Video Quality Experts Group:
Current Results and Future Directions

Ann Marie Rohaly,[a] Philip Corriveau,[b,*] John Libert,[c] Arthur Webster,[d] Vittorio Baroncini,[e] John Beerends,[f] Jean-Louis Blin,[g] Laura Contin,[h] Takahiro Hamada,[i] David Harrison,[j] Andries Hekstra,[f] Jeffrey Lubin,[k] Yukihiro Nishida,[l] Ricardo Nishihara,[m] John Pearson,[k] Antonio Franca Pessoa,[m] Neil Pickford,[n] Alexander Schertz,[o] Massimo Visca,[p] Andrew Watson[q] and Stefan Winkler[r]

[a]Tektronix Inc., Beaverton, Oregon; [b]Communications Research Centre, Ottawa, Canada; [c]National Institute of Standards and Technology, Gaithersburg, Maryland; [d]U.S. Dept. of Commerce, Boulder, Colorado; [e]Fondazione Ugo Bordoni, Rome, Italy; [f]KPN Research, Leidschendam, Netherlands; [g]CCETT/CNET, Sévigné Cedex, France; [h]Centro Studi e Laboratori Telecomunicazioni, Turin, Italy; [i]KDD Media Will Corporation, Tokyo, Japan; [j]Independent Television Commission, United Kingdom; [k]Sarnoff Corporation, Princeton, New Jersey; [l]NHK, Tokyo, Japan; [m]CPqD Foundation, Campinas, Brazil; [n]DCITA, Canberra, Australia; [o]Institut für Rufndunktechnik GmbH, Munich, Germany; [p]RAI, Turin, Italy; [q]NASA Ames Research Center, Moffett Field, California; [r]Swiss Federal Institute of Technology, Lausanne, Switzerland

ABSTRACT

Subjective assessment methods have been used reliably for many years to evaluate video quality. They continue to provide the most reliable assessments compared to objective methods. Some issues that arise with subjective assessment include the cost of conducting the evaluations and the fact that these methods cannot easily be used to monitor video quality in real time. Furthermore, traditional, analog objective methods, while still necessary, are not sufficient to measure the quality of digitally compressed video systems. Thus, there is a need to develop new objective methods utilizing the characteristics of the human visual system. While several new objective methods have been developed, there is to date no internationally standardized method.

The Video Quality Experts Group (VQEG) was formed in October 1997 to address video quality issues. The group is composed of experts from various backgrounds and affiliations, including participants from several internationally recognized organizations working in the field of video quality assessment. The majority of participants are active in the International Telecommunications Union (ITU) and VQEG combines the expertise and resources found in several ITU Study Groups to work towards a common goal. The first task undertaken by VQEG was to provide a validation of objective video quality measurement methods leading to Recommendations in both the Telecommunications (ITU-T) and Radiocommunication (ITU-R) sectors of the ITU. To this end, VQEG designed and executed a test program to compare subjective video quality evaluations to the predictions of a number of proposed objective measurement methods for video quality in the bit rate range of 768 kb/s to 50 Mb/s. The results of this test show that there is no objective measurement system that is currently able to replace subjective testing. Depending on the metric used for evaluation, the performance of eight or nine models was found to be statistically equivalent, leading to the conclusion that no single model outperforms the others in all cases. The greatest achievement of this first validation effort is the unique data set assembled to help future development of objective models.

Keywords: Objective measurement, Subjective assessment, Video Quality Experts Group, International Telecommunications Union, Double Stimulus Continuous Quality Scale, Standards, Video Quality Assessment, Quality Metrics, Evaluation

[*] Email: phil.corriveau@crc.ca Telephone: +1 613-998-7822 Fax: +1 613-990-6844 Website: http://www.crc.ca/vqeg/
VQEG Reflector: ituvidq@its.bldrdoc.gov

1. INTRODUCTION

A growing concern of the video industry, including broadcasters, telecommunications organizations, and video researchers, is the assurance and maintenance of an acceptable level of service quality for the distribution of video programming. This paper discusses current objective and subjective methods of video quality assessment along with the results of a validation test conducted by the Video Quality Experts Group (VQEG).

The long-standing benchmark used for evaluating video quality is subjective assessment. For more than twenty years, researchers world-wide have used the subjective methods standardized in ITU-R Recommendation BT.500 [1] to evaluate video quality in television services. More recently, the ITU-T developed Recommendation P.910 [2] to standardize methods for multimedia quality assessment. Because the premise behind subjective assessment is the use of human observers to rate the quality of video sequences (usually short clips), it is impractical and impossible to use these methods for the in-service continuous evaluation of video quality. Hence, objective methods are required.

Traditional, analog objective measurement systems, while still necessary, are no longer sufficient to measure the quality of digitally compressed video systems. With the shift in technology from analog to digital, a change in the types of visual artifacts has occurred. As a result, the outputs of analog measurement methods do not correlate well with subjective quality assessments of digitally compressed video. In order to properly assess these new artifacts, new objective methods need to be developed and standardized.

As a result, VQEG was formed to address these and related video quality issues. (See reference [3] for more background on VQEG.) VQEG conducted a large validation test that compared subjective results to the objective measurements of ten proponent models. With the completion of this validation test of objective models, a new, large, and invaluable data set will be made available for future model development. It was hoped that the results of this test could be used in the preparation of one or more new ITU Recommendations. The results, however, indicate that it is too soon for an ITU Recommendation to be determined.

2. CURRENT METHODS

Many accepted methods are used to evaluate and examine video quality. There are two classes of assessment methods: subjective and objective. Subjective test methods require human viewers, expert or non-expert, to rate the quality or difference in quality of two video clips. Subjective assessment can be a costly and time-consuming process, but one, however, that yields accurate results for any given evaluation.

Objective test methods do not use human subjects but rather measure and analyze the video signal. Traditional analog methods are able to accurately measure and assess analog impairments of the video signal. However, with the introduction and development of digital technologies, visually noticeable artifacts are manifested differently than analog artifacts. This change has led to the need for new objective test methods.

The new objective measurement methods analyze the video signal, some utilizing knowledge of the human visual system. These methods implement an algorithm that measures video quality based (usually) on the comparison of a source and a processed sequence. The algorithms, referred to as models, may incorporate characteristics of the human visual system in an attempt to systematically measure the perceptible degradation occurring in the video imagery. See reference [3] for more information on subjective and objective testing.

3. STANDARDIZATION

Several quality assessment methods covering different areas of service have been recommended and standardized by the International Telecommunications Union (ITU). For example, ITU-T Recommendations P.910 [2] and P.920 [4] cover subjective quality assessment methods for multimedia applications, ITU-T Recommendations P.800 [5] and P.861 [6] cover methods for subjective and objective assessment of telephone speech quality, and ITU-R Recommendation BT.500 [1] covers methods for the subjective assessment of the quality of television pictures.

With the shift in video technology from analog to digital, there is an urgent need to establish international standards for the objective evaluation of video quality. As with subjective assessment standards that have been developed over the years, objective methods must follow a rigorous standardization process to ensure that the methods are accurate and robust. The current VQEG validation test is the first step in the international standardization process for objective video quality methods.

3.1 Video Quality Experts Group (VQEG)

In October 1997, a meeting was held at Centro Studi e Laboratori Telecomunicazioni (CSELT) in Turin, Italy to discuss the technical procedures needed to validate objective measures of video quality. Experts from ITU-T Study Groups 9 and 12 and ITU-R Study Group 11 took part in the meeting and contributed to the specification of a work plan for this activity. It was established that this group would embark on the task of outlining, designing, and conducting a test to evaluate objective measurement systems. This was the formative meeting for VQEG, which has since conducted most of its work via an e-mail reflector (ituvidq@its.bldrdoc.gov). There are currently over 100 people on this e-mail list. Arthur Webster, U.S. Dept. of Commerce, and Philip Corriveau, Communications Research Centre Canada, are the Co-Chairs of VQEG.

VQEG formally announced and proposed the validation of objective measurement systems and began the process of soliciting submissions of models to be included in an ITU verification process leading to one or more ITU Recommendations. The proposal was submitted to ITU-T Study Groups 9 and 12 and ITU-R Study Group 11 as a contribution and was also sent out on the VQEG e-mail reflector. The proposal stated that all objective models should be capable of receiving as input a processed sequence and its corresponding source sequence. Based on this input, the model must provide one unique figure of merit that correlates with the value obtained from subjective assessment of the processed material.

3.2 Testing Process

An extensive testing process was finalized at the second meeting of VQEG held in May 1998 at the National Institute of Standards and Technology (NIST) facility in Gaithersburg, Maryland. A set of test sequences was selected by the Independent Lab and Selection Committee (ILSC) of VQEG. The sequences were kept confidential until the proponents submitted their final implementations of objective assessment methods to the ILSC (August 7, 1998). The final selection of test conditions (referred to as hypothetical reference circuits or HRCs) was approved by the VQEG members present at the meeting. Details concerning the processing of test sequences and the final distribution of these to both proponents and testing facilities were finalized. The two testing phases (subjective and objective) were executed in parallel and a detailed timetable can be found in the report for the Gaithersburg VQEG meeting [7].

The subjective test was conducted by multiple laboratories who received the source and processed video clips in random order on digital component video viewing tapes. Each facility conducted formal subjective assessments according to the subjective test plan developed and published by VQEG [8]. The subjective test plan followed the currently accepted testing procedures standardized in ITU-R Recommendation BT.500. The Double Stimulus Continuous Quality Scale (DSCQS) method was used for the subjective test. Due to the size of the test (20 test sequences × 16 HRCs) and the need the minimize contextual effects, the test design was broken down into four basic quadrants (50Hz/high quality, 50Hz/low quality, 60Hz/high quality and 60Hz/low quality). In addition, the need to repeat experiments in different laboratories necessitated the use of a large number of subjective testing facilities. The following eight laboratories conducted subjective tests:

CRC (Canada)
RAI (Italy)
CCETT (France)
CSELT (Italy)
NHK (Japan)
DCITA (Australia)
Berkom (Germany)
FUB (Italy)

In the case of objective testing, each proponent of an objective method received the video sequences on computer tapes, analyzed them, and reported the results—one number per HRC/source combination. (For the details of the objective test procedure, see the objective test plan [9].) A randomly selected subset of these sequences (10%) was sent to several independent testing laboratories to confirm the results reported by the proponents. The following four laboratories performed this verification:

CRC (Canada)
IRT (Germany)
FUB (Italy)
NIST (USA)

The proponents of objective methods of video quality participating in this test are listed below (the numbers in brackets were assigned to the proponents for analysis purposes and are referred to later in the Results and Discussion section):

[P0] Peak Signal to Noise Ratio (PSNR)
[P1] CPqD (Brazil)
[P2] Tektronix / Sarnoff (USA)
[P3] NHK (Japan Broadcasting Corporation)/ Mitsubishi Electric Corp. (Japan)
[P4] KDD (Japan)
[P5] Swiss Federal Institute of Technology (EPFL) (Switzerland)
[P6] TAPESTRIES (European Union)
[P7] NASA (USA)
[P8] KPN Research (The Netherlands) / Swisscom CIT (Switzerland)
[P9] NTIA/ITS (USA)
---- Institut für Nachrichtentechnik (Germany). (Results not included [10].)

4. RESULTS AND DISCUSSION

For complete details of the results, statistical analysis and conclusions please refer to the objective test plan [9] and the VQEG Final Report [10]. Included in this section are selected excerpts from this full report.

4.1 Subjective Data

Prior to conducting the full analysis of the data, a post-screening of the subjective test scores was conducted. The first step of this screening was to check the completeness of the data for each viewer. A viewer was discarded if there was more than one missed vote in a single test session. The second step of the screening was to eliminate viewers with unstable scores and viewers with extreme scores (i.e., outliers). The procedure used in this step was that specified in Annex 2, section 2.3.1 of ITU-R Recommendation BT.500 [1] and was applied separately to each test quadrant for each laboratory (i.e., 50 Hz/low quality, 50 Hz/high quality, 60 Hz/low quality, 60 Hz/high quality for each laboratory).

A total of 10 viewers was discarded from the subjective data set, leading to a screened subjective data set including scores from a total of 287 viewers: 140 from the 50 Hz tests and 147 from the 60 Hz tests. The number of viewers by test quadrant is as follows: 50 Hz/low quality – 70 viewers, 50 Hz/high quality – 70 viewers, 60 Hz/low quality – 80 viewers and 60 Hz/high quality – 67 viewers.

To examine the results of the subjective test, an analysis of variance (ANOVA) was conducted on the collected ratings. The purpose of conducting an ANOVA on the subjective data was multi-fold. First, it allowed for the identification of main effects of the test variables and interactions between them that might suggest underlying problems in the data set. Second, it allowed for the identification of differences among the data sets obtained by the eight subjective testing laboratories. Finally, it allowed for the determination of context effects due to the different ranges of quality inherent in the low and high quality portions of the test.

Because the various HRC/source combinations in each of the four quadrants were presented in separate tests with different sets of viewers, individual ANOVAs were performed on the subjective data for each test quadrant. Each of these analyses was a 4 (lab) × 10 (source) × 9 (HRC) repeated measures ANOVA with lab as a between-subjects factor and source and HRC as within-subjects factors. The basic results of the analyses for all four test quadrants are in agreement and demonstrate highly significant main effects of HRC and source sequence and a highly significant HRC × source sequence interaction ($p < 0.0001$ for all effects). As these effects are expected outcomes of the test design, they confirm the basic validity of the design and the resulting data.

For the two low quality test quadrants, 50 and 60 Hz, there is also a significant main effect of lab ($p < 0.0005$ for 50 Hz, $p < 0.007$ for 60 Hz). This effect is due to differences in the difference mean opinion scores (DMOS) measured by each lab. Despite the fact that viewers in each laboratory rated the quality differently on average, the aim here was to use the entire subject sample to estimate global quality measures for the various test conditions and to correlate the objective model outputs to these global subjective scores. Moreover, individual lab to lab correlations are generally high and this is due to the fact that even though the mean scores are statistically different, the scores for each lab vary in a similar manner across test conditions.

Additional analyses were performed on the data obtained for two HRCs that were common to both low and high quality tests. These analyses were 2 (quality) × 10 (source) × 2 (HRC) repeated measures ANOVAs with quality as a between-subjects factor and source and HRC as within-subjects factors. The basic results of the 50 and 60 Hz analyses are in agreement and show no significant main effect of quality range and no significant HRC × quality range interaction ($p > 0.2$ for all effects). Thus, these analyses indicate no context effect was introduced into the data for these two HRCs due to the different ranges of quality inherent in the low and high quality portions of the test.

4.2 Objective Data

Performance of the objective models was evaluated with respect to three aspects of their ability to estimate subjective assessment of video quality:

- prediction accuracy – the ability to predict the subjective quality ratings with low error,
- prediction monotonicity – the degree to which the model's predictions agree with the relative magnitudes of subjective quality ratings and
- prediction consistency – the degree to which the model maintains prediction accuracy over the range of video test sequences, i.e., that its response is robust with respect to a variety of video impairments.

These attributes were evaluated through four performance metrics specified in the objective test plan [9] and are discussed in the following sections. As a visual illustration of the relationship between subjective data and model predictions, scatter plots of DMOS and model predictions are provided below for each model. Figure 1 shows that for many of the models, the points cluster about a common trend, though there may be various outliers.

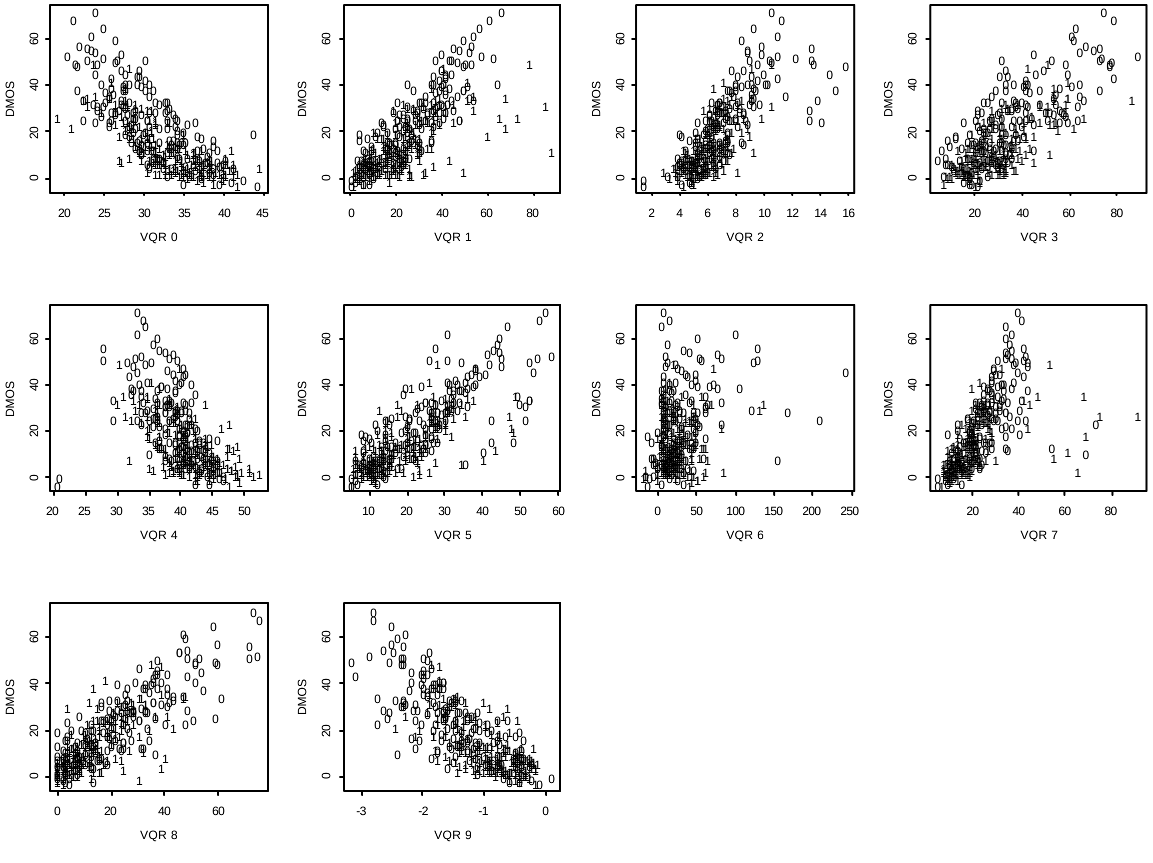

Figure 1. Scatter plots of DMOS vs. model predictions for the complete data set. The 0 symbols indicate scores obtained in the low quality quadrants of the subjective test and the 1 symbols indicate scores obtained in the high quality quadrants of the subjective test.

4.2.1 Variance-weighted regression analysis (modified metric 1 [9])

In developing the VQEG objective test plan, it was observed that regression of DMOS against objective model scores might not adequately represent the relative degree of agreement of subjective scores across the video sequences. Hence, a metric was included in order to factor this variability into the correlation of objective and subjective ratings. On closer examination of this metric, however, it was determined that regression of the subjective differential opinion scores, DOS, with the objective scores would not necessarily accomplish the desired effect, i.e., accounting for unequal variance of the subjective ratings in the correlation to objective scores. Moreover, conventional statistical practice offers a method for dealing with this situation.

Regression analysis assumes homogeneity of variance among the replicates, Y_{ik} , regressed with X_i . When this assumption cannot be met, a weighted least squares analysis can be used. A function of the variance among the replicates can be used to explicitly factor a dispersion measure into the computation of the regression function and the correlation coefficient. Accordingly, rather than applying metric 1 as specified in the objective test plan [9], a weighted least squares procedure was adopted.

The weighted regression was implemented by applying a weighted least squares procedure to minimize the error of the following weighted non-linear function of $\boldsymbol{X_I}$:

$$Y_i^w = w_i \left[\frac{b_1 - b_2}{1 + e^{-\left(\frac{X_i - b_3}{|b_4|}\right)}} + b_2 \right] + e_i^w ,\ i = 1 \ldots n \quad ,$$

where initial estimates of the logistic function [1] parameters are:

$$b_1 = \max(Y_i)$$
$$b_2 = \min(Y_i)$$
$$b_3 = \overline{X}$$
$$b_4 = 1$$

and other parameters are

$$w_i = \frac{1}{\sqrt{s_{Y_i}^{\,2}}}$$
$$Y_i = i^{th} DMOS$$
$$s_i^2 = \text{var}\,iance\ corresponding\ to\ the\ i^{th} DMOS$$
$$Y_i^w = w_i \cdot Y_i$$
$$e_i^w = w_i \cdot e_i$$
$$e_i = i^{th} residual$$

The fitting procedure yielded for each model a set of transformed objective model response values, $\hat{X}_i , i = 1 \ldots n$. The weighted correlation coefficient, $\boldsymbol{r_w}$, between the fitted objective model responses, $\hat{X}_i$, and the DMOS values, Y_i , was then calculated using the following weighted linear correlation function:

$$r^w = \frac{\sum_{i=1}^n w_i \left(\hat{X}_i - \overline{X}^w\right)\left(Y_i - \overline{Y}^w\right)}{\sqrt{\sum_{i=1}^n w_i \left(\hat{X}_i - \overline{X}^w\right)^2 \cdot \sum_{i=1}^n w_i \left(Y_i - \overline{Y}_i\right)^2}},$$

where

$$\hat{X}_i = i^{th}\ fitted\ objective\ score\ from\ weighted\ regression\ as\ previously\ described,$$

$$\overline{X}^w = \frac{\sum_{i=1}^n \hat{X}_i w_i}{\sum_{i=1}^n w_i} \quad (= weighted\ mean\ of\ \hat{X}_i),$$

$$Y_i = i^{th}\ DMOS,$$

$$\overline{Y}^w = \frac{\sum_{i=1}^n Y_i w_i}{\sum_{i=1}^n w_i} \quad (= weighted\ mean\ of\ Y_i),$$

$$w_i = \frac{1}{s_Y^2},$$

s_Y^2 [illegible] of Y_i

Figure 2 shows the variance-weighted regression correlations and their associated 95% confidence intervals for each proponent model calculated over the main partitions of the subjective data. To determine the statistical significance of the correlations obtained for each proponent model for the four main test quadrants (50Hz/high quality, 50Hz/low quality, 60Hz/high quality and 60Hz/low quality), a Tukey's Honest Significant Difference (HSD) posthoc analysis [12] was conducted under a 10-way repeated measures ANOVA. The results of this analysis indicate that

- the performance of P6 is statistically lower than the performance of the remaining nine models and the performance of P0, P1, P2, P3, P4, P5, P7, P8 and P9 is statistically equivalent.

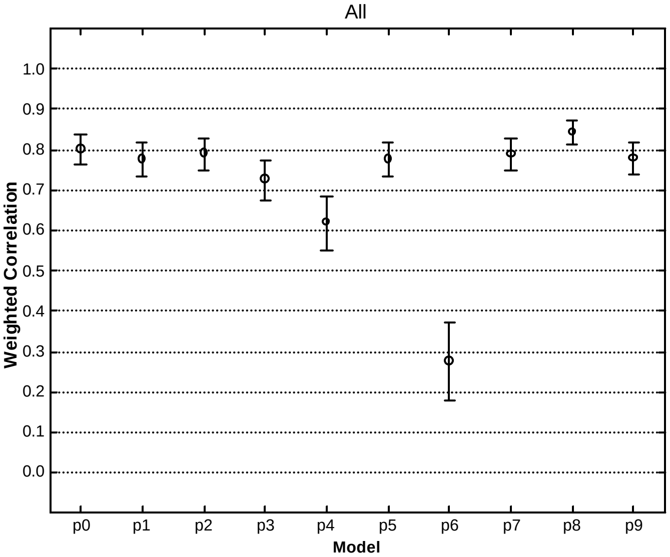

Figure 2. Variance-weighted regression correlations. The figure shows the correlations for each proponent model calculated over the entire subjective data set. The error bars represent 95% confidence intervals.

4.2.2 Non-linear regression analysis (metric 2 [9])

Recognizing the potential non-linear mapping of the objective model outputs to the subjective quality ratings, the objective test plan provided for fitting each proponent's model output with a non-linear function prior to computation of the correlation coefficients. As the nature of the non-linearities was not well known beforehand, it was decided that two different functional forms would be regressed for each model and the one with the best fit (in a least squares sense) would be used for that model. The functional forms used were a 3rd order polynomial and a four-parameter logistic function [1]. The regressions were performed with the constraint that the functions remain monotonic over the full range of the data. For the polynomial function, this constraint was implemented using the procedure outlined in reference [11].

The resulting non-linear regression functions were then used to transform the set of model outputs to a set of predicted DMOS values and correlation coefficients were computed between these predictions and the subjective DMOS. A comparison of the correlation coefficients corresponding to each regression function for the entire data set and the four main test quadrants revealed that in virtually all cases, the logistic fit provided a higher correlation to the subjective data. As a result, it was decided to use the logistic fit for the non-linear regression analysis.

Figure 3 shows the Pearson correlations and their associated 95% confidence intervals for each proponent model calculated over all of the subjective data. To determine the statistical significance of the correlations obtained for each proponent model for the four main test quadrants (50Hz/high quality, 50Hz/low quality, 60Hz/high quality and 60Hz/low quality), a Tukey's HSD posthoc analysis was conducted under a 10-way repeated measures ANOVA. The results of this analysis indicate that

- the performance of P6 is statistically lower than the performance of the remaining nine models and the performance of P0, P1, P2, P3, P4, P5, P7, P8 and P9 is statistically equivalent.

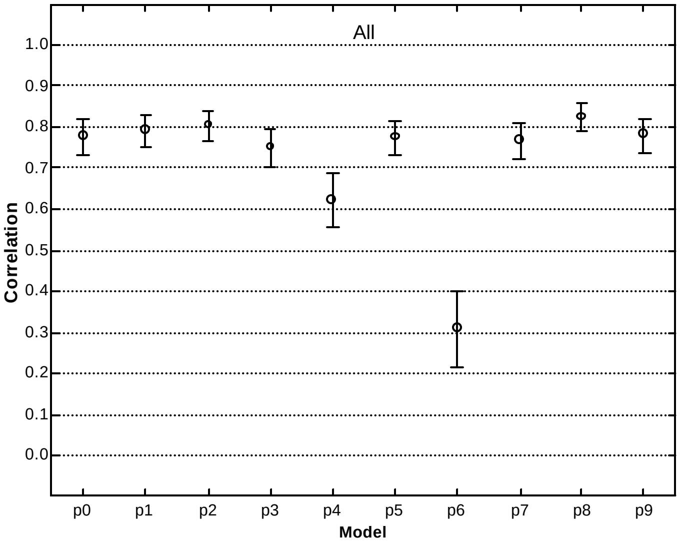

Figure 3. Pearson correlations. The figure shows the correlations for each proponent model calculated over the entire subjective data set. The error bars represent 95% confidence intervals.

4.2.3 Spearman rank order correlation analysis (metric 3 [9])

Spearman rank order correlations test for agreement between the rank orders of DMOS and model predictions. This correlation method only assumes a monotonic relationship between the two quantities. A virtue of this form of correlation is that it does not require the assumption of any particular functional form in the relationship between data and predictions.

Figure 4 shows the Spearman rank order correlations and their associated 95% confidence intervals for each proponent model calculated over all of the subjective data. To determine the statistical significance of the correlations obtained for each proponent model for the four main test quadrants (50Hz/high quality, 50Hz/low quality, 60Hz/high quality and 60Hz/low quality), a Tukey's HSD posthoc analysis was conducted under a 10-way repeated measures ANOVA. The results of this analysis indicate that

- the performance of P6 is statistically lower than the performance of the remaining nine models and the performance of P0, P1, P2, P3, P4, P5, P7, P8 and P9 is statistically equivalent.

4.2.4 Outlier analysis (metric 4 [9])

This metric evaluates an objective model's ability to provide consistently accurate predictions for all types of video sequences and not fail excessively for a subset of sequences, i.e., prediction consistency. The model's prediction consistency can be measured by the number of outlier points (defined as having an error greater than some threshold as a fraction of the total number of points). A smaller outlier fraction means the model's predictions are more consistent.

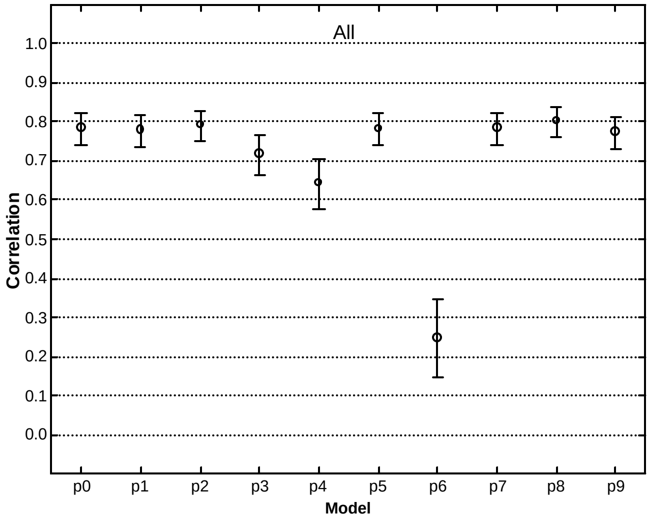

Figure 4. Spearman rank order correlations. The figure shows the correlations for each proponent model calculated over the entire subjective data set. The error bars represent 95% confidence intervals.

The objective test plan [9] specified this metric as follows:

Outlier Ratio = # outliers / N
where an outlier is a point for which
ABS[e_i] > 2 * (DMOS Standard Error)$_i$, $i = 1 \ldots N$
where e_i = i^{th} residual of observed DMOS vs. the predicted DMOS value.

Figure 5 shows the outlier ratios for each proponent model calculated over the main partitions of the subjective data. To determine the statistical significance of the correlations obtained for each proponent model for the four main test quadrants (50Hz/high quality, 50Hz/low quality, 60Hz/high quality and 60Hz/low quality), a Tukey's HSD posthoc analysis was conducted under a 10-way repeated measures ANOVA. The results of this analysis indicate that

- the performance of P6 and P9 is statistically lower than the performance of P8 but statistically equivalent to the performance of P0, P1, P2, P3, P4, P5 and P7 and
- the performance of P8 is statistically equivalent to the performance of P0, P1, P2, P3, P4, P5 and P7.

4.2.5 PSNR performance (reference model)

It is perhaps surprising to observe that PSNR (P0) does so well with respect to the other, more complicated prediction methods. In fact, its performance is statistically equivalent to that of most proponent models for the three performance metrics analyzed thusfar. Some features of the data collected for this effort present possible reasons for this.

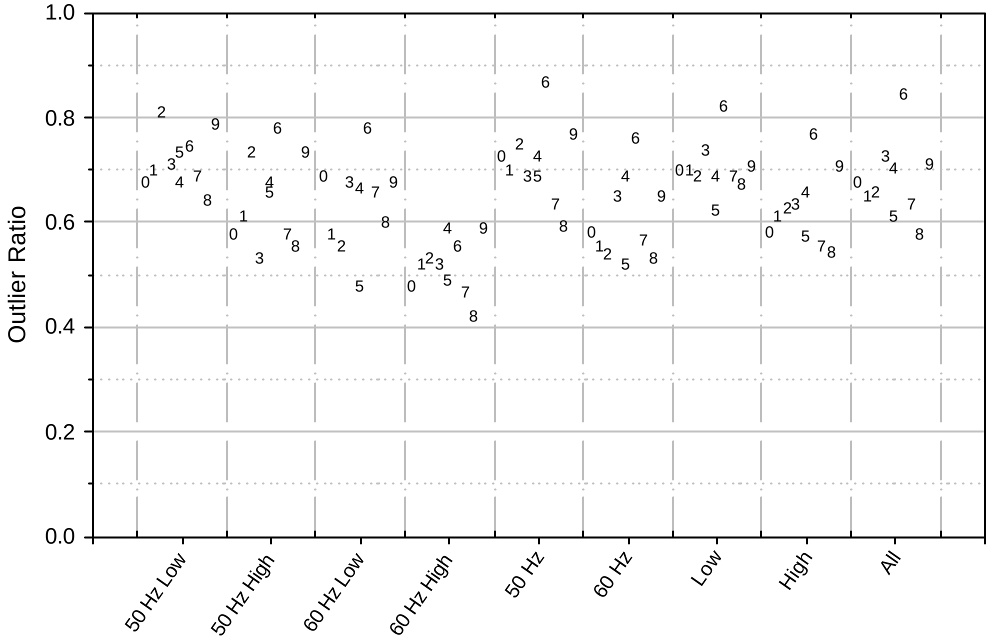

Figure 5. Outlier ratios for each proponent model calculated over different partitions of the subjective data set. The specific data partition is listed along the abscissa while each proponent model is identified by its numerical symbol.

First, it can be noted that in previous smaller studies, various prediction methods have performed significantly better than PSNR. It is suspected that in these smaller studies, the range of distortions (for example, across different scenes) was sufficient to tax PSNR but was small enough so that the alternate prediction methods, tuned to particular classes of visual features and/or distortions, performed better. However, it is believed that the current study represents the largest single video quality study undertaken to date in this range of video quality (768 kb/s to 50 Mb/s). In a large study such as this, the range of features and distortions is perhaps sufficient to additionally tax the proponents' methods, whereas PSNR performs about as well as in the smaller studies.

Another possible factor is that in this study, source and processed sequences were aligned and carefully normalized prior to PSNR and proponent calculations. Because lack of alignment is known to seriously degrade PSNR performance, it could be the case that some earlier results showing poor PSNR performance were due at least in part to a lack of alignment.

Third, it is noted that these data were collected at a single viewing distance (five screen heights, which is quite far), and with a single monitor size and setup procedure. Many proponents' model predictions will change in reasonable ways as a function of viewing distance and monitor size/setup while PSNR by definition cannot. Further, closer viewing distances should differentiate the visible distortions, and hence the models, more strongly. Overall, we expect that broadening the range of viewing conditions will demonstrate better performance from the more complicated models than from PSNR.

5. CONCLUSIONS

There is an urgent need to develop and standardize new objective measurement methods for video quality. The Video Quality Experts Group was formed to combine the expertise and resources from three ITU Study Groups with those from other experts in industry and other standards bodies. The current effort to validate objective methods for video quality utilizes twelve testing laboratories and several of the ten proponent organizations.

Depending on the metric that is used, there are eight or nine models (out of a total of ten) whose performance is statistically indistinguishable. Note that this group of models includes PSNR. PSNR is a measure that was not originally included in the test plans but it was agreed at the third VQEG meeting held in September 1999 at KPN Resesarch in The Netherlands to include it as a reference objective model. It was also discussed and determined at this meeting that three of the proponent models did not generate proper values due to software or other technical problems. Please refer to reference [10] for explanations of their performance.

Based on the analyses presented in its report [10], VQEG is not presently prepared to propose one or more models for inclusion in ITU Recommendations on objective picture quality measurement. Despite the fact that VQEG is not in a position to validate any models, the test was a great success. One of the most important achievements of the VQEG effort is the collection of an important new data set. Prior to the current effort, model developers have had a very limited set of subjectively-rated video data with which to work. Once the VQEG data set is released, future work is expected to substantially improve the state of the art of objective measures of video quality.

With the finalization of this first major effort conducted by VQEG, several conclusions stand out:

- no objective measurement system in the test is able to replace subjective testing,
- no one objective model outperforms the others in all cases,
- the analysis does not indicate that a method can be proposed for ITU Recommendation at this time,
- a great leap forward has been made in the state of the art for objective methods of video quality assessment and
- the data set produced by this test is uniquely valuable and can be utilized to improve current and future objective video quality measurement methods.

6. FUTURE WORK OF VQEG

Concerning the future work of VQEG, there are several areas of interest to participants. These are discussed below. What must always be borne in mind, however, is that the work progresses according to the level of participation and resource allocation of the VQEG members. Therefore, final decisions of future directions of work will depend upon the availability and willingness of participants to support the work.

Because there is still a need for standardized methods of double-ended objective video quality assessment, the most likely course of future work will be to push forward to find a model for the bit rate range covered in this test. This follow-on work will possibly see several proponents working together to produce a combined new model that will, hopefully, outperform any that were in the present test. Likewise, new model proponents are entering the arena anxious to participate in a second round of testing — either independently or in collaboration.

At the same time as the follow-on work is taking place, the investigation and validation of objective and subjective methods for lower bit rate video assessment will be launched. This effort will most likely cover video in the range of 16 kb/s to 2 Mb/s and should include video with and without transmission errors as well as video with variable frame rate, variable temporal alignment, and frame repetition. This effort will validate single-ended (i.e., using no reference or source information) and/or reduced reference objective methods. Since single-ended objective video quality measurement methods are currently of most interest to many VQEG participants, this effort will probably begin quickly.

Another area of particular interest to many segments of the video industry is that of in-service methods for measurement of distribution quality television signals with and without transmission errors. These models could use either single-ended or reduced reference methods [13]. MPEG-2 video would probably be the focus of this effort.

ACKNOWLEDGEMENTS

The authors would like to thank all the participants of VQEG who contributed to the planning and execution of the validation effort.

REFERENCES

1. Recommendation ITU-R BT.500-7, “Methodology for the subjective assessment of the quality of television pictures,” ITU-R, Geneva, 1974-1997.
2. Recommendation ITU-T P.910, “Subjective video quality assessment methods for multimedia applications,” ITU-T, Geneva, 1996.
3. P. Corriveau and A. Webster, “VQEG evaluation of objective methods of video quality assessment,” *SMPTE Journal*, **108,** pp. 645-648, 1999.
4. Recommendation ITU-T P.920, “Interactive test methods for audiovisual communications,” ITU-T, Geneva, 1996.
5. Recommendation ITU-T P.800, “Methods for subjective determination of transmission quality,” ITU-T, Geneva, 1996.
6. Recommendation ITU-T P.861, “Objective quality measurement of telephone-band (300-3400 Hz) speech codecs,” ITU-T, Geneva, 1998.
7. Report of the Gaithersburg VQEG meeting May 1998. Can be accessed (January 2000) at anonymous ftp site: ftp://ftp.its.bldrdoc.gov/dist/ituvidq/vqeg2min.rtf
8. VQEG subjective test plan. Can be accessed (January 2000) at anonymous ftp site: ftp://ftp.its.bldrdoc.gov/dist/ituvidq/subj_test_plan_final.rtf
9. VQEG objective test plan. Can be accessed (January 2000) at anonymous ftp site: ftp://ftp.its.bldrdoc.gov/dist/ituvidq/obj_test_plan_final.rtf
10. A.M. Rohaly, J. Libert, P. Corriveau and A. Webster (eds.), “Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment,” 2000. (Download from ftp://ftp.crc.ca/crc/vqeg)
11. C. Fenimore, J. Libert and M.H. Brill, “Algebraic constraints implying monotonicity for cubics (Technical Report NISTIR 6453),” U.S. DOC/NIST, Gaithersburg, MD, 2000.
12. B.J. Wiener, *Statistical Principles in Experimental Design (2nd edition)*, McGraw-Hill, New York, 1971.
13. ITU-T Study Group 9 Delayed Contribution D.104, “Family of three draft new Recommendations: J.fulref, J.redref, J.noref,” Geneva, 1999.