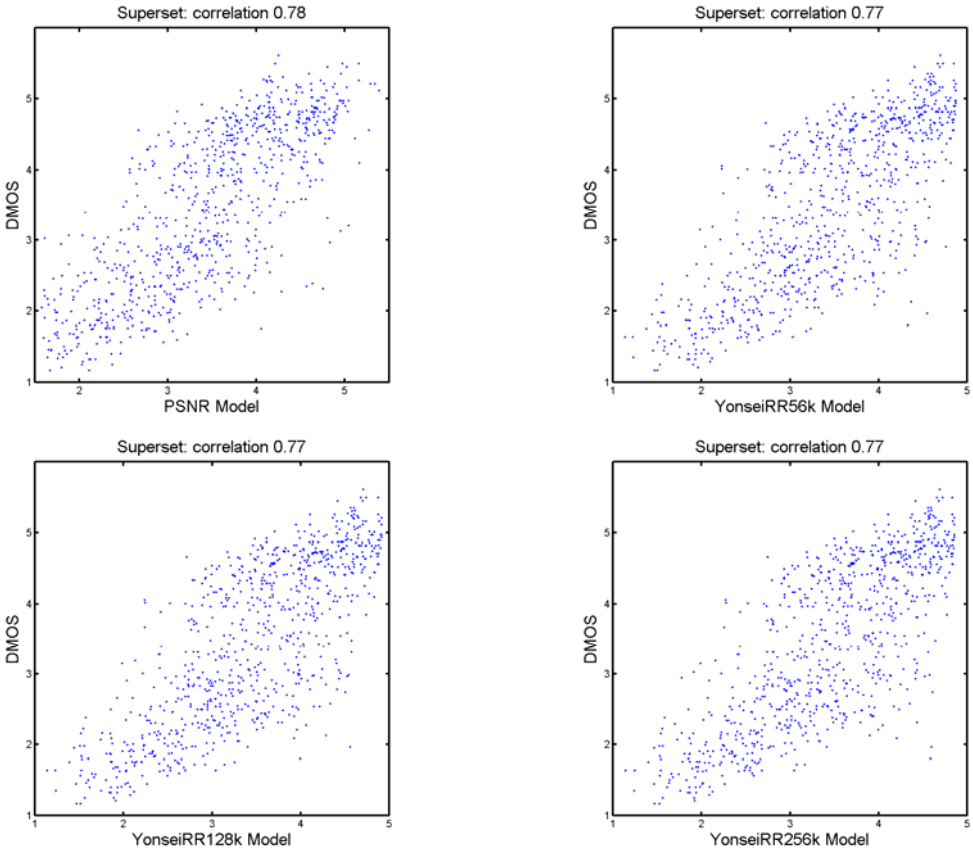
Table 23 identifies the coefficients for the monotonic, polynomial fitting between each model and each dataset. All analysis on the superset for one model uses the same mapping (i.e., superset analysis and the division by HRC type). These coefficients were computed using the following equation:

$$DMOSp = A3 \cdot x^3 + A2 \cdot x^2 + A1 \cdot x + A0$$

Table 23. Monotonic Polynomial Coefficients for RR Models

PSNR				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	-0.0000306550279250	-0.0019879070241439	0.3921369753140520	-5.7280907926448500
vqeghd2	-0.0001334181504514	0.0111764931035134	-0.1877690908303150	1.9953778604294600
vqeghd3	-0.0001932109488014	0.0197073219079854	-0.5006502641704290	5.3895615601423200
vqeghd4	-0.0001935191412653	0.0131947404548850	-0.1153986433564110	0.6574532383968700
vqeghd5	-0.0000148995264923	0.0061596181685647	-0.1807405959360720	3.6252029746604300
vqeghd6	-0.0001813435735504	0.0123033171141708	-0.0165560675513631	-3.0727162004210500
superset	-0.0001427599450031	0.0137910530512573	-0.2912028187262140	3.3051795730244100
Yonsei56k				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	-0.0000701879413315	0.0037050830032694	0.1335361040130230	-2.2402753306053600
vqeghd2	-0.0000354990211253	-0.0001212653331526	0.2537546464691130	-3.9606475541513200
vqeghd3	-0.0001834176223618	0.0181658646237622	-0.4409542946631760	4.6646590423521700
vqeghd4	0.0000188867914360	-0.0086811987375009	0.6605641243404790	-9.2470759358754500
vqeghd5	0.0000031852013340	-0.0049959333564710	0.4386501693559710	-5.6386455711061800
vqeghd6	-0.0002693865304109	0.0279477448057628	-0.7743755024948150	7.2914694188770500
superset	-0.0000982615035905	0.0081043987909922	-0.0642502778917543	0.1049825273901460
Yonsei128k				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	-0.0000626846363523	0.0028549443001271	0.1641146502580400	-2.5872467454270300
vqeghd2	-0.0000355003434871	0.0009214721727024	0.1871565502554760	-2.9176047545811500
vqeghd3	-0.0001887855268650	0.0188546356753588	-0.4695721160486150	5.0526066810967100
vqeghd4	0.0000280195121077	-0.0096965282216900	0.6955691253267590	-9.6180948500647900
vqeghd5	-0.0000002998910098	-0.0046075579022733	0.4241816579500430	-5.4491373284590500
vqeghd6	-0.0002686378820118	0.0278678619663216	-0.7720020815433460	7.2792609064531200
superset	-0.0000913686542645	0.0074705832596161	-0.0454851180411333	-0.0710260566542785
Yonsei256k				
Dataset	**A3**	**A2**	**A1**	**A0**
vqeghd1	-0.0000628064653793	0.0028552911039717	0.1657909010064500	-2.6412693478522100
vqeghd2	-0.0000645499503195	0.0029261867791390	0.1503197706909690	-2.8142319592391300
vqeghd3	-0.0001835902239905	0.0181727127675506	-0.4403445968265440	4.6459870916707500
vqeghd4	0.0000589972911484	-0.0132500850310157	0.8277622127101000	-11.2132909586595000
vqeghd5	0.0000507018141148	-0.0099999001029493	0.6033904762319630	-7.2811874051790300
vqeghd6	-0.0002684338164053	0.0278448506319658	-0.7712314401566040	7.2731102597870900
superset	-0.0000967680720945	0.0078849337323418	-0.0555373849819267	0.0158656819579036

Following are plots for each RR model, showing the model's fitted value on the X-axis and the superset DMOS on the Y-axis. PSNR is plotted for comparison purposes. The superset correlation is shown above the plot. Please note that due to the transformation that was applied to the individual datasets while forming the aggregated superset, some DMOS values are greater than 5.0.



8.4. NR Models

Three NR models were submitted to VQEG. All three were withdrawn.

9. Secondary Data Analysis:

The secondary analysis was performed using techniques described in the ITU temporary document TD 12rev1a "Statistical Evaluation Procedure for P.OLQA v.1.0.a". These techniques are also described in ftp://vqeg.its.bldrdoc.gov/Documents/VQEG_Krakow_Jun10/MeetingFiles/HDTV/Swissqual_Epsilon_Insensitive_RMSE.pdf These statistics were developed within ITU-T P.OLQA project.

Two different methods are carried out in this analysis which will be described in the following sections.

9.1. Overview

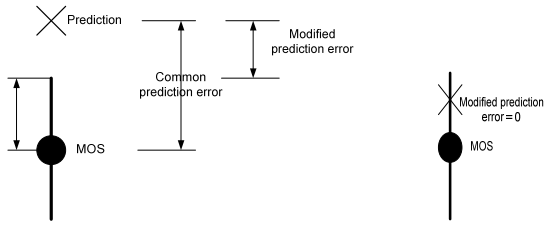
9.1.1. Correlation coefficients

Correlation coefficients as used in this report are not sensitive to the confidence of the underlying subjective data. In addition correlation coefficients are highly dependent on the distribution of the subjective scores across the scale. Finally, the aggregation of correlation coefficients is not in general a useful performance indicator due to their high non-linearity. For these reasons more advanced statistics were applied in the VQEG HDTV Phase I evaluation process.

9.1.2. Epsilon-insensitive root mean square error rmse*

Epsilon-insensitive RMSE (rmse*) is calculated as the traditional RMSE but small differences to the target value will not be counted. This rmse* considers only differences related to epsilon-wide band around the target value. This 'epsilon' is defined as the 95% confidence interval of the subjective MOS value. By definition, the uncertainty of the MOS is taken into account in this evaluation.

The rmse* is calculated on a prediction error as illustrated in the figure below.



9.1.3. Ci95 weighted root mean square error rmse**

As a second metric based on the absolute prediction error, is a rmse** value that is later aggregated to rmse_tot*.

Also the rmse** is based on the traditional RMSE. It is weighted by the ci95% interval for taking into account uncertainties of the individual MOS value used for calculating the prediction error. The resulting rmse** is not a MOS difference but is rather a dimensionless value. In contrast to rmse* it considers prediction errors within the confidence interval but weights them lower than the traditional RMSE.

9.1.4. Other Considerations and Warnings

RMSE values, rmse* values, rmse_tot* values and Pearson correlation values cannot be directly compared, because these are all reported on different scales.

As the number of viewers increases toward infinity, RMSE and rmse* will converge toward the same value.

These are new statistics that are presented here to help people understand and explore these new ways of analyzing objective models' performance using subjective data.

9.2. Description of the Evaluation Based on Epsilon Insensitive RMSE

The "Epsilon Insensitive RMSE" takes the uncertainty of the subjects into account. This is important since the objective models will not be able to predict the average opinion score more accurate than the average subjects themselves. It is calculated similar to the traditional root mean square error but the 95% confidence interval of the subjective MOS value is included into evaluation.

The Epsilon Insensitive RMSE, rmse*, is defined as follows:

$$rmse^* = \sqrt{\left(\frac{1}{N-d}\sum_N Perror(i)^2\right)}$$

whereas

$$Perror(i) = \max(0, \left|MOSLQS(i) - MOSLQO(i)\right| - ci_{95}(i))$$

and

$$ci_{95} = t(0.05, M)\frac{\sigma}{\sqrt{M}}$$

In the above formula MOSLQS represents the subjective DMOS value associated to the video clip i, ci_{95} is the confidence interval and σ the standard deviation related to the subjective DMOS value. t(0.05,M) is the 95 percentile value of the student t distribution for the two tailed test and M the number of viewers. MOSLQO represents the objective DMOS value associated to the video clip. The index i denotes the video sample in the experiment, N the total number of video samples in the experiment and d the number of freedom.

Note that Perror() will be 0 if the predicted Objective DMOS value is within the confidence interval of the subjective test and greater than 0 if outside.

A distance measure, relative to the best performing model, which is the model with the lowest rmse*, is carried out to compare models on an experiment basis. The Distance is defined as:

$$d_{k,v} = \max(0, rmse^*_{k,v}{}^2 - rmse^*_{k,b}{}^2 \times \mathrm{F}(0.05, N_k, N_k))$$

where $rmse^*_{k,b}$ denotes $rmse^*$ of the best performing model for experiment k. The index v denotes the objective model and $\mathrm{F}(0.05, N_k, N_k)$ is the tabulated value of the F-distribution for N_k degrees of freedom and 95% significance level. N_k is set to the number of considered samples in experiment k.

The aggregation across experiments is often done using different weights for known and unknown databases. However in our case all experiments are considered as unknown experiments. Therefore the aggregated value across experiments for each model is the mean of the distances calculated above.

$$p_v = \frac{1}{M}\sum_{k=1}^{M} d_{k,v}$$

With M equal to the number of experiments.

The model which achieves the lowest p_v is defined as the best performing model within this test. To determine if a model is statistically equivalent to the best performing model, a statistical significance test will be applied on the aggregated distance measure.

$$t_v = \max(0, \frac{p_v}{\left(p_{\min} + c\right)} - \mathrm{F}(0.05, K, K))$$

Where p_v is the aggregated distance for model v and $p_{\min}$ is the lowest p_v in the evaluation. To avoid division by 0 the constant c is set to 0.0004. $\mathrm{F}(0.05, K, K)$ is the tabulated value of the F-distribution for K degrees of freedom and 95% significance level. K is set to the total number of experiments in the test.

The model v is considered statistically equivalent to the best performing model if $t_v = 0$. If $t_v > 0$, the model is considered to be statistically different than the best performing model.

9.3. Description of the Evaluation Based on the Statistical Significance of the rmse_tot* Across All Experiments

This section describes a second metric based on the relation between the prediction error and the confidence intervals.

In this test, for each candidate model v the absolute prediction error, rmse_tot* is calculated:

$$rmse\_tot_v^* = \frac{1}{K}\sum_K rmse^{**}_{k,v}$$

with

$$rmse^{**}_{k,v} = \sqrt{\frac{1}{N_k - d}\sum_{N_k}\left(\frac{MOSLQS(i)_k - MOSLQO(i)_{k,v}}{\max(ci_{95}(i)_k, c)}\right)^2}$$

and

$$ci_{95} = t(0.05, M)\frac{\sigma}{\sqrt{M}}$$

In the above function, k represents the experiment number. The index i denotes the sequence number in the experiment k. N_k determines the total number of sequences in experiment k. K denotes the number of experiments and d the number of freedoms. The constant c set to 0.1 avoids divisions by very small ci95 values as they usually appear at the low end of the scale.

MOSLQS represents the subjective DMOS value associated to the video clip. ci_{95} the confidence interval of the subjective DMOS value, t(0.05,M) is the 95 percentile value of the student t distribution for the two tailed test and M the number of viewers. σ stands for the standard deviation of the individual scores associated to the video clip. MOSLQO represents the Objective DMOS value for that video clip.

The best performing model is defined as the model with the lowest rmse_tot*.

In a next step a statistical significance test is carried out to determine whether the models are statistically equivalent to the best performing model.

Let

$$r_v = \max(0, \frac{rmse\_tot_v^{*2}}{\left(rmse\_tot_{\min}^{*2} + c\right)} - \mathrm{F}(0.05, T, T))$$

To avoid division by 0 the constant c is set to 0.0004. $\mathrm{F}(0.05, T, T)$ is the tabulated value of the F-distribution for T degrees of freedom and 95% significance level. T is set to the total number of sequences in the test. $rmse\_tot^*_{min}$ is the rmse_tot* of the best performing model.

The model v is considered statistically equivalent to the best performing model if $r_v = 0$. In case that $r_v > 0$, the model is considered to be statistically different than the best performing model.

9.4. Benchmark of the Subjective Data

The subjective and objective data is described in detail in chapter 11. At this point, for a better overview of the values margin, the minimum, mean and maximum of DMOS, standard deviation σ and confidence interval ci_{95} for each database and the superset are presented in table 24.

Table 24: Benchmark of subjective data

Dataset	DMOS			σ			ci_{95}		
	Min	mean	Max	Min	Mean	Max	Min	Mean	Max
Vqeghd1	1.25	3.31	5.17	0.44	0.77	1.17	0.19	0.33	0.49
Vqeghd2	1.38	2.92	5.00	0.42	0.82	1.24	0.18	0.34	0.52
Vqeghd3	1.58	3.55	5.25	0.46	0.85	1.23	0.20	0.36	0.52
Vqeghd4	1.58	3.60	5.13	0.54	0.83	1.31	0.23	0.35	0.55
Vqeghd5	1.46	3.53	5.33	0.29	0.81	1.18	0.12	0.34	0.50
Vqeghd6	1.25	3.31	5.17	0.44	0.77	1.17	0.19	0.33	0.49
Superset	1.17	3.45	5.61	0.29	0.80	1.40	0.12	0.34	0.59

9.5. FR Models Evaluation Based on Epsilon Insensitive RMSE

The Evaluation based on Epsilon Insensitive RMSE was carried out on the remaining five models presented in this report. For comparison purpose PSNR was included as an additional model. The calculation was done based on the monotonic, polynomial fitted results DMOSp as calculated in section 8, Official ILG Data Analysis. The number of freedom was set to four.

Table 25 lists the Epsilon Insensitive RMSE, rmse*, for PSNR and each FR Model for the six databases. The final line of this table indicates the average value for that model. The top performing models are marked red. For each database the distances between models are presented in table 26, whereas the aggregated distances are presented in table 27. Table 28 contains the statistical significance values for PSNR and the FR Models.

Table 25: Epsilon-insensitive root mean square error RMSE*

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
vqeghd1	154	0.41482	0.46949	0.82403	0.36367	0.57277	0.57747
vqeghd2	135	0.61036	0.57185	0.58548	0.36000	0.49186	0.52828
vqeghd3	154	0.35149	0.37061	0.53868	0.36000	0.42362	0.28000
vqeghd4	155	0.40581	0.43804	0.45085	0.40418	0.32381	0.48033
vqeghd5	155	0.50781	0.59116	0.65770	0.35867	0.37197	0.59115
vqeghd6	155	0.41851	0.49651	0.39054	0.24345	0.29125	0.24270
Average	151.333333	0.45147	0.48961	0.57455	0.34833	0.41255	0.44999

Table26: Distance Measure

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
vqeghd1	154	0.00000	0.04788	0.50648	0.00000	0.15553	0.16093
vqeghd2	135	0.20035	0.15482	0.17060	0.00000	0.06974	0.10689
vqeghd3	154	0.02126	0.03507	0.18790	0.02732	0.07718	0.00000
vqeghd4	155	0.02801	0.05520	0.06659	0.02668	0.00000	0.09404
vqeghd5	155	0.09018	0.18179	0.26488	0.00000	0.00000	0.18177
vqeghd6	155	0.09837	0.16974	0.07574	0.00000	0.00805	0.00000
Overall	908	0.04511	0.08100	0.17139	0.00000	0.01148	0.04378

Table 27: Average Distance over Sets

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR

		1	3	4	5	6	7
Average	908	0.07303	0.10742	0.21203	0.00900	0.05175	0.09061

Table 28: Statistical Significance of the aggregated distance measure

Dataset	NrDatabases	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
Significance	6	2.71861	6.37674	17.50571	0.00000	0.45473	4.58847

All models except the best performing model show statistical differences greater than zero. As a result of this test, none of the models may be considered statistically equivalent to the best performing model.

9.6. FR Models Evaluation Based on the Statistical Significance of rmse_tot* Performed on Individual Datasets

The evaluation based on the statistical significance of rmse_tot* was carried out on the remaining five models presented in this report. For comparison purpose PSNR was included as an additional model. The calculation was done based on the monotonic, polynomial fitted results DMOSp as calculated in section 8, Official ILG Data Analysis.

Table 29 lists the rmse** for PSNR and each FR Model across the six databases. The final line of this table indicates the average value for that model. The top performing models are marked red. In table 30 the aggregated distance measure rmse_tot* for each model is presented. Table 31 shows the results of the statistical significance test applied on the aggregated distance measure.

Table 29: Absolute Prediction Error RMSE** for Individual Datasets

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
vqeghd1	154	2.12949	2.35600	3.78346	1.91698	2.54490	2.69881
vqeghd2	135	2.79188	2.59324	2.65697	1.77764	2.33990	2.44642
vqeghd3	154	1.73316	1.83435	2.41559	1.77764	1.96692	1.53929
vqeghd4	155	1.92021	2.04082	2.07757	1.91495	1.50182	2.23408
vqeghd5	155	2.46343	2.72684	2.93620	1.79184	1.86746	2.67687
vqeghd6	155	2.05458	2.28889	1.91917	1.42696	1.52204	1.46146
Average	151.333333	2.18212	2.30669	2.63149	1.76767	1.95717	2.17616

Table 30: Aggregated Distance Measure rmse_tot* for Individual Datasets

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
rmse_tot*	908	2.18212	2.30669	2.63149	1.76767	1.95717	2.17616

Table 31: Statistical Significance r for Individual Datasets

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
Significance	908	0.36116	0.53456	1.03195	0.00000	0.07240	0.35309

All models except the best performing model show a statistical significance value greater than zero. Therefore, as a result of this statistical significance test based on rmse_tot* none of the models may be considered statistically equivalent to the best performing model.

9.7. FR Models Evaluation Based on the Statistical Significance of rmse_tot* performed on Aggregated Superset

Based on the aggregated FR superset data as used in section 8, official ILG Data Analysis, a second statistical significance of rmse_tot* analysis will be presented in tables 32 to 34.

Table 32 lists the rmse** for PSNR and each FR Model across the aggregated supersets. The top performing models are marked red. In table 33 the aggregated distance measure rmse_tot* is presented. Table 34 shows the results of the statistical significance test applied on the aggregated distance measure.

Table 32: Absolute Prediction Error RMSE** for aggregated superset

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
Superset	828	2.33431	2.37304	2.88700	1.77973	2.06492	2.32867

Table 33: Aggregated distance measure rmse_tot* for aggregated superset

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
rmse_tot*	828	2.33431	2.37304	2.88700	1.77973	2.06492	2.32867

Table 34: Statistical Significance r for aggregated superset

Dataset	NrSamples	PSNR	NTT	Opticom	Swissqual	Tektronix	YonseiFR
		1	3	4	5	6	7
Significance	828	0.54640	0.60219	1.42959	0.00000	0.18369	0.53834

Only the best performing model shows a statistical significance value equal to 0.
As a result of the statistical rmse_tot* analysis performed on the aggregated superset none of the models may be considered statistically equivalent to the best performing model.

9.8. RR Models Evaluation Based on Epsilon Insensitive RMSE

The evaluation based on Epsilon Insensitive RMSE was carried out on the three submitted models presented in this report. For comparison purpose PSNR was included as an additional model. The calculation was done based on the monotonic, polynomial fitted results DMOSp as calculated in section 8, Official ILG Data Analysis. The number of freedom was set to four.

For each database the Epsilon Insensitive RMSE for PSNR and each RR Model is presented in table 35. The final line of this table indicates the average value for that model. The top performing models are marked red. For each database, the distances between models are presented in table 36, whereas the aggregated distances are presented in table 37. Table 38 contains the statistical significance values for PSNR and the RR Models.

Table 35: Epsilon-insensitive root mean square RMSE*

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10
Vqeghd1	154	0.41482	0.54395	0.54314	0.53840
Vqeghd2	135	0.61036	0.56821	0.56745	0.56668
Vqeghd3	154	0.35149	0.33206	0.33775	0.32932
Vqeghd4	155	0.40581	0.47355	0.47427	0.47428
Vqeghd5	155	0.50781	0.64223	0.63628	0.63651
Vqeghd6	155	0.41851	0.24415	0.24381	0.24401
Average	151.33	0.45147	0.46736	0.46712	0.46487

Table 36: Distance Measure

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10

Vqeghd1	154	0.00000	0.07139	0.07051	0.06538
Vqeghd2	135	0.00000	0.00000	0.00000	0.00000
Vqeghd3	154	0.00000	0.00000	0.00000	0.00000
Vqeghd4	155	0.00000	0.00958	0.01026	0.01028
Vqeghd5	155	0.00000	0.07633	0.06872	0.06901
Vqeghd6	155	0.09766	0.00000	0.00000	0.00000
Overall	908	0.00000	0.00000	0.00000	0.00000

Table 37: Average Distance

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10
Average	908	0.01628	0.02622	0.02492	0.02411

Table 38: Statistical Significance of the aggregated distance measure

Dataset	NrDatabases	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10
Significance	6	0.00000	0.00000	0.00000	0.00000

According to table 36 the significance of the aggregated distance measure is 0 for all four models and PSNR. As the result of the "epsilon insensitive RMSE" test, all tested models may be considered statistically equivalent.

9.9. RR Models Evaluation Based on the Statistical Significance of rmse_tot* performed on Individual Databases

The evaluation based on the statistical significance of rmse_tot* was carried out on the three submitted models presented in this report. For comparison purpose PSNR was included as an additional model. The calculation was done based on the monotonic, polynomial fitted results DMOSp as calculated in section 8, Official ILG Data Analysis.

Table 39 lists the rmse** for PSNR and the RR Model across the six databases. The final line of this table indicates the average value for that model. The top performing models are marked red. In table 40 the aggregated distance measure rmse_tot* for each model is presented. Table 41 shows the results of the statistical significance test applied on the aggregated distance measure.

Table 39: Absolute Prediction Error RMSE** for individual datasets

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10
Vqeghd1	154	2.12949	2.57128	2.56713	2.54575
Vqeghd2	135	2.79188	2.61577	2.60202	2.60749
Vqeghd3	154	1.73316	1.69658	1.71059	1.68949
Vqeghd4	155	1.92021	2.20600	2.21057	2.20551
Vqeghd5	155	2.46343	2.89568	2.87222	2.89252
Vqeghd6	155	2.05458	1.46912	1.46611	1.46507
Average	151.33	2.18212	2.24240	2.23810	2.23431

Table 40: Aggregated distance measure rmse_tot* for individual datasets

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10
rmse_tot*	908	2.18212	2.24240	2.23810	2.23431

Table 41: Statistical Significance r for individual datasets

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10

Significance	908	0.00000	0.00000	0.00000	0.00000

The statistical significance is 0 for all four models and PSNR. As a result of this statistical significance test based on rmse_tot* all four models and PSNR may be considered statistically equivalent.

9.10. RR Models Evaluation Based on the Statistical Significance of rmse_tot* performed on Aggregated Superset

Based on the aggregated RR superset data as used in section 8, Official ILG Data Analysis, a second statistical significance of rmse_tot* analysis will be presented in tables 42 to 43.

Table 42 lists the rmse** for PSNR and each RR Model across the aggregated supersets. The top performing models are marked red. In table 43 the aggregated distance measure rmse_tot* is presented. Table 44 shows the results of the statistical significance test applied on the aggregated distance measure.

Table 42: Absolute Prediction Error RMSE** for aggregated superset

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10
Supperset	828	2.33431	2.31135	2.30361	2.31427

Table 43: Aggregated distance measure rmse_tot* for aggregated superset

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10
rmse_tot*	828	2.33431	2.31135	2.30361	2.31427

Table 44: Statistical Significance r for aggregated superset

Dataset	NrSamples	PSNR	YonseiRR56k	YonseiR128k	YonseiRR256k
		1	8	9	10
Significance	828	0.00000	0.00000	0.00000	0.00000

All models show a statistical significance value equal to 0.
As a result of the statistical rmse_tot* analysis performed on the aggregated superset all of the models may be considered statistically equivalent to the best performing model.

9.11. Secondary analysis for NR Models

Three NR models were submitted to VQEG. All three were withdrawn.

10. Subjective and Objective Data:

The subjective data (MOS and DMOS) and objective data for each model presented are available in companion document “VQEG_HDTV_Final_Report_Data.xls”.

Appendix I: Model Descriptions

Note: The proponent comments are not endorsed by VQEG. They are presented in this Appendix to give the Proponents a chance to discuss their results and should not be quoted out of this context.

Appendix I.1 : NTT

1. Brief description of model

The NTT model accurately estimates subjective quality using a precise alignment process and a video quality algorithm that reflects human visual characteristics based on the effect of codecs, bit-rate, frame-rate, and video quality distorted by packet loss. It is divided into two units: a video-alignment unit, and subjective quality-estimation unit. The video-alignment unit filters the video sequences based on the effect of video capturing and post-processing of the decoder and matches pixels between reference-video and processed video sequences in the spatial temporal directions. Afterwards, it matches frames between reference and processed video sequences based on the effect of video frame skipping and freezing. The subjective quality-estimation unit calculates the objective video quality that reflects human visual characteristics by using (i) a spatial degradation parameter based on four parameters that reflect the presence of overall noise, spurious edges, localized motion distortion, and localized spatial distortions caused by packet loss and (ii) a temporal degradation parameter, which reflects frame-rate freezing and variation.

2. Model performance

The NTT model performed more poorly than the PSNR model in some tests in the VQEG validation process. The primary reason for this is that there are particular source sequences we did not expect (e.g. the video sequences involving telop or rotation of objects). If these particular sequences[3] are removed, the NTT model outperforms the PSNR model, as shown in Table I.1.

Table I.1: Results omitted in particular video sequences

Correlation	PSNR	Method1	NTT	Method2	Method3	Method4	Method5
vqeghd1	0.84	0.84 ③	0.85 ②	0.59 ⑥	0.87 ①	0.75 ④	0.72 ⑤
vqeghd2	0.53	0.52 ⑥	0.70 ②	0.56 ⑥	0.83* ①	0.68 ③	0.66 ④
vqeghd3	0.87	0.53 ⑥	0.89 ③	0.71 ⑤	0.94 ①	0.87 ④	0.91 ②
vqeghd4	0.88	0.87 ②	0.83 ③	0.74 ⑤	0.82 ④	0.88 ①	0.72 ⑥
vqeghd5	0.72	0.72 ③	0.69 ④	0.49 ⑥	0.88 ①	0.82 ②	0.58 ⑤
vqeghd6	0.87	0.90 ②	0.82 ⑥	0.87 ⑤	0.92 ①	0.92 ④	0.90 ③
all	0.78	0.73 ⑤	0.80 ③	0.66 ⑥	0.88 ①	0.82 ②	0.75 ③
superset	0.67	0.19 ⑥	0.69 ⑤	0.92 ①	0.80 ③	0.72 ④	0.82 ②
H.264 Coding	0.70	0.51 ⑥	0.81 ①	0.75 ③	0.72* ⑤	0.78 ②	0.81 ①
H.264 Error	0.69	0.59 ⑥	0.72 ③	0.63 ⑤	0.82* ①	0.71 ③	0.76 ②
mpeg-2 Coding	0.65	0.66 ④	0.71 ③	0.64 ⑥	0.75* ②	0.60 ⑤	0.77 ①
mpeg-2 Error	0.86	0.82 ③	0.87 ①	0.63 ⑥	0.87* ①	0.79 ⑤	0.79 ④

RMSE	PSNR	Method1	NTT	Method2	Method3	Method4	Method5
vqeghd1	0.66	0.68 ③	0.64 ②	0.99 ⑥	0.60 ①	0.80 ④	0.84 ⑤
vqeghd2	0.87	0.88 ⑥	0.74 ②	0.86 ⑤	0.57* ①	0.75 ③	0.78 ④
vqeghd3	0.56	0.96 ⑥	0.54 ③	0.81 ⑤	0.41 ①	0.57 ④	0.48 ②
vqeghd4	0.54	0.56 ②	0.63 ③	0.75 ⑤	0.64 ④	0.52 ①	0.77 ⑥
vqeghd5	0.76	0.76 ③	0.81 ④	0.97 ⑥	0.53 ①	0.63 ②	0.90 ⑤
vqeghd6	0.54	0.42 ③	0.60 ⑥	0.53 ⑤	0.41 ②	0.40 ①	0.44 ④
all	0.66	0.71 ⑤	0.66 ③	0.82 ⑥	0.52 ①	0.61 ②	0.70 ④
superset	0.87	1.22 ⑥	0.91 ⑤	0.69 ①	0.72 ③	0.78 ④	0.70 ②
H.264 Coding	0.74	0.92 ⑥	0.67 ②	0.78 ④	0.76* ⑤	0.66 ①	0.69 ③
H.264 Error	0.67	0.72 ④	0.69 ③	0.95 ⑥	0.60* ①	0.61 ②	0.82 ⑤
mpeg-2 Coding	0.63	0.60 ①	0.71 ④	0.77 ⑤	0.63* ②	0.63 ②	0.80 ⑥
mpeg-2 Error	0.63	0.73 ③	0.63 ①	0.96 ⑥	0.63* ①	0.75 ③	0.76 ⑤

[3] The particular sequences follow six video sequences. [vqeghd1_src02, vqeghd2_src04, vqeghd2_src09, vqeghd3_src06, vqeghd5_src06, vqeghd6_src04]

Appendix I.2 : Proponent B, Opticom

Overview

PEVQ is a very robust model which is designed to predict the effects of transmission impairments on the video quality as perceived by a human subject. Its main targets are Mobile Multimedia applications and IPTV. The key features of PEVQ are:

- (fast and reliable) temporal alignment of the input sequences based on multi dimensional feature correlation analysis with limits that reach far beyond those tested by VQEG, especially with regard to the amount of time clipping, frame freezing and frame skipping which can be handled.
- Full frame spatial alignment
- Color alignment algorithm based on cumulative histograms
- Enhanced framerate estimation and rating
- Detection and perceptually correct weighting of frame freezes and frame skips.
- Only four indicators are used to detect the video quality. Those indicators operate in different domains (temporal, spatial, chrominance) and are motivated by the Human Visual System. Perceptual masking properties of the HVS are modelled at several stages of the algorithm. These indicators are integrated using a sophisticated spatial and temporal integration algorithm.

In its first stage the algorithm performs alignment steps in various domains and collects information on frozen or skipped frames. In a second step the now synchronized and equalized images are compared for visual differences in the luminance as well as in the chrominance domain, taking masking effects and motion into account. This results in a set of indicators which all describe certain quality aspects. The last step is finally the integration of the individual indicators by non-linear functions in order to derive the final MOS.

Due to the low number of indicators and the resulting low degree of freedom the model can hardly be over trained and is very robust. PEVQ was developed for Multimedia applications by Roland Bitto of OPTICOM and is built on an earlier TV quality measure developed by Dr. John Beerends and Andries Hekstra from KPN. PEVQ can be efficiently implemented without sacrificing the prediction accuracy and has been widely adopted by the Mobile telecom industry.

Comment on Results for HD Content

Due to a limited number of subjective HD databases, PEVQ V3.4, which was tested by VQEG, was mostly trained on much smaller resolutions than required for the HD test and it was therefore certainly performing far below its potential. The real potential of the algorithm can be seen by looking at the results of the VQEG Multimedia project, which resulted in the standardization of PEVQ in ITU-T Rec. J.247. Consequently, now that more HD databases became available, OPTICOM further improved PEVQ. By applying some very minor modifications to the algorithm, the performance could be increased significantly. The accuracy now achieved by PEVQ V4 and above is equal to the performance of the best model in the VQEG test. Since those results were achieved after the validation databases became available, those results can of course not be taken into account within the current benchmark.

Appendix I.3 : Swissqual

VQuad-HD : Comments on Model Performance

Model description:

The model takes as input a reference and a processed video sequence. Score estimation is based on the following steps:

1. First, the video sequences are preprocessed. In particular, the frames are subsampled.
2. A spatio-temporal frame alignment between reference and processed video sequence is performed.
3. A local similarity and a difference measure inspired by visual perception, a jerkiness measure, and a blockiness measure are computed.
4. The quality score is estimated based on a non-linear aggregation of the above features.

Model Performance

Model evaluation shows that the **VQuad-HD** model has a **high performance** and is **very robust**. This can best be seen on the aggregated dataset, where the root mean square error (RMSE) is 0.56 on the MOS scale of [1 5] and the correlation coefficient is 87%.

On the individual datasets the performance reaches up to 92% of correlation, with an RMSE as low as 0.45.

Robustness is demonstrated by a **high value of lowest performance** (worst case performance): 82% of correlation with an RMSE of 0.65 on database 4.

Compared to the standard measure "PSNR", the VQuad-HD model has **always** a **higher performance** than PSNR on all individual datasets. On 4 out of 6 datasets and on the aggregated data this difference is **statistically significant**.

Additional Advantages

The **VQuad-HD** model as several additional **advantages**:

1. The model allows for an **efficient** implementation, as the model's computational complexity is kept as low as possible.
2. Furthermore, it computes **additional** quality related **features**.
3. It automatically handles interlaced and progressive video sequences.

Appendix I.4 : Tektronix HDTV FR Model

Model Description

The model includes highly accurate, verified, adaptive, configurable components for spatial alignment, simulation of display, view, perception, cognition and summary [1]. Temporal alignment capable of handling multiple frozen and skipped frames was under development at the time of the model submission deadline.

Model Performance Verification

Simulation of display, view, perception, cognition and summary was verified as per [1]. For example, the perceptual model was calibrated and verified to be within tight specification for threshold and perceived equal supra-threshold differences using over 1500 respective simulated light stimuli as per human vision science experiments. DMOS prediction accuracy was verified via accuracy analysis relative to ITU-R BT.500 compliant studies conducted by the Communications Research Centre (Ottawa) including H.264 encoded HD [2]. The model is configurable to specific viewers, display models, viewing conditions, viewing applications (sports, talking heads, general, etc.) [3].

VQEG HDTV Test Results

As can be seen in the plots of predicted DMOS vs DMOS, many of the model results have been clipped (see Figure I.1). This clipping has two causes: 1) combined skipped and frozen frames beyond capability of extended alignment algorithm implementation and 2) the ITU-R BT.500 recommended practice of a trial or "training" run is normally simulated by using example worst case video. This training was not accommodated by the VQEG HDTV Test Plan. Normally, specific applications have specific quality ranges, which set not only a scale, but also a non-linear mapping [1]. In contrast, see Figure I.2 which plots predicted DMOS vs. DMOS for codec impairments only.

The direct model outputs (predicted DMOS), before the VQEG HDTV Test Plan's "mapping to the subjective scale" under "Evaluation Procedure" plotted against DMOS demonstrates a coherent relationship apart from non-temporally aligned outliers and the aforementioned clipping (Figure I.1). However, since a 3rd order fitting function is applied (as per the VQEG HDTV Test Plan's "mapping to the subjective scale" in order to optimize over correlation and error metric) to linearize this clipped data, the coherence is reduced as seen in the final plots.

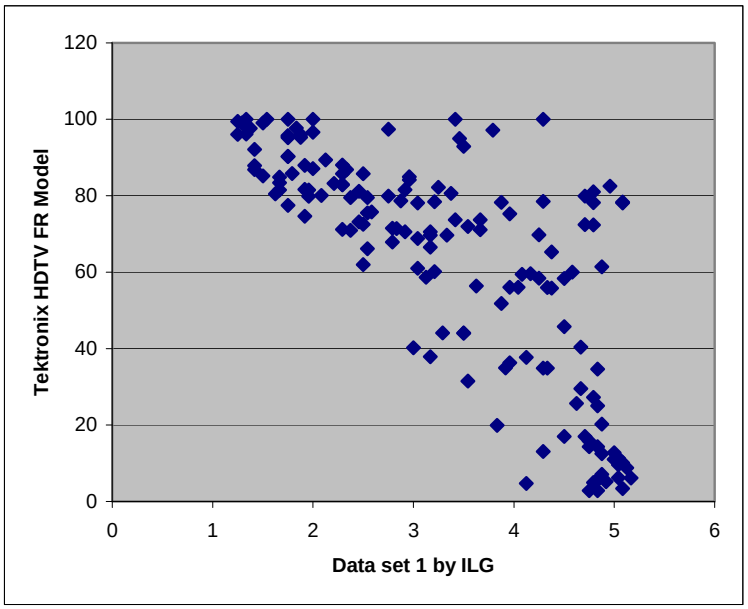
A minor accuracy reduction is also to be expected due to ILG's using different displays. Each proponent's model is evaluated against subjective scores obtained with a mix of displays, and so no one display model was sufficient to accurately simulate the light output of each used.

In summary

Although the given conditions were not fully supportive for our model, the correlation coefficient is 0.85 with Data set 6, which is assumed to have no transmission errors in the PVS without the non-linear map process described in "9.5 Mapping to the Subjective Scale". A common display and worst case training sequences would allow the model to show a higher correlation.

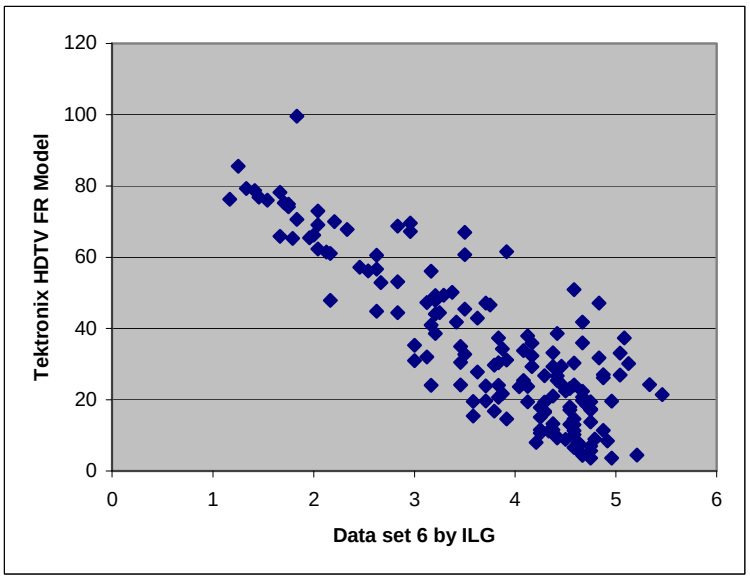
References

[1] Kevin Ferguson, "An adaptable human vision model for subjective video quality rating prediction among cif, sd, hd and e-cinema," VPQM 2007 Proceedings, http://vpqm.org/ (http://enpub.fulton.asu.edu/resp/vpqm2007/PDF_icon.gif)
[2] "Objective Measurements and Subjective Assessments,"
http://www2.tek.com/cmswpt/tidetails.lotr?ct=TI&cs=apn&ci=16509&lc=EN
[3] http://www.tek.com/products/video_test/pqa500/



Correlation coefficient : -0.76

Figure I.1: Vast majority of HRC's have transmission errors: "Transmission errors were simulated in accorance with ITU-T Rec. G.1050, "Network Model for Evaluating Multimedia Transmission Performance Over Internet Protocol." – NTIA. Note ILG uses reversed DMOS (greater magnitude is better).



Correlation coefficient : -0.85

Figure I.2: HRC's have no transmission errors, only codec impairments.

Appendix I.5 : Yonsei

Yonsei FR model

In the Yonsei FR models, an edge detection algorithm is first applied to the source video sequence to locate the edge areas. Features are extracted from these edge areas and transmitted along with other features. Then, the degradation of those edge areas is measured by computing the mean squared error. From this mean squared error, the edge PSNR is computed. Furthermore, the model uses the additional features and post-processing to adjust the EPSNR to produce the final video quality metric.

The models are efficient in terms of speed and can be implemented in real time consuming a small portion of CPU time.

Although some unexpected impairments lowered the overall performance, it can be easily taken care of, resulting in performance improvement.

Yonsei RR models

In the Yonsei RR models, an edge detection algorithm is first applied to the source video sequence to locate the edge areas. Features are extracted from these edge areas and transmitted along with other features. Then, the degradation of those edge areas is measured by computing the mean squared error. From this mean squared error, the edge PSNR is computed. Furthermore, the model uses the additional features to adjust the EPSNR to produce the final video quality metric.

The models are efficient in terms of speed and can be implemented in real time consuming a small portion of CPU time.

Although some unexpected impairments lowered the overall performance, it can be easily taken care of, resulting in performance improvement.

Appendix II: Experiment Designs

Note: The opinions expressed in this section are not endorsed by VQEG. This Appendix gives the subjective testing labs a chance to discuss their experiment's results and should not be quoted out of this context.

Appendix II.1. HRCs Associated with Each Individual Sequence in VQEGHD1

Laboratories: Ghent University – IBBT and NTIA

Test	Lab	HRC #	Codec	Bit Rate	Frame Rate	PLR	Other
HDTVPool1	NTIA	0	None	-	30	-	Reference
HDTVPool1	NTIA	1	MPEG-2	6 Mbps	30	133c	Bursty
HDTVPool1	NTIA	2	MPEG-2	6 Mbps	30	133e	Bursty
HDTVPool1	NTIA	3	MPEG-2	6 Mbps	30	133f	Bursty
HDTVPool1	NTIA	4	MPEG-2	6 Mbps	30	133g	Bursty
HDTVPool1	NTIA	5	MPEG-2	6 Mbps	30	0 %	Coding only
HDTVPool1	NTIA	6	MPEG-2	8 Mbps	30	133d	Bursty
HDTVPool1	NTIA	7	MPEG-2	8 Mbps	30	133e	Bursty
HDTVPool1	NTIA	8	MPEG-2	8 Mbps	30	133f	Bursty
HDTVPool1	NTIA	9	MPEG-2	8 Mbps	30	0 %	Coding only
HDTVPool1	NTIA	10	MPEG-2	12 Mbps	30	133c	Bursty
HDTVPool1	NTIA	11	MPEG-2	12 Mbps	30	133d	Bursty
HDTVPool1	NTIA	12	MPEG-2	12 Mbps	30	133e	Bursty
HDTVPool1	NTIA	13	MPEG-2	12 Mbps	30	133f	Bursty
HDTVPool1	NTIA	14	MPEG-2	12 Mbps	30	133g	Bursty
HDTVPool1	NTIA	15	MPEG-2	18 Mbps	30	133c	Bursty

Notes:

All HRCs were created with a hardware decoder receiving video streamed over an IP network.

Transmission errors were simulated in accorance with ITU-T Rec. G.1050, "Network Model for Evaluating Multimedia Transmission Performance Over Internet Protocol." This standard describes a statistical model in which likelihood of occurance values are assigned to all network elements and impairments. PLR of "133" indicates Test Case #133. The letter after "133" (c, d, e, f, or g) indicates the severity and type of transmission impairments imposed. The following table provides more information on severities c, d, e, f, and g. This information is taken from Table 10 in G.1050.

	Severity=>	A	B	C	D	E	F	G	H*
Source Location (A) Parameters									
LAN A Occupancy	%	1	2	3	5	8	12	16	20
Access A Occupancy	%	0	1	2	4	8	15	30	50
MTU A	bytes	512	512	1508	1508	1508	1508	1508	1508
Core Network Impairments									
Route flap interval	seconds	0	3600	1800	900	480	240	120	60
Route flap delay	ms	0	2	4	8	16	32	64	128
Delay (regional)	ms	4	8	16	32	64	128	256	512
Delay (intercontinental)	ms	16	32	64	128	196	256	512	768
Jitter (peak to peak)	ms	5	10	24	40	70	100	150	500
Link fail interval	seconds	0	3600	1800	900	480	240	120	60
Link fail duration	ms	0	64	128	256	400	800	1600	3000
Packet loss	%	0	0.01	0.02	0.04	0.1	0.2	0.5	1
Reorder factor	%	0	1	2	3	4	6	8	10
Destination Location (B) Parameters									
Access B Occupancy	%	0	1	2	4	8	15	30	50
MTU B	bytes	512	512	1508	1508	1508	1508	1508	1508
LAN B Occupancy	%	1	2	3	5	8	12	16	20

Appendix II.2. HRCs Associated with Each Individual Sequence in VQEGHD2

Laboratories : IRCCyN and Ericsson

Pool 2 test design			1080i30 test			
HRC	**Res**	**Codec**	**QP**	**MUX**	**Packet loss**	**Comment**
0	1080i	Uncomp				Reference
1	1080i	H.264	QP26		No	JM, average bitrate over all content 13.5Mb/s
2	1080i	H.264	QP32		No	JM, average bitrate over all content 5.1Mb/s
3	1080i	H.264	QP38		No	JM, average bitrate over all content 2.3Mb/s
4	1080i	H.264	QP44		No	JM, average bitrate over all content 1.2Mb/s
5	1080i	H.264	QP26		Short burst, 0.7%	JM, average bitrate over all content 13.5Mb/s
6	1080i	H.264	QP26		Long burst, 4.2%	JM, average bitrate over all content 13.5Mb/s
7	1080i	H.264	QP26		Short burst, 0.7%	JM, average bitrate over all content 13.5Mb/s
8	1080i	H.264	QP26		Short burst, 0.7%	JM, average bitrate over all content 13.5Mb/s
9	720p	H.264	QP26		No	JM, average bitrate over all content 11.3Mb/s
10	720p	H.264	QP38		No	JM, average bitrate over all content 2.5Mb/s
11	1080i	MPEG 2/H.264	QP15/38		No	FFMPEG/JM, average bitrate over all content 2.3Mb/s
12	1080i	MPEG 2	QP10		No	FFMPEG, average bitrate over all content 10.0Mb/s
13	1080i	MPEG 2	QP15		No	FFMPEG, average bitrate over all content 6.5Mb/s
14	1080i	MPEG 2	QP25		No	FFMPEG, average bitrate over all content 2.8Mb/s
15	1080i	MPEG 2	QP10		3 Bursts of biterrors	JM, average bitrate over all content 2.5Mb/s

Appendix II.3. HRCs Associated with Each Individual Sequence in VQEGHD3

Laboratories: Ghent Unversity – IBBT and Acreo

Test	Lab	HRC #	Codec	Bit Rate	Frame Rate	PLR	Other
HDTVPool3	Ghent University – IBBT	0	None	-	30	-	Reference
HDTVPool3	Ghent University – IBBT	2	MPEG-2	10 Mbps	30	0,015 %	Bursty
HDTVPool3	Ghent University – IBBT	4	H.264	15 Mbps	30	-	Coding only
HDTVPool3	Ghent University – IBBT	5	MPEG-2	5 Mbps	30	0,30 %	Bursty
HDTVPool3	Ghent University – IBBT	7	H.264	10 Mbps	30	-	Coding only
HDTVPool3	Ghent University – IBBT	8	H.264	10 Mbps	30	0,015 %	Bursty
HDTVPool3	Ghent University – IBBT	9	H.264	5 Mbps	30	0,024 %	Bursty
HDTVPool3	Ghent University – IBBT	10	H.264	5 Mbps	30	0,30 %	Bursty
HDTVPool3	Ghent University – IBBT	12	H.264	3 Mbps	30	0,035 %	Bursty
HDTVPool3	Ghent University – IBBT	13	H.264	3 Mbps	30	0,50 %	Bursty
Test	**Lab**	**HRC #**	**Codec**	**Bit Rate**	**Frame Rate**	**PLR**	**VBR Maximum Bitrate**
HDTVPool3	NTIA	16	H.264	1.0 Mbps	30	0%	3.0 Mbps
HDTVPool3	NTIA	17	H.264	1.2 Mbps	30	0%	2.5 Mbps
HDTVPool3	NTIA	18	H.264	1.5 Mbps	30	0%	2.5 Mbps
HDTVPool3	NTIA	19	H.264	2.25 Mbps	30	0%	5.5 Mbps
HDTVPool3	NTIA	20	H.264	3.4 Mbps	30	0%	7.0 Mbps
HDTVPool3	NTIA	21	H.264	5.0 Mbps	30	0%	15.0 Mbps

Note: HRCs 16-21 were created using Variable Bit Rate (VBR) encoding applied to one file that contained all 9 source sequences.

Appendix II.4. HRCs Associated with Each Individual Sequence in VQEGHD4

Laboratories : Acreo, AGH University, CRC, Ericsson, and NTIA

Pool 4 test design			**1080i25 test**			
HRC	**Res**	**Codec**	**Bitrates**	**MUX**	**Packet loss**	**Comment**
0	1080i	Uncomp				Reference
1	1080i	H.264	20Mbps	24Mbps	Low	TX1: Encoder: Tandberg EN8092, Network Simulator: Anue GEM Network Emulator,Decoder: Tandberg RX1290
2	1080i	H.264	20Mbps	24Mbps	Moderate	TX2: Encoder: Tandberg EN8092, Network Simulator: Anue GEM Network Emulator,Decoder: Tandberg RX1290
3	1080i	H.264	20Mbps	24Mbps	High	TX3: Encoder: Tandberg EN8092, Network Simulator: Anue GEM Network Emulator,Decoder: Tandberg RX1290
4	1080i	H.264	10Mbps	12Mbps	Low to High	TX4: Encoder: Tandberg EN8092, Network Simulator: Anue GEM Network Emulator,Decoder: Tandberg RX1290
5	1080i	H.264	5Mbps	6Mbps	Low	TX5: Encoder: Tandberg EN8092, Network Simulator: Anue GEM Network Emulator,Decoder: Tandberg RX1290
6	1080i	H.264	5Mbps	6Mbps	Moderate to High	TX6: Encoder: Tandberg EN8092, Network Simulator: Anue GEM Network Emulator,Decoder: Tandberg RX1290
7	1080i	H.264	2.5Mbps	6Mbps	Low	TX7: Encoder: Tandberg EN8092, Network Simulator: Anue GEM Network Emulator,Decoder: Tandberg RX1290
8	1080i	H.264	2.5Mbps	6Mbps	Moderate to High	TX8: Encoder: Tandberg EN8092, Network Simulator: Anue GEM Network Emulator,Decoder: Tandberg RX1290
9	1080i	H.264	3Mbps		No	
10	1080i	H.264	6Mbps		No	
11	720p	H.264	4Mbps		No	
12	720p	MPEG 2	5Mbps		No	
13	720p	MPEG 2	8Mbps		No	
14	720p	MPEG	11Mbps		No	

		2				
15	720p	MPEG 2	15Mbps		No	

Appendix II.5. HRCs Associated with Each Individual Sequence in VQEGHD5

Laboratories : Psytechnics and Deutsche Telekom

SRCs	SRC1-SRC9
Codecs	H264, MPEG2
Bit rates	2 Mbps - 16 Mbps
PLR	

Experimental design:	
SRCs	9
HRCs	16
PVSs	144
Common_PVSs	24
Total PVSs	168

Trials	SRC #	Filename	CODEC	Bit Rate	FPS	PLR	NOTES
1	SRC1	hdtv5_src01_hrc00.avi	N/A	N/A	25	0	Hidden reference
2	SRC2	hdtv5_src02_hrc00.avi	N/A	N/A	25	0	Hidden reference
3	SRC3	hdtv5_src03_hrc00.avi	N/A	N/A	25	0	Hidden reference
4	SRC4	hdtv5_src04_hrc00.avi	N/A	N/A	25	0	Hidden reference
5	SRC5	hdtv5_src05_hrc00.avi	N/A	N/A	25	0	Hidden reference
6	SRC6	hdtv5_src06_hrc00.avi	N/A	N/A	25	0	Hidden reference
7	SRC7	hdtv5_src07_hrc00.avi	N/A	N/A	25	0	Hidden reference
8	SRC8	hdtv5_src08_hrc00.avi	N/A	N/A	25	0	Hidden reference
9	SRC9	hdtv5_src09_hrc00.avi	N/A	N/A	25	0	Hidden reference
10	SRC1	hdtv5_src01_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)
11	SRC2	hdtv5_src02_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)
12	SRC3	hdtv5_src03_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)
13	SRC4	hdtv5_src04_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)
14	SRC5	hdtv5_src05_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)
15	SRC6	hdtv5_src06_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)
16	SRC7	hdtv5_src07_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)
17	SRC8	hdtv5_src08_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)
18	SRC9	hdtv5_src09_hrc01.avi	H264	8M	25	0	Compression errors (2-pass encoding)

19	SRC1	hdtv5_src01_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
20	SRC2	hdtv5_src02_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
21	SRC3	hdtv5_src03_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
22	SRC4	hdtv5_src04_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
23	SRC5	hdtv5_src05_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
24	SRC6	hdtv5_src06_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
25	SRC7	hdtv5_src07_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
26	SRC8	hdtv5_src08_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
27	SRC9	hdtv5_src09_hrc02.avi	H264	4M	25	0	Compression errors (2-pass encoding)
28	SRC1	hdtv5_src01_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
29	SRC2	hdtv5_src02_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
30	SRC3	hdtv5_src03_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
31	SRC4	hdtv5_src04_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
32	SRC5	hdtv5_src05_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
33	SRC6	hdtv5_src06_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
34	SRC7	hdtv5_src07_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
35	SRC8	hdtv5_src08_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
36	SRC9	hdtv5_src09_hrc03.avi	H264	2M	25	0	Compression errors (2-pass encoding)
37	SRC1	hdtv5_src01_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
38	SRC2	hdtv5_src02_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
39	SRC3	hdtv5_src03_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
40	SRC4	hdtv5_src04_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
41	SRC5	hdtv5_src05_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
42	SRC6	hdtv5_src06_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
43	SRC7	hdtv5_src07_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
44	SRC8	hdtv5_src08_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
45	SRC9	hdtv5_src09_hrc04.avi	MPEG2	8M	25	0	Compression errors (2-pass encoding)
46	SRC1	hdtv5_src01_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)
47	SRC2	hdtv5_src02_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)
48	SRC3	hdtv5_src03_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)
49	SRC4	hdtv5_src04_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)
50	SRC5	hdtv5_src05_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)
51	SRC6	hdtv5_src06_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)
52	SRC7	hdtv5_src07_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)
53	SRC8	hdtv5_src08_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)

54	SRC9	hdtv5_src09_hrc05.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding)
55	SRC1	hdtv5_src01_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
56	SRC2	hdtv5_src02_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
57	SRC3	hdtv5_src03_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
58	SRC4	hdtv5_src04_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
59	SRC5	hdtv5_src05_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
60	SRC6	hdtv5_src06_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
61	SRC7	hdtv5_src07_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
62	SRC8	hdtv5_src08_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
63	SRC9	hdtv5_src09_hrc06.avi	H264	2M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
64	SRC1	hdtv5_src01_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
65	SRC2	hdtv5_src02_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
66	SRC3	hdtv5_src03_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
67	SRC4	hdtv5_src04_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
68	SRC5	hdtv5_src05_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
69	SRC6	hdtv5_src06_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
70	SRC7	hdtv5_src07_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
71	SRC8	hdtv5_src08_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
72	SRC9	hdtv5_src09_hrc07.avi	MPEG2	4M	25	0	Compression errors (2-pass encoding using downscaled 720p version)
73	SRC1	hdtv5_src01_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
74	SRC2	hdtv5_src02_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
75	SRC3	hdtv5_src03_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
76	SRC4	hdtv5_src04_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
77	SRC5	hdtv5_src05_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
78	SRC6	hdtv5_src06_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
79	SRC7	hdtv5_src07_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
80	SRC8	hdtv5_src08_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
81	SRC9	hdtv5_src09_hrc08.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (slicing errors)
82	SRC1	hdtv5_src01_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
83	SRC2	hdtv5_src02_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
84	SRC3	hdtv5_src03_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
85	SRC4	hdtv5_src04_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
86	SRC5	hdtv5_src05_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
87	SRC6	hdtv5_src06_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
88	SRC7	hdtv5_src07_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)

89	SRC8	hdtv5_src08_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
90	SRC9	hdtv5_src09_hrc09.avi	H264	16M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
91	SRC1	hdtv5_src01_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
92	SRC2	hdtv5_src02_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
93	SRC3	hdtv5_src03_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
94	SRC4	hdtv5_src04_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
95	SRC5	hdtv5_src05_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
96	SRC6	hdtv5_src06_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
97	SRC7	hdtv5_src07_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
98	SRC8	hdtv5_src08_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
99	SRC9	hdtv5_src09_hrc10.avi	H264	16M	25	0.125	Compression (1-pass encoding) + bursty packet loss (freezing errors)
100	SRC1	hdtv5_src01_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
101	SRC2	hdtv5_src02_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
102	SRC3	hdtv5_src03_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
103	SRC4	hdtv5_src04_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
104	SRC5	hdtv5_src05_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
105	SRC6	hdtv5_src06_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
106	SRC7	hdtv5_src07_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
107	SRC8	hdtv5_src08_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
108	SRC9	hdtv5_src09_hrc11.avi	H264	16M	25	0.25	Compression (1-pass encoding) + bursty packet loss (freezing errors)
109	SRC1	hdtv5_src01_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
110	SRC2	hdtv5_src02_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
111	SRC3	hdtv5_src03_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
112	SRC4	hdtv5_src04_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
113	SRC5	hdtv5_src05_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
114	SRC6	hdtv5_src06_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
115	SRC7	hdtv5_src07_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
116	SRC8	hdtv5_src08_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
117	SRC9	hdtv5_src09_hrc12.avi	H264	16M	25	0.50	Compression (1-pass encoding) + bursty packet loss (freezing errors)
118	SRC1	hdtv5_src01_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
119	SRC2	hdtv5_src02_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
120	SRC3	hdtv5_src03_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
121	SRC4	hdtv5_src04_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
122	SRC5	hdtv5_src05_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
123	SRC6	hdtv5_src06_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)

124	SRC7	hdtv5_src07_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
125	SRC8	hdtv5_src08_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
126	SRC9	hdtv5_src09_hrc13.avi	H264	4M	25	2	Compression (1-pass encoding) + bursty packet loss (slicing errors)
127	SRC1	hdtv5_src01_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
128	SRC2	hdtv5_src02_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
129	SRC3	hdtv5_src03_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
130	SRC4	hdtv5_src04_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
131	SRC5	hdtv5_src05_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
132	SRC6	hdtv5_src06_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
133	SRC7	hdtv5_src07_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
134	SRC8	hdtv5_src08_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
135	SRC9	hdtv5_src09_hrc14.avi	H264	4M	25	1	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (slicing errors)
136	SRC1	hdtv5_src01_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)
137	SRC2	hdtv5_src02_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)
138	SRC3	hdtv5_src03_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)
139	SRC4	hdtv5_src04_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)
140	SRC5	hdtv5_src05_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)
141	SRC6	hdtv5_src06_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)
142	SRC7	hdtv5_src07_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)
143	SRC8	hdtv5_src08_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)
144	SRC9	hdtv5_src09_hrc15.avi	H264	4M	25	0.25	Compression (1-pass encoding using downsized 720p version) + bursty packet loss (freezing errors)

Appendix II.6. HRCs Associated with Each Individual Sequence in VQEGH6

Lead Experimentor : Vittorio Baroncini(FUB)

HRC Descriptions

The Pool 6 HRCs have been generated trying to provide two different sub-set of impairments.

The first set simulates cases where high coding resources are available and the best possible quality with high care to psycho visual impact is desired.

The second set simulates cases where coding time is a constrain and standard quality (with no particular attention to psycho visual impact) may be acceptable.

The first set HRCs were given odd numbers, while the second set were given even numbers.

It is quite easy to identify the first and second sub-sets making a graph of the results for each SRC and looking at the two curves got form the even or odd points.

All HRC produced using a SW AVC encoder by Ateme (Version 1.3.3.19).

(I know it, this is old! But is all I have and it doesn't work that bad!)

HRC	Profile@level	Entropy coding	Deblocking	Quality	Psychovisual	Bit rate
01	High@3.1	CABAC	enabled	Full	Best	1,5 Mbps
02	High@3.1	CABAC	enabled	Normal	None	1,5 Mbps
03	High@3.1	CABAC	enabled	Full	Best	3 Mbps
04	High@3.1	CABAC	enabled	Normal	None	3 Mbps
05	High@3.1	CABAC	enabled	Full	Best	4 Mbps
06	High@3.1	CABAC	enabled	Normal	None	4 Mbps
07	High@3.1	CABAC	enabled	Full	Best	5 Mbps
08	High@3.1	CABAC	enabled	Normal	None	5 Mbps
09	High@3.1	CABAC	enabled	Full	Best	6 Mbps
10	High@3.1	CABAC	enabled	Normal	None	6 Mbps
11	High@3.1	CABAC	enabled	Full	Best	7 Mbps
12	High@3.1	CABAC	enabled	Normal	None	8 Mbps
13	High@3.1	CABAC	enabled	Full	Best	9 Mbps
14	High@3.1	CABAC	enabled	Normal	None	11 Mbps
15	High@3.1	CABAC	enabled	Normal	None	12 Mbps

Some remarks on the ACR test method

The results of both subsets confirm my arguments about the use of an ACR test method (which-ever it could be) in the HDTV case.

Some SRCs showed evident noise due to the camera.

This was interpreted as an impairment (even if ALL viewing subjects were trained NOT to consider noise as a valuable impairment!) and produced lower value MOS points (i.e. most of the test conditions are not statistically different from each other).

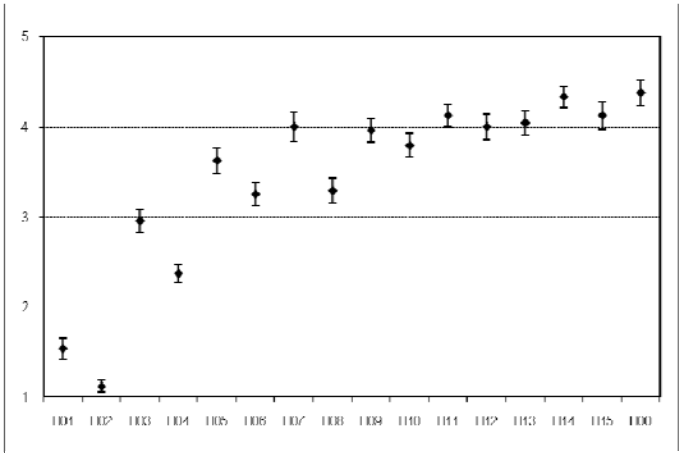
It might be more efficient to adopt DSIS (not to waste time with DSCQS) to have better results.

Some remarks on SRC04

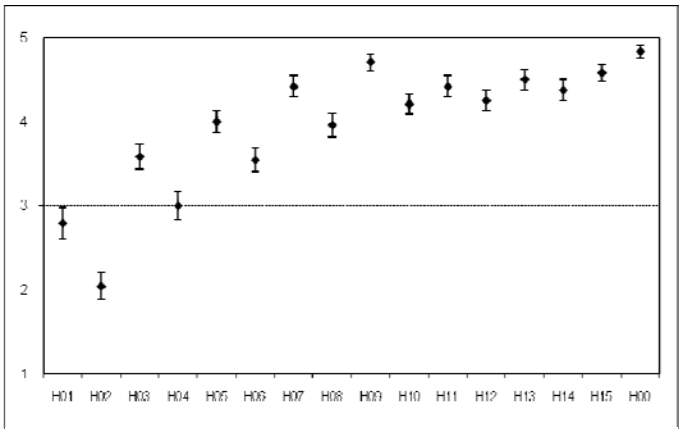
SRC04 contains noise in the original SRC. Quite all the compressed clips were evaluated better than the original SRC! I sit meaningful to compare metrics with these MOS?

MOS Range for each HRC, Plotted by SRC

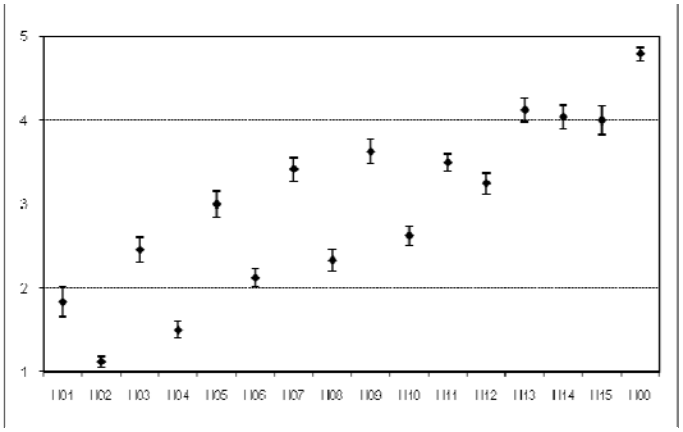
SRC01



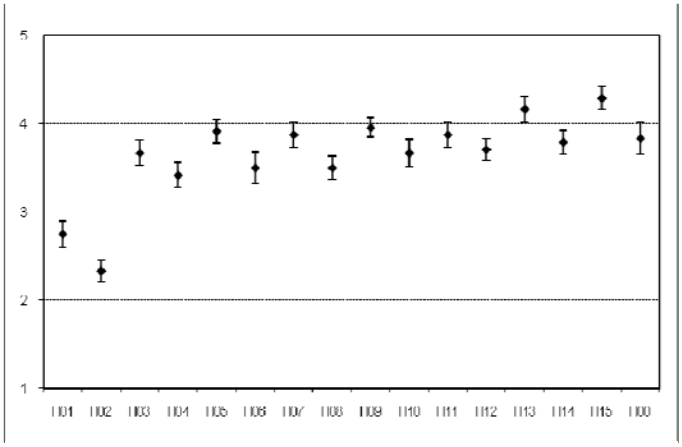
SRC02



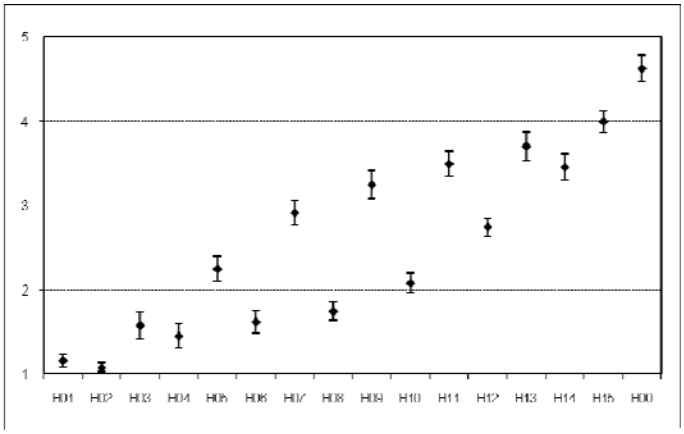
SRC03



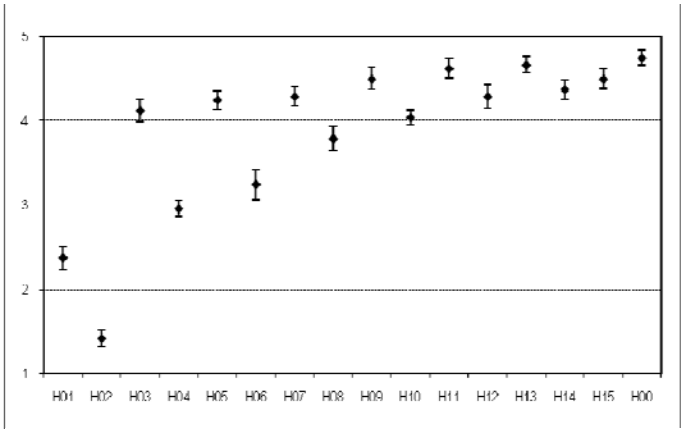
SRC04



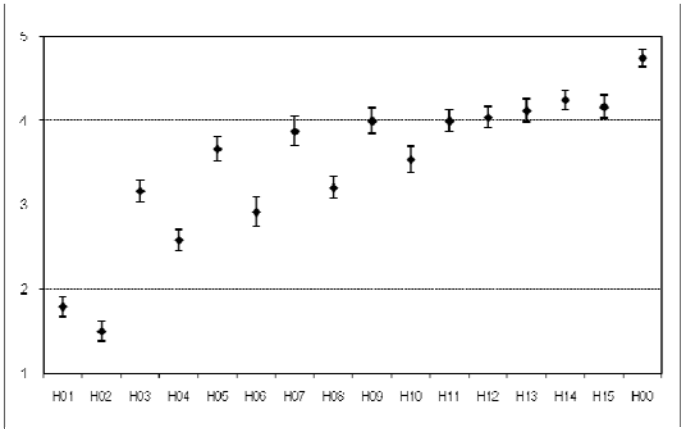
SRC05



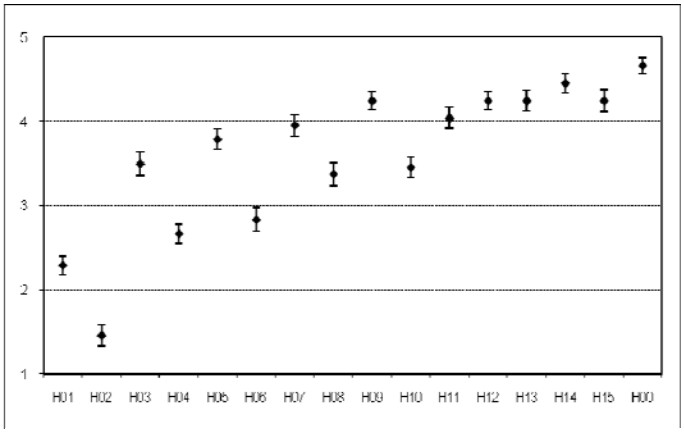
SRC06



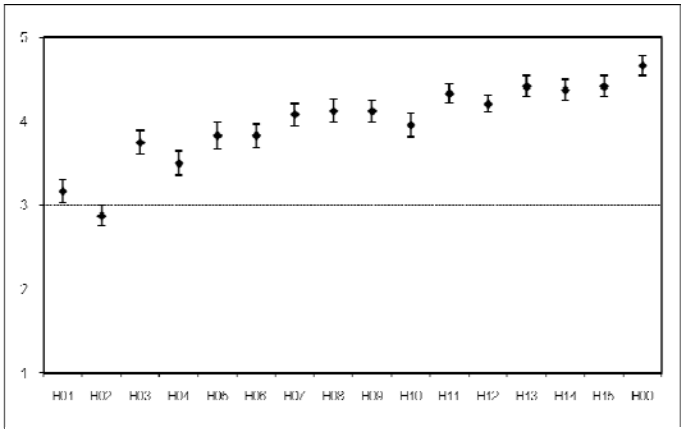
SRC07



SRC08



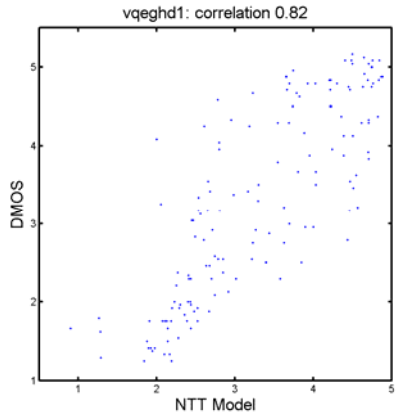
SRC09

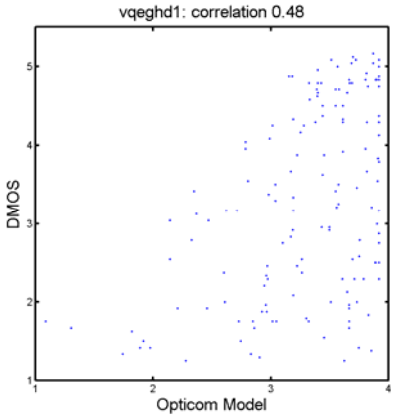


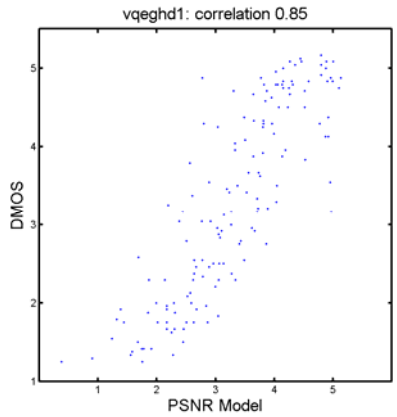
Appendix II.7. HRCs Associated with Each Individual Sequence in the Common Set

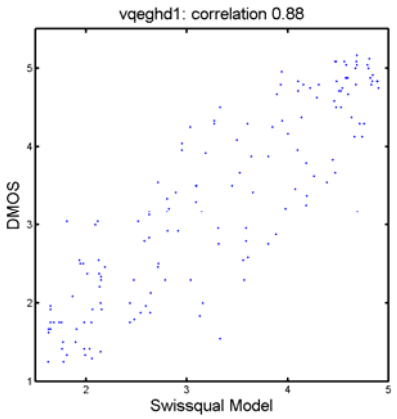
Unfortunately, the details pairing the common set PVSs and HRCs to bit rate has been lost. In compliance with the HDTV Test Plan, the common set PVSs were created using default settings of MPEG-2 and H.264 coders (i.e., no unusual coder settings). All PVSs were created within the bit rate range of 1 Mbps to 30 Mbps.

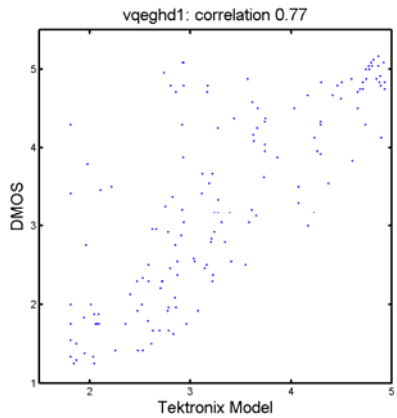
Appendix III: Plots Depicting Each Model & Dataset

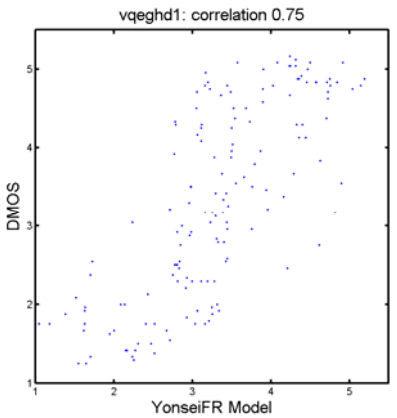


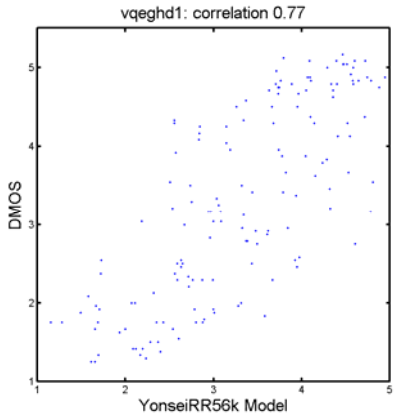


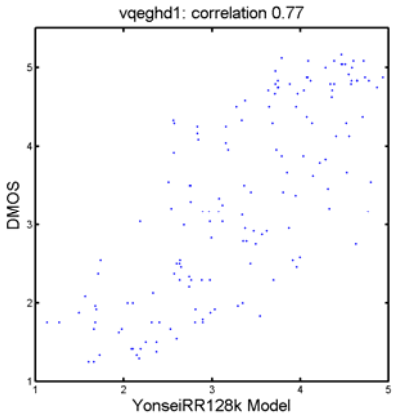


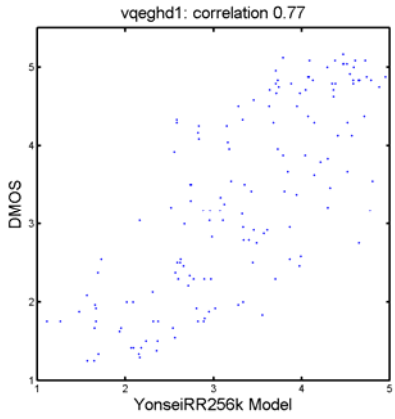


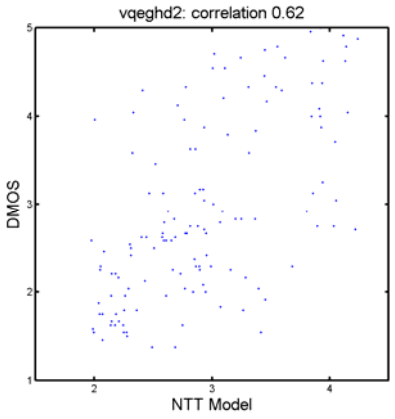


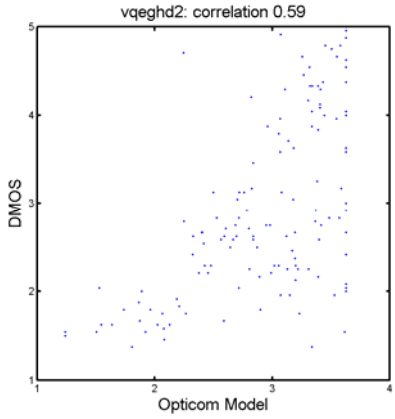


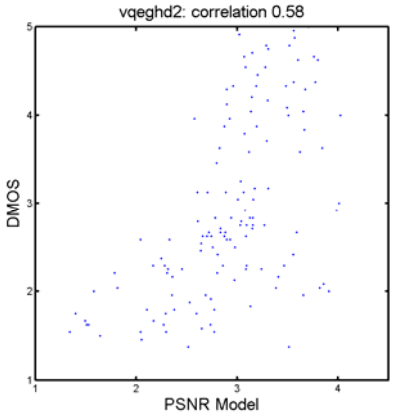

vqeghd1: correlation 0.77
DMOS
YonseiRR56k Model

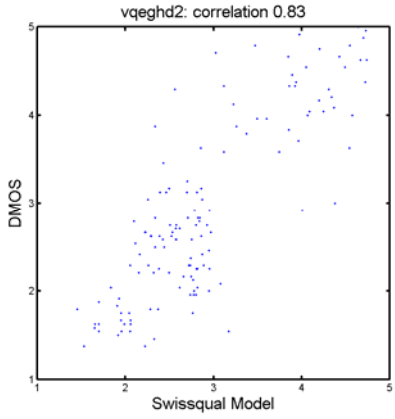

vqeghd1: correlation 0.77
DMOS
YonseiRR128k Model

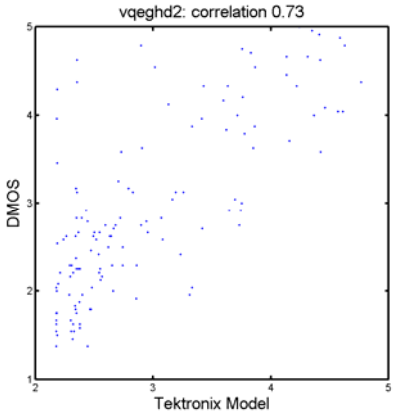

vqeghd1: correlation 0.77
DMOS
YonseiRR256k Model

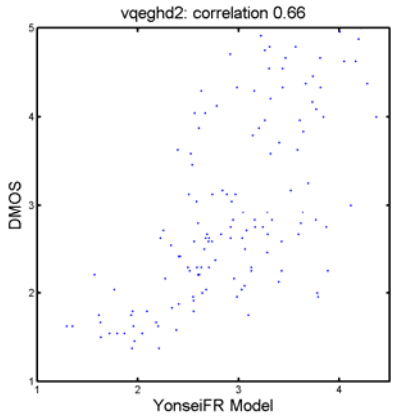

vqeghd2: correlation 0.62
DMOS
NTT Model

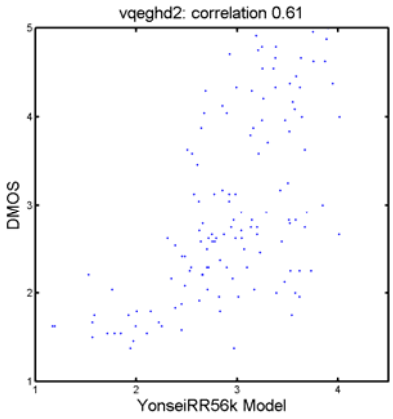

vqeghd2: correlation 0.59
DMOS
Opticom Model

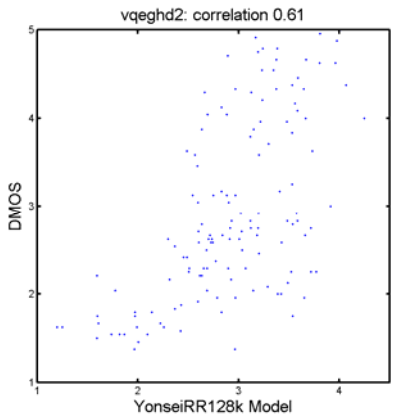

vqeghd2: correlation 0.58
DMOS
PSNR Model

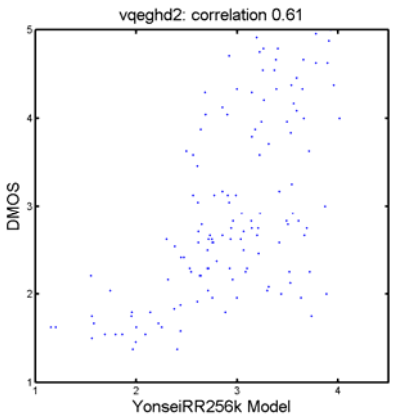

vqeghd2: correlation 0.83
DMOS
Swissqual Model

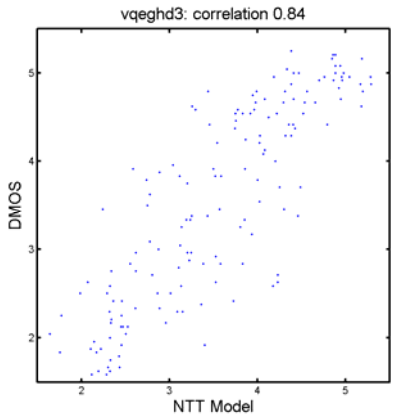

vqeghd2: correlation 0.73
DMOS
Tektronix Model

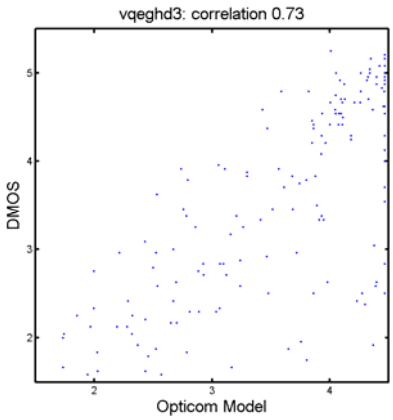

vqeghd2: correlation 0.66
DMOS
YonseiFR Model

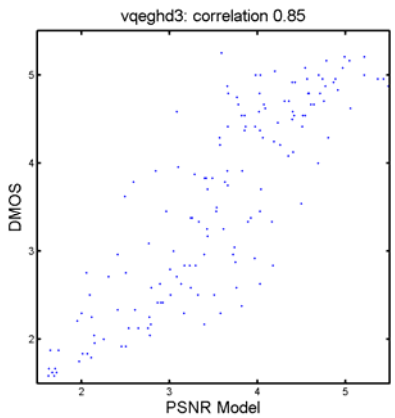

vqeghd2: correlation 0.61
DMOS
YonseiRR56k Model

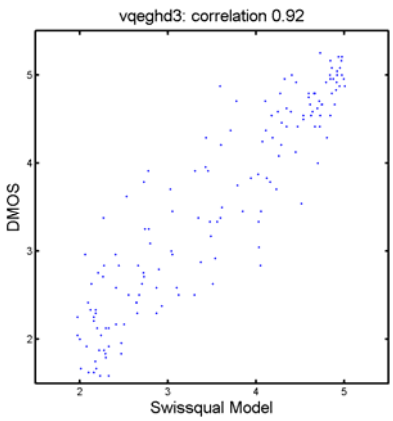

vqeghd2: correlation 0.61
DMOS
YonseiRR128k Model

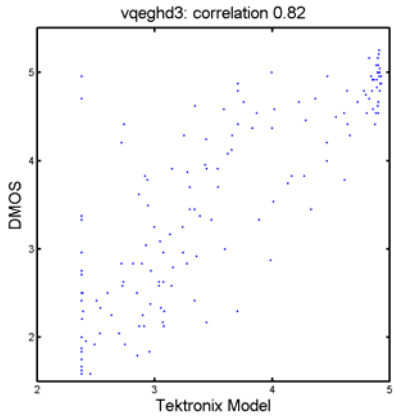

vqeghd2: correlation 0.61
DMOS
YonseiRR256k Model

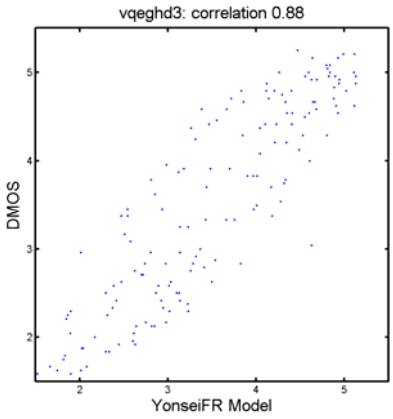

vqeghd3: correlation 0.84
DMOS
NTT Model

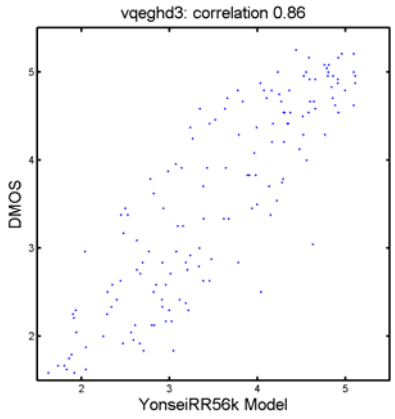

vqeghd3: correlation 0.73
DMOS
Opticom Model

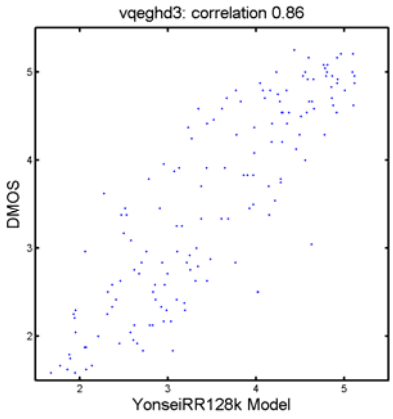

vqeghd3: correlation 0.85
DMOS
PSNR Model

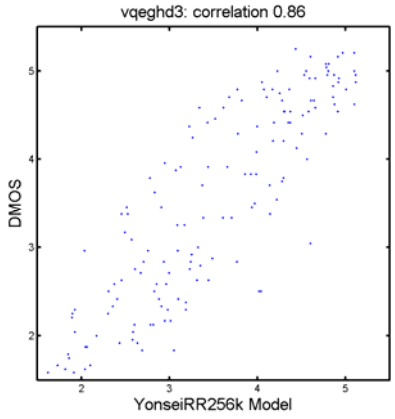

vqeghd3: correlation 0.92
DMOS
Swissqual Model

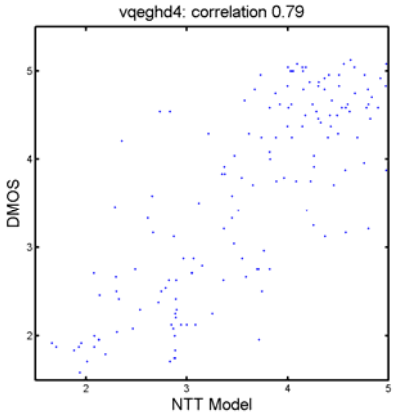

vqeghd3: correlation 0.82
DMOS
Tektronix Model

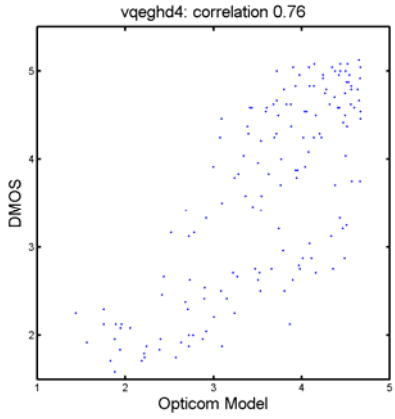

vqeghd3: correlation 0.88
DMOS
YonseiFR Model

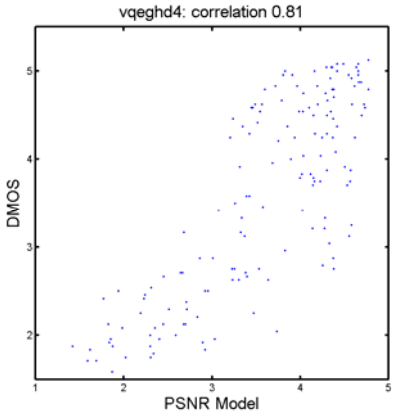

vqeghd3: correlation 0.86
DMOS
YonseiRR56k Model

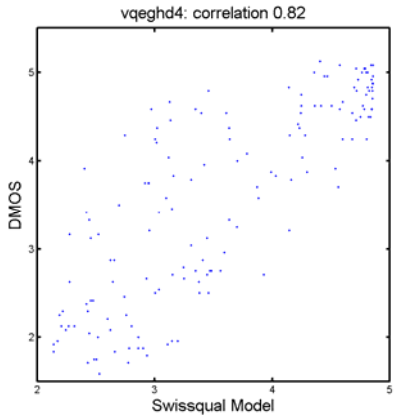

vqeghd3: correlation 0.86
DMOS
YonseiRR128k Model

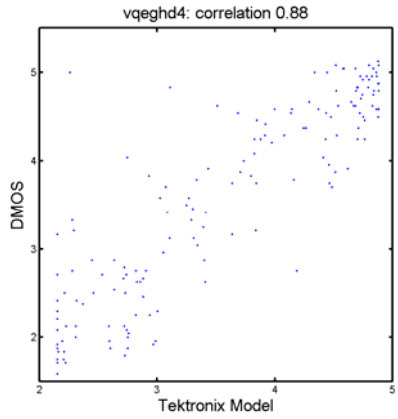

vqeghd3: correlation 0.86
DMOS
YonseiRR256k Model

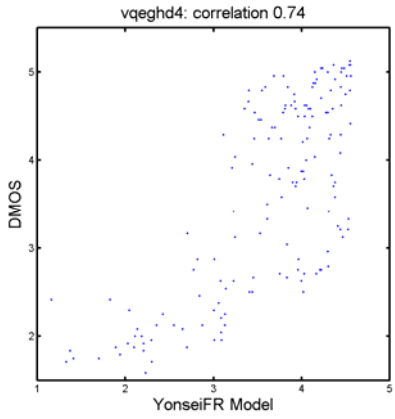

vqeghd4: correlation 0.79
DMOS
NTT Model

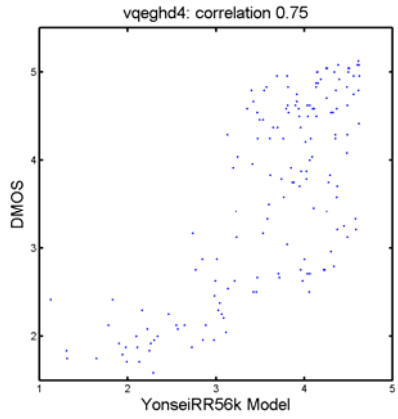

vqeghd4: correlation 0.76
DMOS
Opticom Model

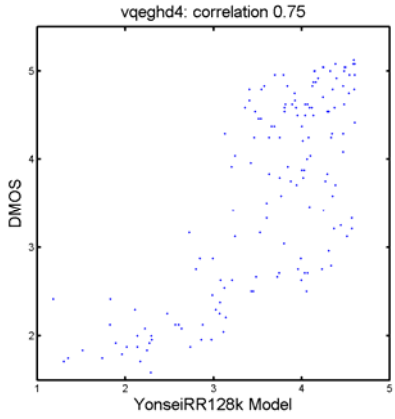

vqeghd4: correlation 0.81
DMOS
PSNR Model

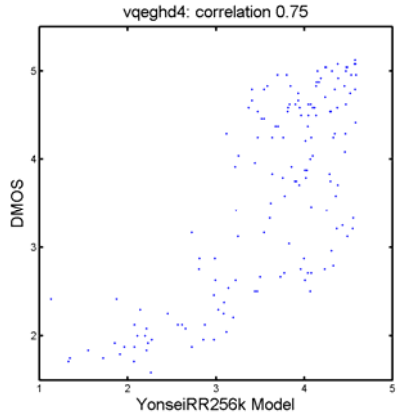

vqeghd4: correlation 0.82
DMOS
Swissqual Model

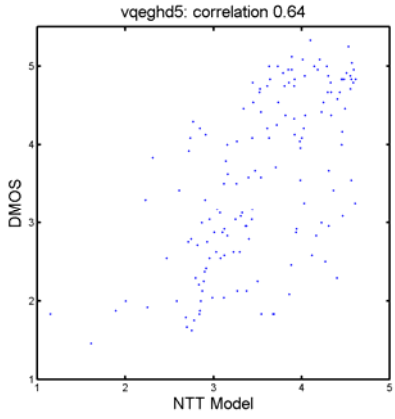

vqeghd4: correlation 0.88
DMOS
Tektronix Model

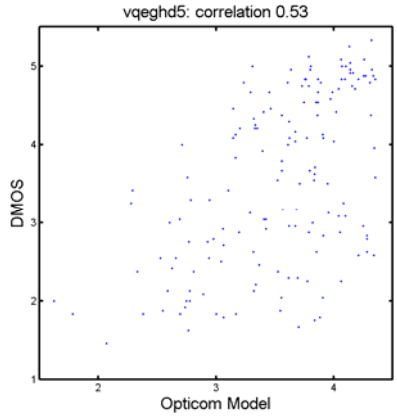

vqeghd4: correlation 0.74
DMOS
YonseiFR Model

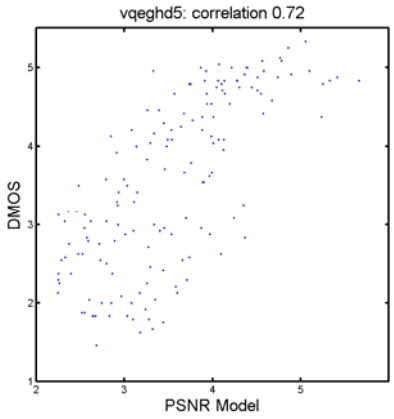

vqeghd4: correlation 0.75
DMOS
YonseiRR56k Model

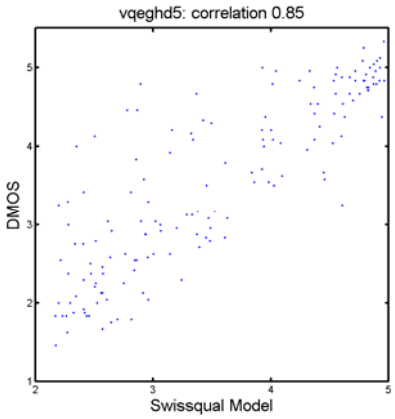

vqeghd4: correlation 0.75
DMOS
YonseiRR128k Model

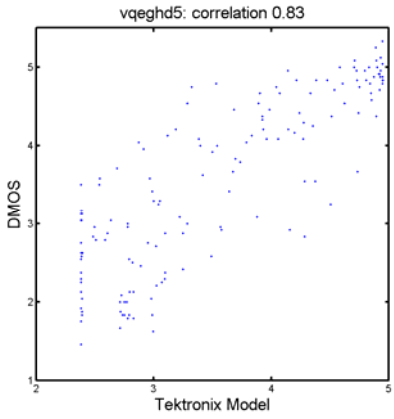

vqeghd4: correlation 0.75
DMOS
YonseiRR256k Model

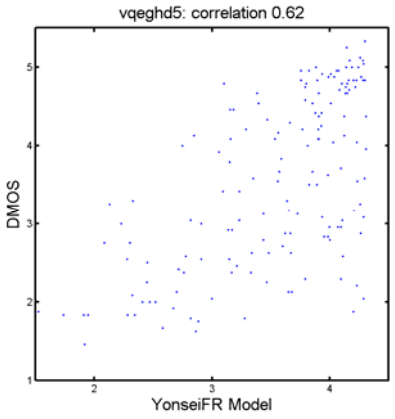

vqeghd5: correlation 0.64
DMOS
NTT Model

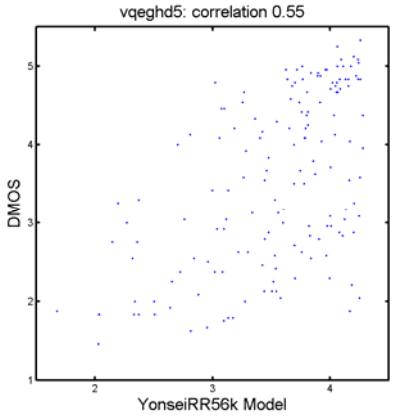

vqeghd5: correlation 0.53
DMOS
Opticom Model

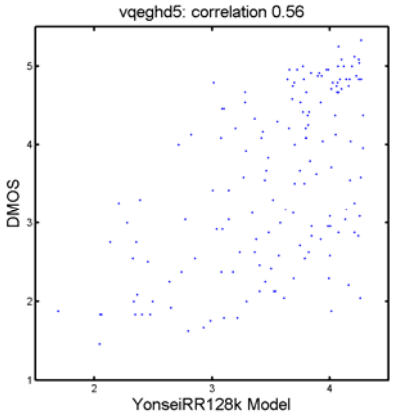

vqeghd5: correlation 0.72
DMOS
PSNR Model

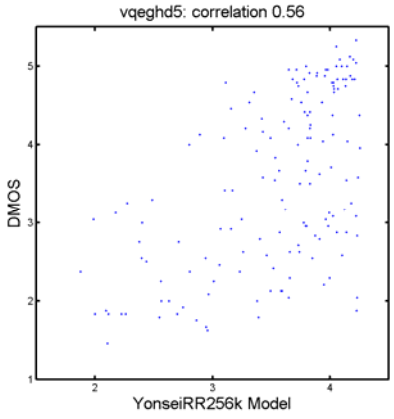

vqeghd5: correlation 0.85
DMOS
Swissqual Model

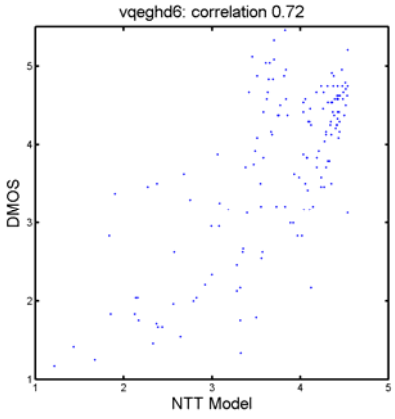

vqeghd5: correlation 0.83
DMOS
Tektronix Model

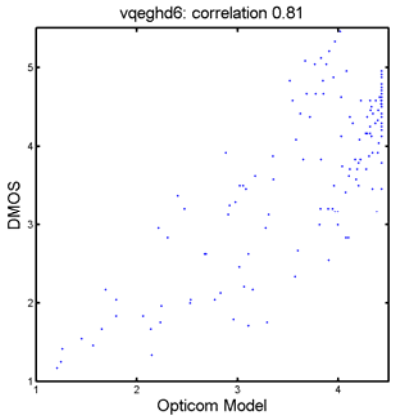

vqeghd5: correlation 0.62
DMOS
YonseiFR Model

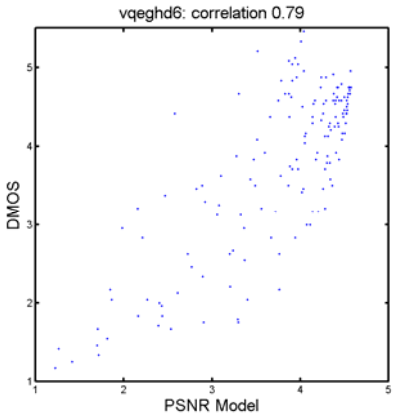

vqeghd5: correlation 0.55
DMOS
YonseiRR56k Model

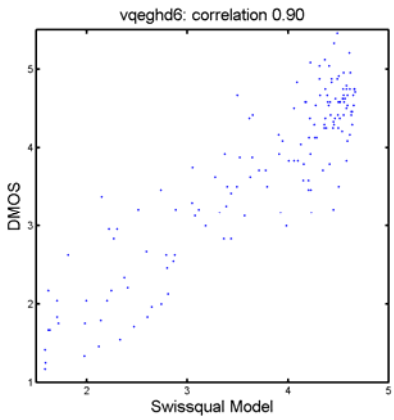

vqeghd5: correlation 0.56
DMOS
YonseiRR128k Model

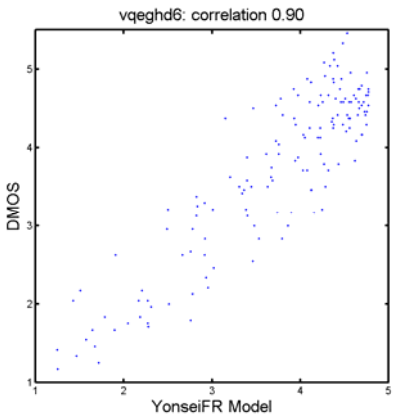

vqeghd5: correlation 0.56
DMOS
YonseiRR256k Model

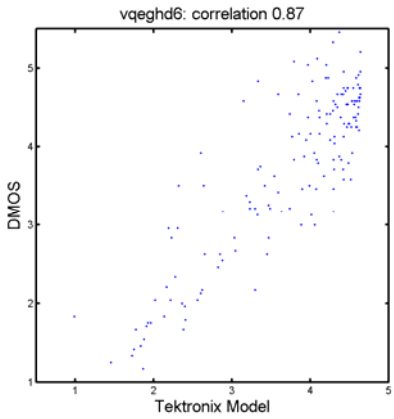

vqeghd6: correlation 0.72
DMOS
NTT Model

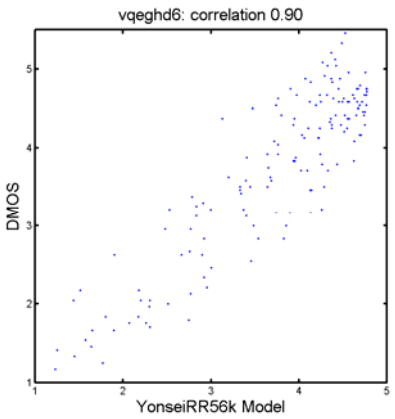

vqeghd6: correlation 0.81
DMOS
Opticom Model

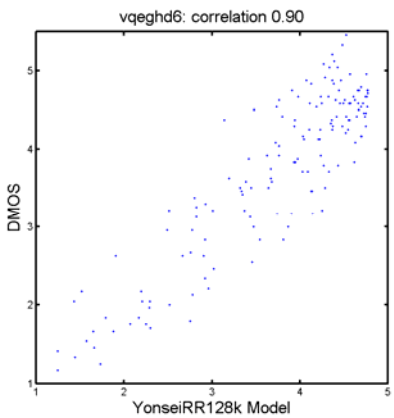

vqeghd6: correlation 0.79
DMOS
PSNR Model

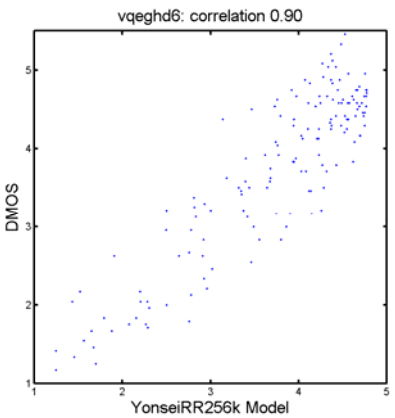

vqeghd6: correlation 0.90
DMOS
Swissqual Model


vqeghd6: correlation 0.90
DMOS
YonseiFR Model


vqeghd6: correlation 0.87
DMOS
Tektronix Model


vqeghd6: correlation 0.90
DMOS
YonseiRR56k Model


vqeghd6: correlation 0.90
DMOS
YonseiRR128k Model


vqeghd6: correlation 0.90
DMOS
YonseiRR256k Model

Appendix IV: Common Video Clip Analysis and Interpretation

Warning: This appendix presents a new method for the analysis of subjective data similarity. This secondary analysis more strictly examines the equivalence of subjective test data obtained by different laboratories than that presented in the main body of this report. Further investigation of the subjective data and analysis techniques applied in this section is necessary before any concrete conclusions can be reached.

Statistics by Lucjan Janowski

Comparing two different subjective test is a difficult task. We know that some differences can be caused by cultural or language differences, which should not be taken into consideration. On the other hand, if two experiments are run differently (different light conditions, distance, displays, ...) they should not be combined into a single set. Since in VQEG experiment a single common set should be used the experiments consistencies have to be checked.

HDTV test plan decided that the main common set analysis is based on correlation. The obtained results are shown in Section 7. This methodology was based on the previous experience, nevertheless it is not a formal prove that the results obtained by different ILGs are similar. In this appendix we present ANOVA and Pearson χ^2 test analysis are shown and discussed.

ANOVA Analysis

Note that we are interested in using MOS value which is the mean of subjects' scores. Therefore ANOVA analysis, which is created to compare the mean values obtained for different populations, seems to be perfect for this task.

The task is to compare different ILGs nevertheless we can test such a difference on a different levels by answering different questions

1. Is the global (computed for all opinion scores) mean value obtained for an ILG different from the global mean obtained for all other ILGs opinion scores?
2. Is the global (computed for all opinion scores) mean value obtained for an ILG different from the global mean obtained for another ILG opinion scores?
3. How many PVSes are different for two different ILGs?

The First Question

For each ILG the answer is negative. The closed (p-value higher than 0.025) is ILG2 what confirms the correlation analysis. The first impression is that the experiments are different and they should not be compared, nevertheless from this analysis we cannot say if a single ILG or numerous ILGs are different. This can be answered by the second question.

The Second Question

The obtained p-values are shown in Table IV.1. p-value higher than 0.05 indicates that both ILGs have the same global mean values.

Table IV.1. Lab-to-Lab p-value for MOSes

	ILG2	**ILG3**	**ILG4**	**ILG5**	**ILG6**
ILG1	0.000	0.000	**0.229**	**0.668**	0.000
ILG2		0.006	0.001	0.000	**0.506**
ILG3			0.000	0.000	0.037
ILG4				**0.439**	0.000
ILG5					0.000

The obtained results show that ILG1, ILG4 and ILG5 have statistically the same global mean values. The same conclusion can be made for ILG2 and ILG6. It means we have three groups and in order to be statistically correct we should not join those results differently than only to those separate groups.

The Third Question

Note that we are focusing on each PVS. The model and MOS are obtained for single PVS not all PVSes. Since both previous analysis aggregate the PVSes in order to have a more detailed view each PVS comparison is presented. Common set has 24 PVSes therefore for each ILG pair *n* of them can be statistically the same and 24-*n* are not statistically the same. The obtained results for each pair are shown in Table IV.2. The values under diagonal indicate how many PVSes are statistically different for the particular ILG pair. For example there is only one PVS having statistically different MOSes for ILG1 and ILG4.

Table IV.2. Number of PVSes that are statistically the same (over diagonal) and statistically different (under diagonal) for particular ILG pair for MOSes

	ILG1	**ILG2**	**ILG3**	**ILG4**	**ILG5**	**ILG6**
ILG1	-	11	6	23	19	9
ILG2	13	-	16	13	13	22
ILG3	18	8	-	10	8	15
ILG4	1	11	14	-	23	11
ILG5	5	11	16	1	-	13
ILG6	15	2	9	13	11	-

Table IV.2 analysis shows that in general PVSes for different ILGs are statistically the same (higher numbers over diagonal). Of course it is not true for all ILG pairs.

The ANOVA analysis assumes that the samples have normal distribution. Since five point scale was used the assumption about normal distribution cannot be true. Especially for extreme PVSes (very good or very bad) the answers' distribution is far from normal. In case of discrete distributions comparison a proper statistical tool is Pearson χ^2 test which is analyzed in the following part of this appendix.

Pearson χ^2 test

In this case the same questions are analyzed. Since the results are similar they are presented without too detailed description.

The First Question

Each ILG has statistically different opinion scores distribution than the rest opinion score distribution. The closed (p-value 0.016) is ILG4.

The Second Question

The obtained p-values are shown in Table IV.3. p-value higher than 0.05 indicates that both ILGs have the same opinion score distribution.

Table IV.3. Lab-to-Lab p-value for MOSes

	ILG2	**ILG3**	**ILG4**	**ILG5**	**ILG6**
ILG1	0.000	0.000	**0.433**	**0.724**	0.000
ILG2		0.000	0.004	0.004	**0.142**
ILG3			0.000	0.000	0.001
ILG4				**0.454**	0.000
ILG5					0.000

The obtained results show the same group of ILGs as ANOVA analysis.

The Third Question

In Table IV.4 number of PVSes with the same and different opinion score distribution are shown for each ILG pair.

Table IV.4. Number of PVSes which have statistically the same distribution (over diagonal) and statistically different distribution (under diagonal) for particular ILG pair for opinion scores

	ILG1	**ILG2**	**ILG3**	**ILG4**	**ILG5**	**ILG6**
ILG1	-	12	10	22	21	10
ILG2	12	-	14	14	16	21
ILG3	14	10	-	15	10	16
ILG4	2	10	9	-	22	10
ILG5	3	8	14	2	-	15
ILG6	14	3	8	14	9	-

Table IV.4 analysis confirms the results obtained by ANOVA analysis. Since Pearson χ^2 test is more restrictive (if distributions are the same the mean values also are the same but the opposite conclusion is not true) in this case less PVSes are statistically the same.

Conclusions

The first conclusion is that Pearson χ^2 test and ANOVA analysis results are similar. The most significant difference between those analysis is more restrictive PVSes similarity in case of Pearson χ^2 test. Such stronger restriction is an obvious consequence of Pearson χ^2 test methodology.

The formal analysis shows that the differences between different ILGs are statistically significant. Nevertheless, comparing PVSes we see that most of them are statistically the same and we have to remember that PVSes are the most important from the analysis point of view. Therefore, we prefer to say that using formal test in order to decide if two ILGs are similar is incorrect. Therefore, we decided to use correlation criteria as the final one.

Appendix V Method for Post-Experiment Screening of Subjects

A statistical criterion for rejecting a subject's data is that it correlates with the average of the other subjects' data no better than chance. The linear Pearson correlation coefficient per PVS for one viewer vs. all viewers is defined as:

$$r1(x,y) = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right)\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right)}}$$

Where
xi = MOS of all viewers per PVS
yi = individual score of one viewer for the corresponding PVS
n = number of PVSs
i = PVS index.

Rejection criterion

1. Calculate r1 for each viewer
2. Exclude a viewer if (r1<0.75) for that subject

Appendix VI Expansion of Scope to Include CRT Monitors

Experimentors: Mr. Akira Takahashi, Mr. Taichi Kawano, and Mr. Jun Okamoto (NTT)

1. Introduction

The HDTV Test Plan indicated that an LCD monitor should be used in subjective tests. Concerns were raised that the video quality of a CRT monitor might differ from that of an LCD monitor. However, such a subjective test result had not been obtained. Therefore, NTT conducted subjective tests to compare video-quality characteristics between CRT and LCD monitors. These results show that the HDTV subjective video quality was not affected by monitor type.

2. Experiment

To verify that subjective video quality is not affected by monitor type, we conducted four experiments (See Table VI.1). We used two data sets (vqeghd2 and vqeghd4). The subjective video-quality characteristics between LCD and CRT monitors (See Table VI.2) are shown in Fig. VI.1. Their correlation coefficients (Rs) and the root mean square errors (RMSEs) are listed in Table VI.3. These results of the four experiments show that R is high and RMSE is low, regardless of frame rate. Additionally, the results of a t-test show that there is no significant difference between video qualities for these monitors, where a significant level is 5%.

3. Conclusions

These results show that the HDTV subjective video quality is not affected by monitor type (LCD and CRT). The scope of Recommendations resulting from the VQEG HDTV Phase I experiment can be expanded to include head-end monitoring, because the subjective video-quality characteristics for an LCD monitor are statistically equivalent to that for a CRT monitor. Thus, the results of this HDTV Final Report can be used for monitoring applications where the CRT monitor is required.

4. Reference

[1] VQEG, “Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content,” 2009.

5. Acknowledgment

This activity was supported by the Strategic Information and Communications R&D Promotion Programme (SCOPE) of the Ministry of Internal Affairs and Communication of Japan (No.073103002).

Table VI.1 Experimental conditions

		Monitor type	
		LCD	CRT
Frame rate	60i (vqeghd2)	Experiment 1	Experiment 2
	50i (vqeghd4)	Experiment 3	Experiment 4

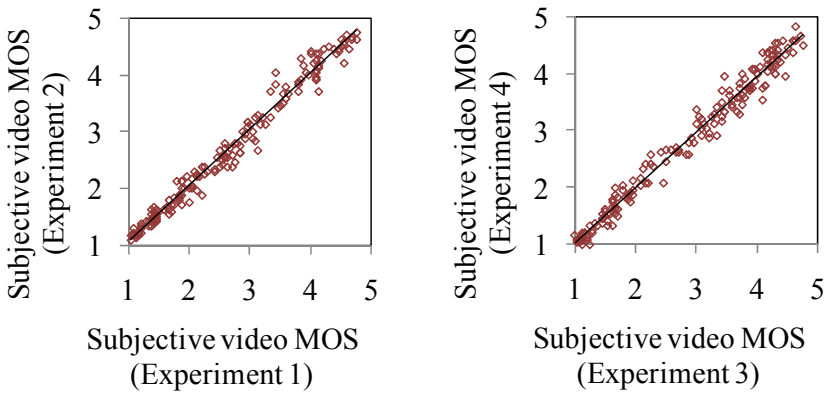


Figure VI.1 Subjective video-quality characteristics between LCD and CRT monitors

Table VI.2 Monitor information

	Monitor type	
	LCD	CRT
Manufacture name	Sony	Sony
Model number	LMD-4250W	BVM-D32E1WJ

Table VI.3 Statistical results

	R	RMSE
Experiments 1 and 2	0.990	0.176
Experiments 3 and 4	0.989	0.175