Available online at www.sciencedirect.com
SCIENCE DIRECT®
Signal Processing: *Image Communication* 19 (2004) 101–107
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Accuracy and cross-calibration of video quality metrics: new methods from ATIS/T1A1

Michael H. Brill[a,*], Jeffrey Lubin[b], Pierre Costa[c], Stephen Wolf[d], John Pearson[e]

[a] *Datacolor, 5 Princess Road, Lawrenceville, NJ 08648, USA*
[b] *Sarnoff Corporation, Princeton, NJ, USA*
[c] *SBC, Austin, TX, USA*
[d] *NTIA/ITS, Boulder, CO, USA*
[e] *Siemens Corporate Research, Princeton, NJ, USA*

Abstract

Video quality metrics (VQMs) have often been evaluated and compared using simple measures of correlation to subjective mean opinion scores from panels of observers. However, this approach does not fully take into account the variability implicit in the observers. We present techniques for determining the statistical resolving power of a VQM, defined as the minimum change in the value of the metric for which subjective test scores show a significant change. Resolving power is taken as a measure of accuracy. These techniques have been applied to the video quality experts group (VQEG) data set and incorporated into the recent Alliance for Telecommunications Industry Solutions (ATIS) Committee T1A1 series of technical reports (TRs), which provide a comprehensive framework for characterizing and validating full-reference VQM. These approved TRs, while not standards, will enable the US telecommunications industry to incorporate VQMs into contracts and tariffs for compressed video distribution. New methods for assessing VQM accuracy and cross-calibrating VQMs are an integral part of the framework. These methods have been applied to two VQMs at this point: peak-signal-to-noise ratio and the version of Sarnoff's just noticeable difference metric (JNDmetrix™) tested by VQEG (Rapporteur Q11/12 (VQEG): Final report from the VQEG on the validation of objective models of video quality assessment, June 2000). The framework is readily extensible to additional VQMs.

Keywords: Video quality; Quality metrics; VQEG; Standards; Telecommunication

1. Introduction

US telecommunication companies use digital techniques of compression for the transport of video over long distances. The transport of a given digital video stream from source to final destination often requires the services of more than one telecommunications company. The companies involved need a mutually accepted way of measuring video quality at the points where the video is transmitted from one network to the next. Only in this way can end-to-end video quality be managed and assured to the end-user. Since these are publicly regulated companies, industry-sanctioned methods for measuring video quality are required. Such objective video quality metrics (VQMs) are

*Corresponding author. Tel.: +1-609-895-7432; fax: +1-609-895-7472.
E-mail address: mbrill@datacolor.com (M.H. Brill).

0923-5965/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0923-5965(03)00073-0

needed to help design and control digital transmission systems.

Many VQMs exist, including several commercial products now being used in varying degrees by industry (for example, see [6] for a description of 10 different VQMs). To assure quality of service across many platforms and VQM choices, there is a need to quantify the accuracy of each VQM and also a need to translate (or cross-calibrate) from the output of one VQM to that of another. VQMs have traditionally been evaluated and compared using simple measures of correlation to subjective mean opinion scores from panels of observers. However, this approach does not fully take into account the variability implicit in the observers. The present paper describes the recent effort of one US telecommunication standards organization, the Alliance for Telecommunications Industry Solutions (ATIS), to develop a better means of quantifying VQM accuracy and cross-calibration. Computation of the quantitative relationships requires the use of an existing subjective data set and corresponding videos (undistorted and distorted). We will describe the new methods in detail and present example computations using a selected subset of subjective data from the video quality experts group (VQEG) [6]. We will also present a summary of possible future improvements that can be made to the method.

2. Background on ATIS T1A1.1

T1A1.1 (Multi-Media Communications Coding and Performance) was a working group under Committee T1 (Telecommunications) of ATIS. (T1A1.1 has recently disbanded and merged with T1A1.3, Performance of Networks and Services). In response to the perceived urgent need for sanctioned guidance on video quality, T1A1.1 decided in February 2001 to develop a series of technical reports (TRs) rather than standards, which can take much longer time to develop. Four of these TRs [1–4] were approved in October 2001 and a fifth TR was approved in January 2002 [5]. The first TR [1] provides an extensible framework into which any properly documented VQM can be incorporated and quantitatively related to other VQMs that are already part of the report series. The second TR [2] provides a description of normalization methods for correcting calibration problems in the distorted video stream (e.g., spatial/temporal shifts, non-unity gains, level shifts) before making any VQM measurements. The third and fourth TRs [3,4] provide full implementation details for two VQMs currently used by industry, peak-signal-to-noise ratio (PSNR) and Sarnoff's JNDmetrix as tested by VQEG, respectively. The fifth TR [5] provides sample data and program code for implementing the VQM accuracy and cross-calibration methods described in [1] for PSNR [3] and Sarnoff's JNDmetrix [4]. All of the TRs are available from www.atis.org/atis/docstore.

3. Initial data analysis

The primary data used in this analysis are subjective scores of various video-source materials subjected to various kinds of digital-processing distortion. These data must be brought into a form in which they can be used to evaluate VQMs.

Let each video-source/distortion combination in a data set be called a "situation", and let N be the number of situations in this data set. A subjective score for situation i and viewer l will be denoted as S_{il}, and an objective score for situation i will be denoted as O_i. Averaging over a variable such as viewer will be denoted with a dot in that variable location. For instance, the mean opinion score of a situation will be denoted as $S_{i\bullet}$. The subjective-score statistics from each pair (i,j) of these situations are to be assessed for significance of VQM difference, and then used to arrive at a resolving power for the VQM difference, as a function of the VQM value.

Prior to any statistical analysis, the original subjective mean opinion scores $S_{i\bullet}$ are linearly transformed to the interval [0,1], defined as the *common scale*, where 0 represents no-impairment and 1 represents most impairment. If *best* represents the no-impairment value of the original subjective scale and *worst* represents the maximum

impairment of the original subjective scale, then the scaled scores $\hat{S}_{i\bullet}$ are given by

$$\hat{S}_{i\bullet} = (S_{i\bullet} - \text{best})/(\text{worst} - \text{best}). \quad (1)$$

Next, the VQM scores are transformed to this common scale as a byproduct of the process of fitting the VQM scores to the subjective data. Fitting removes systematic differences between the VQM and the subjective data (e.g., dc shift) that do not provide any useful quality discrimination information. In addition, fitting all VQMs to one common scale will provide a method for cross-calibration of those VQMs.

The simplest method of data fitting is linear correlation and regression. For subjective video quality scores, this may not be the best method. Experience with other video quality data sets [6] indicates chronically poor fits of VQM to subjective scores at the extremes of the ranges. This problem can be ameliorated by allowing the fitting algorithm to use nonlinear, but still monotonic (order-preserving), methods. If a good nonlinear model is used, the objective-to-subjective errors will be smaller and have a central tendency closer to zero.

Nonlinear methods can be constrained to effectively transform the VQM scale to the [0,1] common scale. Besides improving the fit of data with a VQM, a fitting curve also offers an additional advantage over the straight-line fit implied by the native scale (i.e., the original scale of the VQM): the distribution of objective-to-subjective errors around the fitted model curve is less dependent on the VQM score. Of course, the nonlinear transformation may not remove all the score dependency of objective-to-subjective errors. To capture the residual dependence, it would ideally have been useful to record objective-to-subjective error as a function of VQM value. However, our database was too small to divide among VQM bins in a statistically robust way. Therefore, as will be clear in Section 4, we compute a sort of average measure over the VQM range.

We denote the original (native scale) objective scores as O_i, and the common scale objective scores as $\hat{O}_i$. A fitting function F (depending on some fitting parameters) connects the two. The function used to fit the objective VQM data (O_i) to the scaled subjective data ($\hat{S}_{i\bullet}$) must have the following three attributes: (a) a specified domain of validity, which should include the range of VQM data for all the situations used to define the accuracy metric; (b) a specified range of validity, defined as the range of Common Scale scores (a sub-range of [0,1]) to which the function maps; and (c) monotonicity (the property of being either strictly increasing or strictly decreasing) over the specified domain of validity. Of course, the fitting function would be most useful as a cross-calibration tool if it were monotonic over the entire theoretical domain of VQM scores, covered the entire subjective common scale from 0 to 1, and mapped to zero the VQM score that corresponds to a perfect video sequence (no degradations, hence a null distortion). However, this ideal may not be attainable for certain VQMs and function families used to perform the fit.

One possible family of fitting functions is the set of polynomials of order M. There are also two kinds of logistic functions. All these are discussed in [1]. The selection of a fitting-function family (including a priori setting of some of the parameters) depends on the asymptotic (best and worst) scores of the particular VQM.

4. VQM accuracy algorithm

We define a new quantitative measure of VQM accuracy, called *resolving power*, defined as the ΔVQM value above which the conditional subjective-score distributions have means that are statistically different from each other (typically at the 0.95 significance level). Such an "error bar" measure is needed in order for video service operators to judge the significance of VQM fluctuations.

Of several ways to assess a VQM's resolving power, the Student's t-test was chosen. This test was applied to the measurements in all pairs i and j of situations. Emerging from the test are the ΔVQM (i.e., the difference between the greater and lesser VQM scores of i and j) and the *significance* from the t-test. This *significance* is the probability p that, given i and j, the greater VQM score is

associated with the situation that has the greater true underlying mean subjective score. Thus, p is the probability that the observed difference in sample means of the subjective scores from i and j did not come from a single population mean, nor from population means that were ordered oppositely to the associated VQM scores. To capture this ordering requirement, the t-test must be one-tailed. For simplicity, the t-test was approximated by a z-test. This approximation is a close one when the number of viewers is large, as was the case for the VQEG data set [6].

An analysis of variance (ANOVA) test might seem better than the t-test method. However, although a single application of ANOVA will determine whether a statistical separation exists among a set of categories, further paired comparisons are needed to determine the magnitudes and conditions of the statistically significant differences. Also, ANOVA assumes equal category-data variances (which may not be true). Finally, although ANOVA resides in many software packages, finding the right software package may not be easy (e.g., not all ANOVA routines will accept different quantities of data in different categories).

Ref. [5] provides data and program code to fully implement the VQM accuracy algorithm that is described here. The algorithm has the following steps:

Step 1: Start with an input data table with N rows, each row representing a different situation (i.e., a different source video and distortion). Each row i consists of the following: the source number, the distortion number, the VQM scores O_i, the number of viewer responses N_i, the mean subjective opinion score $S_{i\bullet}$, and the sample variance of the subjective scores V_i.

Step 2: Transform the subjective scores $S_{i\bullet}$ to the common scale $\hat{S}_{i\bullet}$ as described in Section 3. The variance V_i of the subjective scores must also be scaled accordingly as

$$\hat{V}_i = V_i/(\text{worst} - \text{best})^2. \tag{2}$$

Note that transforming the subjective scores and their variances is optional. It will not change the z-statistic defined below, but it may change the VQM fitting process. Next, transform the VQM scores O_i to the common scale using a fitting function as discussed in Section 3. The result of the fitting process is a set of common scale VQM scores $\hat{O}_i$. Display the coefficient values used in the fit, and also the VQM domain over which the fit was done (domain of validity).

Step 3: For each pair of distinct situations i and j $(i \neq j)$, use a one-tailed z-test to assign a probability of *significance* to the difference between the greater and lesser VQM ($\hat{O}_i$ and $\hat{O}_j$, respectively). The significance is the probability that the greater VQM score comes from the situation with the greater true underlying mean subjective score. The z-score is

$$z = (\hat{S}_{i\bullet} - \hat{S}_{j\bullet})/\sqrt{(\hat{V}_i/N_i + \hat{V}_j/N_j)} \tag{3}$$

and the probability of significance of the z-score $p(z)$ is just the cumulative distribution function of z:

$$p(z) = c\,df(z) = (2\pi) - 0.5 \int_{-\infty}^{z} \exp(-z^2/2)\,dz. \tag{4}$$

Step 4: Create a scatter plot of $p(z)$ (ordinate) versus ΔVQM score (abscissa). Given N situations, record each pair (i,j) with $i > j$, record the VQM difference $\hat{O}_i - \hat{O}_j$ in a vector of length $N(N-1)/2$ called **ΔVQM** (with index k), and record the corresponding z-score in a vector called **Z** with length $N(N-1)/2$ (with the same index k). It is desired that $\Delta\text{VQM}(k)$ is always nonnegative, which can be ensured by definition of the otherwise arbitrary ordering of the endpoints i and j. To ensure that this is so, if $\Delta\text{VQM}(k)$ is negative, then replace $Z(k)$ by $-Z(k)$ and $\Delta\text{VQM}(k)$ by $-\Delta\text{VQM}(k)$.

Fig. 1 provides a scatter plot of the subjective **Z** vector versus the **ΔVQM** vector for the JNDmetrix VQM on the common scale, before computation of $p(z)$ as given in Step 3.

Step 5: Consider 19 bins (indexed by m) of ΔVQM, each one of which spans 1/10 of the total range of ΔVQM. The bins overlap by 50%. Associate ΔVQM_m with the midpoint of each bin and associate p_m with the mean of $p(z)$ for all z in bin m.

Step 6: Draw a curve through the points $(\Delta\text{VQM}_m, p_m)$ to produce a graph of p versus

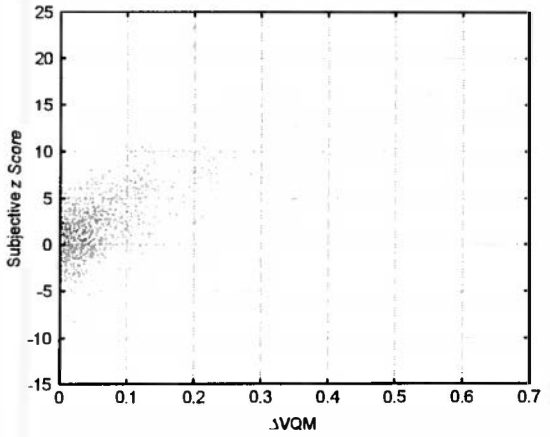

Fig. 1. Scatter plot of subjective z-score versus ΔVQM for the JNDmetrix of [4] on the common scale.

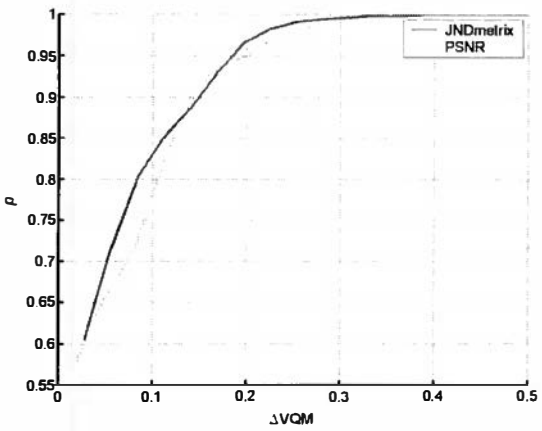

Fig. 2. Confidence versus common-scale ΔVQM score, for PSNR (dashed curve) and JNDmetrix (solid curve).

ΔVQM. Note that p can be interpreted as the average probability of significance.

Step 7: Select a threshold probability p, draw a horizontal line at the ordinate value p, and let its intercept with the curve of step 6 determine the threshold ΔVQM, defined as the accuracy. For an average probability of significance p or greater, the ΔVQM should exceed this threshold. Common choices of p are 0.68, 0.75, 0.90, and 0.95.

Having found a value of ΔVQM for a chosen p, one can use it directly in common scale, as would be appropriate for cross-calibration in Section 6. Alternatively, for other purposes, one has the option of mapping this ΔVQM value back to the native scale to give a native scale resolving power R as a function of the native objective score O:

$$R(O) = |F^{-1}[F(O) + \Delta\mathrm{VQM}] - O|, \quad (5)$$

where F is the fitting function defined in Section 3.

Fig. 2 shows a plot of average probability of significance (or confidence) versus common scale ΔVQM score, for PSNR [3] and JNDmetrix [4]. By choosing an acceptable confidence, one arrives at a value of ΔVQM which is adopted as the common scale resolving power of the VQM. Note that, at 0.90 confidence, the resolving powers of PSNR and JNDmetrix are 0.14 and 0.15, respectively, whereas at 0.95 confidence, the respective resolving powers are 0.20 and 0.18. For this VQEG subjective data subset [5], there is no clear trend indicating that either PSNR or JNDmetrix has more resolving power in general.

5. VQM classification errors

Ref. [5] also provides sample data and program code for computation of classification errors, which is another way to evaluate the effectiveness of a VQM. A classification error is made when the subjective test and the VQM lead to different conclusions on a pair of situations. There are three different types of classification errors that can arise when using a VQM. The "false tie" error is probably the least offensive error. This occurs when the subjective test says two situations are different but the VQM says they are the same. A "false differentiation" error is usually more offensive. This occurs when the subjective test says two situations are the same but the VQM says they are different. The "false ranking" error would generally be the most offensive error. In false ranking, the subjective test says situation i is better than situation j, but the VQM says just the opposite.

Fig. 3 presents an example plot of the relative frequencies for the different classification errors versus ΔVQM. For the plot, the common scale was used for both the objective (PSNR) and subjective scores (VQEG data subset) and the z-score

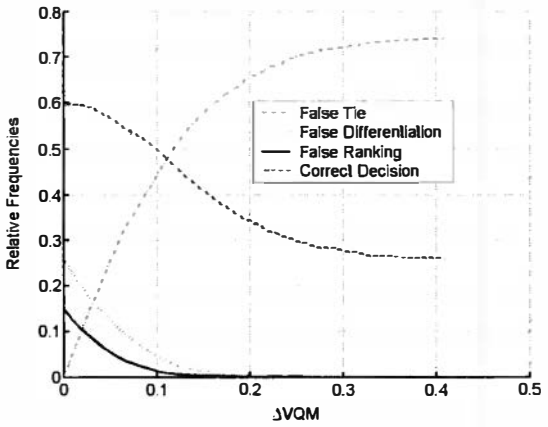

Fig. 3. Frequencies for the different types of classification errors for PSNR of [3] where the subjective and objective scores are both on the common scale.

threshold was selected to give an estimated 95% confidence in the subjective classifications. Note that as ΔVQM is increased, the VQM will declare more and more pairs of situations as equivalent. This reduces the occurrences of false differentiations and false rankings, but increases the occurrence of false ties. One might use a graph like this to select an appropriate value of ΔVQM. For example, one might select ΔVQM to maximize the probability of making correct decisions, or to minimize some weighted sum of the different classification errors. Note that the four outcomes shown in Fig. 3 are exhaustive and thus the relative frequencies will add up to 1.0 for each value of ΔVQM.

6. VQM cross-calibration

The need to relate two VQMs is met by the transformation to a common scale described in Section 3. Once two VQMs (say, PSNR and JNDmetrix) are transformed to the common scale (through an agreed-upon subjective data set), the transformation from PSNR to JNDmetrix is the forward transformation from PSNR to the common scale, composed with the inverse transformation from JNDmetrix to the common scale. If the domains or ranges of the mapping mismatch, the cross-calibration is undefined.

Cross-calibration of two VQMs does not mean one of the VQMs can be substituted error-free for the other. One reason is that the present cross-calibration depends on the particular subjective data that define the common scale. More fundamentally, even within the chosen set of situations, there are likely four situations (call them 1–4) such that both VQM scores change in the same direction going from 1 to 2, but in opposite directions going from 3 to 4. Such behavior makes one VQM better than another, and cannot be captured in any cross-calibration method.

7. Outlook

The TIAI methods promise to be useful in an arena even larger than US telecommunications. For example, they are included in the test plans for future VQEG efforts [http://www.its.bldrdoc.gov/vqeg/]. However, there is room for improvement.

In step 5 of Section 4, the use of bins and bin averages produces the possible (though perhaps rare) problem of a low occupation number in a bin. Such an occurrence would be less likely if one were able to merge bins as much as possible. A different method that does not use bins follows. First, select a z-score threshold (i.e., percent confidence, as defined in step 3 of Section 4) that is appropriate for the application. For example, select $z = 2.4$ to get a confidence of 99%. Next, accumulate z-scores above and below this z-score threshold to the right of a particular ΔVQM score, and adjust that ΔVQM score to its minimum value such that, say, at least 95% of the data points lie above the z-score threshold. In this example, the resulting value of ΔVQM gives the minimum VQM difference required so that 95% of the situation pair decisions have at least 99% confidence.

Another useful exercise would be to define the accuracy of the cross-calibration between two VQMs. This would provide an automatic safeguard against too much trust in cross-calibrating from say a good VQM to a bad one, apropos of the caveat in Section 6.

Acknowledgements

Thanks are due to John Grigg for chairing the T1A1.1 effort, and to Karen Pitts for her statistical advice. Finally, we thank Margaret Pinson and Steve Voran of NTIA, Hui Cheng at Sarnoff, and Dave Fibush of Tektronix, for help and clarifying discussions.

References

[1] ATIS Technical Report T1.TR.72-2001, Methodological framework for specifying accuracy and cross-calibration of video quality metrics. Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC 20005, October 2001.

[2] ATIS Technical Report T1.TR.73-2001, Video normalization methods applicable to objective video quality metrics utilizing a full reference technique, October 2001.

[3] ATIS Technical Report T1.TR.74-2001, Objective video quality measurement using a peak-signal-to-noise-ratio (PSNR) full reference technique, October 2001.

[4] ATIS Technical Report T1.TR.75-2001, Objective video quality measurement using a JND-based full reference technique, October 2001.

[5] ATIS Technical Report T1.TR.77-2002, Data and sample program code to be used with the method specified in T1.TR.72-2001 for the calculation of resolving power of the video quality metrics in T1.TR.74-2001 and T1.TR.75-2001, January 2002.

[6] ITU-T COM 9-80-E, Rapporteur Q11/12 (VQEG): final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, June 2000.