

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

J.149

(03/2004)

SERIES J: CABLE NETWORKS AND TRANSMISSION
OF TELEVISION, SOUND PROGRAMME AND OTHER
MULTIMEDIA SIGNALS

Measurement of the quality of service

Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)

Recommendation ITU-T J.149

Recommendation ITU-T J.149

Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)

Summary

Video quality metrics are intended to provide calculated values that are strongly correlated with viewer subjective assessments. This Recommendation provides methods for curve fitting VQM objective values to subjective data in order to facilitate the accuracy calculation, an algorithm to quantify the accuracy of a given VQM, a simplified root mean square error calculation to quantify the accuracy of a VQM when the subjective data has roughly equal variance across the VQM scale, and a method to plot classification errors to determine the relative frequencies of "false tie", "false differentiation", "false ranking", and "correct decision" for a given VQM.

Source

Recommendation ITU-T J.149 was approved on 15 March 2004 by ITU-T Study Group 9 (2001-2004) under Recommendation ITU-T A.8 procedure.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications, information and communication technologies (ICTs). The ITU Telecommunication Standardization Sector (ITU-T) is a permanent organ of ITU. ITU-T is responsible for studying technical, operating and tariff questions and issuing Recommendations on them with a view to standardizing telecommunications on a worldwide basis.

The World Telecommunication Standardization Assembly (WTSA), which meets every four years, establishes the topics for study by the ITU-T study groups which, in turn, produce Recommendations on these topics.

The approval of ITU-T Recommendations is covered by the procedure laid down in WTSA Resolution 1.

In some areas of information technology which fall within ITU-T's purview, the necessary standards are prepared on a collaborative basis with ISO and IEC.

NOTE

In this Recommendation, the expression "Administration" is used for conciseness to indicate both a telecommunication administration and a recognized operating agency.

Compliance with this Recommendation is voluntary. However, the Recommendation may contain certain mandatory provisions (to ensure e.g. interoperability or applicability) and compliance with the Recommendation is achieved when all of these mandatory provisions are met. The words "shall" or some other obligatory language such as "must" and the negative equivalents are used to express requirements. The use of such words does not suggest that compliance with the Recommendation is required of any party.

INTELLECTUAL PROPERTY RIGHTS

ITU draws attention to the possibility that the practice or implementation of this Recommendation may involve the use of a claimed Intellectual Property Right. ITU takes no position concerning the evidence, validity or applicability of claimed Intellectual Property Rights, whether asserted by ITU members or others outside of the Recommendation development process.

As of the date of approval of this Recommendation, ITU had not received notice of intellectual property, protected by patents, which may be required to implement this Recommendation. However, implementers are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database at <http://www.itu.int/ITU-T/ipr/>.

© ITU 2009

All rights reserved. No part of this publication may be reproduced, by any means whatsoever, without the prior written permission of ITU.

CONTENTS

Page

1	Scope	1
2	Informative references	1
3	Abbreviations.....	2
4	Accuracy of a VQM.....	2
4.1	Nomenclature and coordinate scales	2
4.2	Fitting VQM values to subjective data.....	3
4.3	Metric 1: VQM accuracy based on statistical significance	5
4.4	Metric 2: VQM RMSE calculation.....	7
4.5	Classification plots	7
5	Cross-calibrating two VQMs.....	10
Appendix I – Application of this Recommendation in the evaluation and validation of proposed VQMs.....		11
I.1	Elements of a full VQM disclosure	11
I.2	Scope/limitations of a VQM.....	11
Appendix II – MATLAB Source Code.....		14
Appendix III – Data-fitting to a common scale of VQM.....		20
III.1	Polynomial of order M	20
III.2	Logistic function I	20
III.3	Logistic function II.....	20
Bibliography.....		22

Recommendation ITU-T J.149

Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM)

1 Scope

Video quality metrics are intended to provide calculated values that are strongly correlated with viewer subjective assessments. This Recommendation provides:

- a) methods for curve fitting VQM objective values to subjective data in order to facilitate the accuracy calculation and to produce a normalized objective value scale that can be used for cross-correlation between different VQMs;
- b) an algorithm (based on statistical analysis relative to subjective data) to quantify the accuracy of a given VQM;
- c) a simplified root mean square error calculation to quantify the accuracy of a VQM when the subjective data has roughly equal variance across the VQM scale;
- d) a method to plot classification errors to determine the relative frequencies of "false tie", "false differentiation", "false ranking", and "correct decision" for a given VQM.

The methods specified in this Recommendation are based on objective and subjective evaluation of component video such as defined by ITU-R Rec. BT.601 using methods such as described in ITU-R Rec. BT.500-11. A data set for a VQM will consist of objective values and mean subjective scores for a variety of motion video sources (SRC) processed by a variety of hypothetical reference circuits (HRC). An example of such a data set is given in the ITU-T Tutorial (see Appendix I).

The methods specified in this Recommendation are directly applicable to a defined data set as described above. For measurements not specifically part of the data set, the methods specified in this Recommendation provide a reasonable estimate of accuracy and cross-calibration for applications that can be considered to be similar to and within the scope of the defined data set.

The methods specified in this Recommendation are appropriate for use in combination with other statistical calculations in order to evaluate the usefulness of a VQM. Informative information regarding the use of the methods is presented in Appendix I. A complete verification process by suitable independent laboratories is required for a VQM to be considered for inclusion as a normative part of an ITU-R Recommendation.

NOTE – The structure and content of this Recommendation have been organized for ease of use by those familiar with the original source material; as such, the usual style of ITU-T recommendations has not been applied.

2 Informative references

- ANSI T1.801.01-1995*, *Digital Transport of Video Teleconferencing/Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment*.
- ANSI T1.801.02-1996, *Digital Transport of Video Teleconferencing/Video Telephony Signals – Performance Terms, Definitions and Examples*.
- ANSI T1.801.03-2003, *Digital Transport of One-Way Digital Signals – Parameters for Objective Performance Assessment*.
- IEEE Standard No. 205-2001, *Measurement of Luminance Signal Levels*.

* T1 standards are maintained since November 2003 by ATIS.

- ITU-T Tutorial (2004), *Objective perceptual assessment of video quality: Full reference television* (www.itu.int/ITU-T/studygroups/com09/docs/tutorial_opavc.pdf)
- ITU-R Recommendation BT.500-11 (2002), *Methodology for the subjective assessment of the quality of television pictures*.
- U.S. Standards Committee T1 Technical Report T1.TR.73-2001, *Video Normalization Methods Applicable to Objective Video Quality Metrics Utilizing a Full Reference Technique*.
- U.S. Standards Committee T1 Technical Report T1.TR.74-2001, *Objective Video Quality Measurement Using Peak-Signal-to-Noise Ratio Full Reference Technique*.
- U.S. Standards Committee T1 Technical Report T1.TR.75-2001, *Objective Perceptual Video Quality Measurement Using a JND-Based Full Reference Technique*.
- U.S. Standards Committee T1 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*.

3 Abbreviations

This Recommendation uses the following abbreviations:

FR-TV	Full Reference Television
HRC	Hypothetical Reference Circuit
RMSE	Root Mean Squared Error
SRC	Source
VQEG	Video Quality Experts Group
VQM	Video Quality Metrics

4 Accuracy of a VQM

In order to use an objective video-quality metric (VQM), one must know whether the score difference between two processed videos is statistically significant. Hence, a quantification is needed of the accuracy (or resolving power) of the VQM. To visualize this resolving power, it helps to begin with a scatter plot in which the abscissa of each point is a VQM score from a particular video source (SRC) and distortion (hypothetical reference circuit, or HRC), and the ordinate is a subjective score from a particular viewing of the SRC/HRC. Each SRC/HRC combination (associated with a particular VQM score) contains a distribution of mean subjective scores, S , based on a number of viewers, which represents (approximately) the relative probabilities of S for the particular SRC/HRC combination. The resolving power of a VQM can be defined as the difference between two VQM values for which the corresponding subjective-score distributions have means that are statistically different from each other (typically at the 0.95 significance level).

Given this qualitative picture, two metrics for resolving power will be described in this clause, each one being useful in a different context. The metrics are described in clauses 4.3 and 4.4. Also, in clause 4.5, a method is described for evaluating the frequencies of different kinds of errors made by the VQM. As an example of implementation of all the methods, a computer source code in MATLAB (The Mathworks, Inc., Natick, MA) is provided in Appendix II.

4.1 Nomenclature and coordinate scales

Let each SRC/HRC 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 score and *worst* represents the maximum impairment of the original subjective scale, then the scaled scores $\hat{S}_{i\bullet}$ are given by:

$$\hat{S}_{i\bullet} = \frac{S_{i\bullet} - \text{best}}{\text{worst} - \text{best}}$$

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, which will be discussed in the following clause.

4.2 Fitting VQM values to 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 (see ITU-T Tutorial) 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 non-linear, but still monotonic (order-preserving), methods. If a good non-linear model is used, the objective-to-subjective errors will be smaller and have a central tendency closer to zero.

Non-linear 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 non-linear 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, typical data sets are too small to divide among VQM bins in a statistically robust way. Therefore, as will be clear in clause 4.3, a sort of average measure over the VQM range is computed.

Figure 1 shows the improved fit of model to data incurred by transforming the objective scores using a fitting function. It can be seen that, besides improving the fit of data with VQM, the curve also offers an additional advantage over the straight-line fit implied by the Native Scale: the distribution of model-to-data errors around the fitted model curve is less dependent on the VQM score.