INTERNATIONAL TELECOMMUNICATION UNION

ITU-T J.144

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

(03/2001)

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

Measurement of the quality of service

Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference

ITU-T Recommendation J.144

(Formerly CCITT Recommendation)

ITU-T J-SERIES RECOMMENDATIONS

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

General Recommendations	J.1–J.9
General specifications for analogue sound-programme transmission	J.10–J.19
Performance characteristics of analogue sound-programme circuits	J.20–J.29
Equipment and lines used for analogue sound-programme circuits	J.30–J.39
Digital encoders for analogue sound-programme signals	J.40–J.49
Digital transmission of sound-programme signals	J.50–J.59
Circuits for analogue television transmission	J.60–J.69
Analogue television transmission over metallic lines and interconnection with radio-relay links	J.70–J.79
Digital transmission of television signals	J.80–J.89
Ancillary digital services for television transmission	J.90–J.99
Operational requirements and methods for television transmission	J.100–J.109
Interactive systems for digital television distribution	J.110–J.129
Transport of MPEG-2 signals on packetised networks	J.130–J.139
Measurement of the quality of service	**J.140–J.149**
Digital television distribution through local subscriber networks	J.150–J.159
IPCablecom	J.160–J.179
Miscellaneous	J.180–J.199
Application for Interactive Digital Television	J.200–J.209

For further details, please refer to the list of ITU-T Recommendations.

ITU-T Recommendation J.144

Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference

Summary

This Recommendation provides guidelines on the selection of appropriate objective perceptual video quality measurement equipment designed for use in digital cable television applications when the full reference video signal is available.

Source

ITU-T Recommendation J.144 was prepared by ITU-T Study Group 9 (2001-2004) and approved under the WTSA Resolution 1 procedure on 9 March 2001.

FOREWORD

The International Telecommunication Union (ITU) is the United Nations specialized agency in the field of telecommunications. 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.

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, implementors are cautioned that this may not represent the latest information and are therefore strongly urged to consult the TSB patent database.

CONTENTS

Introduction

Digital television produces new quality of service considerations, with complex relationships between objective parameter measurements and subjective picture quality. While objective measurements with good correlation to subjective quality assessment are desirable in order to attain optimal quality of service in the operation of cable television systems, it must be realized that objective measurements are not a direct replacement for subjective quality assessment.

Subjective quality assessments are carefully designed procedures intended to determine the average opinion of human viewers to a specific set of video sequences for a given application. Results of such tests are valuable in basic system design and benchmark evaluations. Subjective quality assessments for a different application with different test conditions will still provide meaningful results; however, opinion scores for the same set of video sequences are likely to have different values. Objective measurements are intended for use in a broad set of applications producing the same results with a given set of video sequences. The choice of video sequences to use and the interpretation of the resulting objective measurements are some of the factors varied for a specific application.

Therefore objective measurements and subjective quality assessment are complementary rather than interchangeable. Where subjective assessment is appropriate for research related purposes, objective measurements are required for equipment specifications and day-to-day system performance measurement and monitoring.

The following terminology convention is adopted for the purpose of this Recommendation:

– The term "subjective assessment" refers to the determination of the quality or impairment of programme-like pictures presented to a panel of human assessors in viewing sessions.

– The term "objective perceptual measurement" refers to the measurement of the performance of a programme chain by the use of programme-like pictures and objective (instrumental) measurement methods to obtain an indication that approximates the rating that would be obtained from a subjective assessment test.

– The term "signal measurement" refers to the measurement of the performance of a programme chain by the use of test signals and objective (instrumental) measurement methods.

In this Recommendation the terms objective measurement and perceptual measurement may be used interchangeably to mean objective perceptual measurement.

There are three basic methods to perform objective measurements:

• FR – A method applicable when the full reference video signal is available. This is a double-ended method and is the subject of this Recommendation.

• RR – A method applicable when only reduced video reference information is available. This is also a double-ended method and is the subject of a separate Recommendation [under study].

• NR – A method applicable when no reference video signal or information is available. This is a single-ended method and the subject of a separate Recommendation [under study].

The three methods have different applications, and they provide different degrees of measurement accuracy, expressed in terms of correlation with subjective assessment results.

ITU-T Recommendation J.144

Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference

1 Scope

This Recommendation provides guidelines on the selection of appropriate perceptual video quality measurement equipment for use in digital cable television applications when the full reference measurement method can be used.

The full reference measurement method is intended for use when the unimpaired reference video signal is readily available at the measurement point, as may be the case of measurements on individual equipment or a chain in the laboratory or in a closed environment such as a cable television head-end.

2 References

2.1 Normative references

The following ITU-T Recommendations and other references contain provisions which, through reference in this text, constitute provisions of this Recommendation. At the time of publication, the editions indicated were valid. All Recommendations and other references are subject to revision; users of this Recommendation are therefore encouraged to investigate the possibility of applying the most recent edition of the Recommendations and other references listed below. A list of the currently valid ITU-T Recommendations is regularly published.

– ITU-R BT.500-9 (1998), *Methodology for the subjective assessment of the quality of television pictures*.

2.2 Informative references

– ITU-T J.140 (1998), *Subjective picture quality assessment for digital cable television systems*.

– ITU-T J.143 (2000), *User requirements for objective perceptual video quality measurements in digital cable television*.

– ITU-T P.910 (1996), *Subjective video quality assessment methods for multimedia applications*.

– ITU-T Study Group 9, Contribution COM 9-80 (2000), *Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment*.

3 Terms, definitions and acronyms

This Recommendation defines the following terms:

3.1 subjective assessment: The determination of the quality or impairment of programme-like pictures presented to a panel of human assessors in viewing sessions.

3.2 objective perceptual measurement: The measurement of the performance of a programme chain by the use of programme-like pictures and objective (instrumental) measurement methods to obtain an indication that approximates the rating that would be obtained from a subjective assessment test.

3.3 signal measurement: The measurement of the performance of a programme chain by the use of test signals and objective (instrumental) measurement methods.

4 User requirements

User requirements for perceptual measurement methods of picture quality are given in ITU-T J.143.

5 Description of the full reference measurement method

The double-ended measurement method with full reference, for objective measurement of perceptual video quality, evaluates the performance of systems by making a comparison between the undistorted input, or reference, video signal at the input of the system, and the degraded signal at the output of the system (Figure 1).

Figure 1 shows an example of application of the full reference method to test a codec in the laboratory.

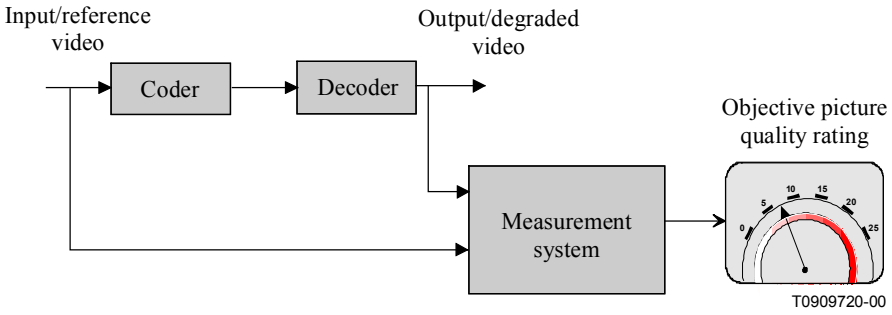


Figure 1/J.144 – Application of the full reference perceptual quality measurement method to test a codec in the laboratory

The comparison between input and output signals may require a spatial and temporal alignment process to compensate for any vertical or horizontal picture shifts or cropping. It also may require correction for any offsets or gain differences in both the luminance and the chrominance channels. The objective picture quality rating is then calculated, typically by applying a perceptual model of human vision.

As the diagnostic tool is based on a human vision model, rather than on the measurement of specific coding artefacts, it is in principle equally valid for analogue systems and for digital systems. It is also in principle valid for chains where analogue and digital systems are mixed, or where digital compression systems are concatenated.

Figure 2 shows an example of the application of the full reference method to test a transmission chain.

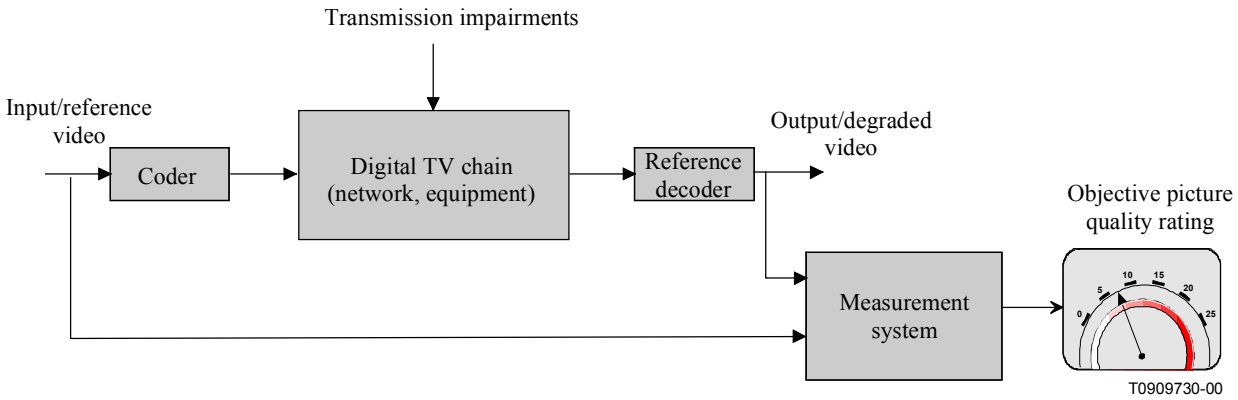


Figure 2/J.144 – Application of the full reference perceptual quality measurement method to test a transmission chain

In this case a reference decoder is fed from various points in the transmission chain, e.g. the decoder can be located at a point in the network, as in Figure 2, or directly at the output of the encoder as in Figure 1. If the digital transmission chain is transparent, the measurement of objective picture quality rating at the source is equal to the measurement at any subsequent point in the chain.

It is generally accepted that the full reference method provides the best accuracy for perceptual picture quality measurements. The method has been proven to have the potential for high correlation with subjective assessments made in conformity with the DSCQS methods specified in ITU-R BT.500.

6 Findings of the Video Quality Experts Group (VQEG)

Studies of perceptual video quality measurements are conducted in an informal group, called Video Quality Experts Group (VQEG), which reports to ITU-T Study Groups 12 and 9 and ITU-R Study Group 6. The first task of VQEG was to assess the performance of proposed double-ended perceptual video quality measurement algorithms.

VQEG issued a comprehensive final draft report on the first phase of its work in March 2000.

Readers are advised to study that report to gain complete insight on the work performed by VQEG until that time. In a nutshell, the report shows the results of tests performed on ten models submitted to VQEG by ten different proponents, used in the calculation of objective scores compared with subjective evaluation over a broad range of video systems and source sequences. The tests compared the performance of the proponent models against subjective assessment tests of the same images, and also against the PSNR (peak signal-to-noise ratio) "reference" algorithm. The aim was to check proponent models in terms of:

- prediction accuracy (the model's ability to predict the subjective quality);
- prediction monotonicity (the degree to which the model's predictions agree with the rank ordering of subjective quality ratings);
- prediction consistency (the degree to which the model maintains prediction accuracy over the range of video test sequences and video systems, i.e. that its response is robust with respect to a variety of video impairments).

Over 26 000 subjective opinion scores were generated based on 20 different source sequences processed by 16 different video systems and evaluated at eight independent laboratories worldwide.

The subjective tests were organized into four quadrants: 50 Hz high quality, 50 Hz low quality, 60 Hz high quality and 60 Hz low quality. High quality in this context refers to production quality video and low quality refers to distribution quality. The high quality quadrants included video at bit

rates between 3 Mbit/s and 50 Mbit/s. The low quality quadrants included video at bit rates between 768 kbit/s and 4.5 Mbit/s.

Strict adherence to ITU-R BT.500-9 procedures for the Double Stimulus Continuous Quality Scale (DSCQS) method was followed in the subjective evaluation. The subjective and objective test plans included procedures for validation analysis of the subjective scores and four metrics for comparing the objective data to the subjective results.

In addition to analysis based on the total data set, subsets based on the four subjective test quadrants and the total data with exclusion of certain video processing systems were analysed to determine sensitivity of results to various application-dependent parameters.

The results obtained from the two algorithms that were not fully tested or were found to have implementation problems were discarded. The VQEG test results based on the analysis obtained for the four individual subjective test quadrants essentially show the following:

- No objective measurement system in the test is able to replace subjective testing.
- No objective model statistically outperforms the others in all reference conditions.
- No objective model statistically outperforms PSNR in all reference conditions.
- Based on present evidence, no single method can be recommended to ITU at this time.

On the positive side, the work performed by VQEG has resulted in a much better understanding of the problem of perceptual video quality testing, and of the users' requirements. This will likely lead to the development of improved perceptual models, implemented in commercial equipment.

Studies are planned inside the IEEE Subcommittee G-2.1.6 to provide a pool of test scenes degraded in a controlled way. Each scene will have a corresponding perceptual scale associated with it, which is calibrated in successive steps of just-noticeable-differences of impairment. These scenes are expected to represent a good pool of reference material to test the forthcoming systems.

7 Conclusions

Since no one method of measurement can be recommended at this time, this clause will list some general advice on the models for video quality assessment utilizing the full reference methodology. Current models evaluated by VQEG are detailed in the appendices. It is intended, based on future work by VQEG and others, to adopt one or more of these models (or new models that may be proposed) as normative. Future VQEG work will also likely consider other test conditions, for example, closer viewing distances and additional types and ranges of distortion, that may allow better discrimination among the objective models and between each model and PSNR.

General advice

When perceptual video quality measurements are performed, using the full reference method described in this Recommendation, operators should first analyse how their specific application and user requirements translate in terms of measuring equipment characteristics and performance.

Some aspects to be taken into consideration are listed below:

- ownership cost of the perceptual measurement equipment;
- vendor's after-sale support;
- ease of operation;
- reliability;
- size, weight, power requirements;
- real-time or non-real-time measurement speed;

- online (in-service) operation;
- prediction accuracy, monotonicity and consistency.

When reporting the results of perceptual video quality measurements, operators should always indicate the brand, model and settings of the perceptual measurement equipment and the test pictures used. This will allow operators to compare the results of those tests with tests performed by other operators.

This precaution is necessary because the full reference perceptual measuring equipment can be expected to provide a degree of correlation with subjective assessment tests that depends, among other factors, on the set of selected test pictures, on the degree of compression applied to the video bit stream under test, and on a number of implementation choices that the manufacturer may have made in its design.

When in doubt on the choice of full reference perceptual measurement equipment among the models available on the market, or before they decide to choose a new model, operators would be well advised to perform a set of tests with the new equipment, checking the correlation of its indications with those obtained by means of subjective assessment tests on an appropriate set of test pictures or sequences.

Objective video quality models – Pathway to future revisions

Finally, as help in selecting the perceptual measurement model that best fits their requirements, operators may consult Appendix I. Appendix I is based on the final report of VQEG in ITU-T Study Group 9 Contribution COM 9-80, June 2000.

Appendix I will be regularly updated to reflect ongoing work in VQEG and elsewhere, as well as the operating experience that participants in the work of ITU-T may gain in the use of perceptual measurement equipment.

As the methods in Appendix I (or others that may be proposed later) improve, are fully disclosed, and gain further validity, they may be adopted as normative sections of this Recommendation. For any model to become normative, it must be verified by an open independent body (such as VQEG) which will do the technical evaluation within the guidelines and performance criteria set out by Study Group (SG) 9. The intention of SG 9 is to eventually recommend only one normative full reference method for cable television.

APPENDIX I

Full reference perceptual video quality measurement models

I.1 Model descriptions

Appendices I and IX describe the eight models that were validated by VQEG and documented in the VQEG final report dated March 2000. Brief details of these, together with a description of PSNR, are included below.

I.1.1 PSNR

The PSNR is defined according to the following formulae:

$$PSNR = 10 \log_{10}\left(\frac{255^2}{MSE}\right)$$

$$MSE = \frac{1}{(P2-P1+1)(M2-M1+1)(N2-N1+1)} \sum_{p=P1}^{p=P2} \sum_{m=M1}^{m=M2} \sum_{n=N1}^{n=N2} (d(p,m,n)-o(p,m,n))^2$$

Where *d(p,m,n)* and *o(p,m,n)* represent respectively degraded and original pixel value at frame p, row m and column n.

NOTE – PSNR requires a very high degree of normalization to be used with confidence. The normalization requires both spatial and temporal alignment as well as corrections for gain and offset. The normalization method is the subject of another Recommendation (under study).

I.1.2 CPqD

The CPqD's model presented to VQEG tests has temporarily been named CPqD-IES (Image Evaluation based on Segmentation) version 2.0. The first version of this objective quality evaluation system, CPqD-IES v.1.0, was a system designed to provide quality prediction over a set of predefined scenes.

CPqD-IES v.1.0 implements video quality assessment using objective parameters based on image segmentation. Natural scenes are segmented into plane, edge and texture regions, and a set of objective parameters is assigned to each of these contexts. A perceptual-based model that predicts subjective ratings is defined by computing the relationship between objective measures and results of subjective assessment tests, applied to a set of natural scenes processed by video processing systems. In this model, the relationship between each objective parameter and the subjective impairment level is approximated by a logistic curve, resulting an estimated impairment level for each parameter. The final result is achieved through a combination of estimated impairment levels, based on their statistical reliabilities.

A scene classifier was added to the CPqD-IES v.2.0 in order to get a scene independent evaluation system. Such classifier uses spatial information (based on DCT analysis) and temporal information (based on segmentation changes) of the input sequence to obtain model parameters from a twelve-scene (525/60 Hz) database (Appendix II).

I.1.3 Tektronix/Sarnoff

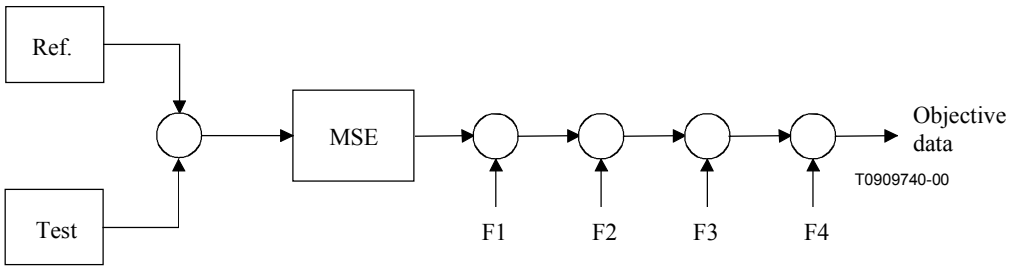
The Tektronix/Sarnoff submission is based on a visual discrimination model that simulates the responses of human spatio-temporal visual mechanisms and the perceptual magnitudes of differences in mechanism outputs between source and processed sequences. From these differences, an overall metric of the discriminability of the two sequences is calculated. The model was designed under the constraint of high-speed operation in standard image processing hardware and thus represents a relatively straightforward, easy-to-compute solution (Appendix III).

I.1.4 NHK/Mitsubishi Electric Corp.

The model emulates human-visual characteristics using 3D (spatio-temporal) filters, which are applied to differences between source and processed signals. The filter characteristics are varied based on the luminance level. The output quality score is calculated as a sum of weighted measures from the filters. The hardware version now available, can measure picture quality in real-time and will be used in various broadcast environments such as real-time monitoring of broadcast signals (Appendix IV).

I.1.5 KDD

See Figure I.1.



F1 Pixel based spatial filtering
F2 Block based filtering (noise masking effect)
F3 Frame based filtering (gaze point dispersion)
F4 Sequence based filtering (motion vector + object segmentation, etc.)

Figure I.1/J.144 – Model description

MSE is calculated by subtracting the Test signal from the Reference signal (Ref). And MSE is weighted by Human Visual Filter F1, F2, F3 and F4.

Submitted model is F1+F2+F4 (Version 2.0, August 1998) (Appendix V).

I.1.6 EPFL

The perceptual distortion metric (PDM) submitted by EPFL is based on a spatio-temporal model of the human visual system. It consists of four stages, through which both the reference and the processed sequences pass. The first converts the input to an opponent-colours space. The second stage implements a spatio-temporal perceptual decomposition into separate visual channels of different temporal frequency, spatial frequency and orientation. The third stage models effects of pattern masking by simulating excitatory and inhibitory mechanisms according to a model of contrast gain control. The fourth and final stage of the metric serves as pooling and detection stage and computes a distortion measure from the difference between the sensor outputs of the reference and the processed sequence (Appendix VI).

I.1.7 NASA

The model proposed by NASA is called DVQ (Digital Video Quality) and is version 1.08b. This metric is an attempt to incorporate many aspects of human visual sensitivity in a simple image processing algorithm. Simplicity is an important goal, since one would like the metric to run in real time and require only modest computational resources. One of the most complex and time-consuming elements of other proposed metrics are the spatial filtering operations employed to implement the multiple, bandpass spatial filters that are characteristic of human vision. We accelerate this step by using the Discrete Cosine Transform (DCT) for this decomposition into spatial channels. This provides a powerful advantage since efficient hardware and software are available for this transformation, and because in many applications the transform may have already been done as part of the compression process.

The input to the metric is a pair of colour image sequences: reference and test. The first step consists of various sampling, cropping, and colour transformations that serve to restrict processing to a region of interest and to express the sequences in a perceptual colour space. This stage also deals with de-interlacing and de-gamma-correcting the input video. The sequences are then subjected to a blocking and a Discrete Cosine Transform, and the results are then transformed to local contrast. The next steps are temporal and spatial filtering, and a contrast masking operation. Finally the masked differences are pooled over spatial temporal and chromatic dimensions to compute a quality measure (Appendix VII).

I.1.8 KPN/Swisscom CT

The Perceptual Video Quality Measure (PVQM) as developed by KPN/Swisscom CT uses the same approach in measuring video quality as the Perceptual Speech Quality Measure (PSQM [1], ITU-T P.861 [2]) in measuring speech quality. The method was designed to cope with spatial, temporal distortions, and spatio-temporally localized distortions like those found in error conditions. It uses ITU-R BT.601.5 [3] input format video sequences (input and output) and resamples them to 4:4:4, Y, Cb, Cr format. A spatio-temporal-luminance alignment is included in the algorithm. Because global changes in the brightness and contrast only have a limited impact on the subjectively perceived quality, PVQM uses a special brightness/contrast adaptation of the distorted video sequence. The spatio-temporal alignment procedure is carried out by a kind of block matching procedure. The spatial luminance analysis part is based on edge detection of the Y signal, while the temporal part is based on different frames analysis of the Y signal. It is well known that the Human Visual System (HVS) is much more sensitive to the sharpness of the luminance component than that of the chrominance components. Furthermore, the HVS has a contrast sensitivity function that decreases at high spatial frequencies. These basics of the HVS are reflected in the first pass of the PVQM algorithm that provides a first order approximation to the contrast sensitivity functions of the luminance and chrominance signals. In the second step the edginess of the luminance Y is computed as a signal representation that contains the most important aspects of the picture. This edginess is computed by calculating the local gradient of the luminance signal (using a Sobel like spatial filtering) in each frame and then averaging this edginess over space and time. In the third step, the chrominance error is computed as a weighted average over the colour error of both the Cb and Cr components with a dominance of the Cr component. In the last step the three different indicators are mapped onto a single quality indicator, using a simple multiple linear regression, which correlates well the subjectively perceived overall video quality of the sequence (Appendix VIII).

I.1.9 NTIA

This video quality model uses reduced bandwidth features that are extracted from spatial-temporal (S-T) regions of processed input and output video scenes. These features characterize spatial detail, motion, and colour present in the video sequence. Spatial features characterize the activity of image edges, or spatial gradients. Digital video systems can add edges (e.g. edge noise, blocking) or reduce edges (e.g. blurring). Temporal features characterize the activity of temporal differences, or temporal gradients between successive frames. Digital video systems can add motion (e.g. error blocks) or reduce motion (e.g. frame repeats). Chrominance features characterize the activity of colour information. Digital video systems can add colour information (e.g. cross colour) or reduce colour information (e.g. colour sub-sampling). Gain and loss parameters are computed by comparing two parallel streams of feature samples, one from the input and the other from the output. Gain and loss parameters are examined separately for each pair of feature streams since they measure fundamentally different aspects of quality perception. The feature comparison functions used to calculate gain and loss attempt to emulate the perceptibility of impairments by modelling perceptibility thresholds, visual masking, and error pooling. A linear combination of the parameters is used to estimate the subjective quality rating (Appendix IX).

I.2 References

[1] BEERENDS (J.G.), STEMERDINK (J.A.): A perceptual speech quality measure based on a psychoacoustic sound representation, *J. Audio Eng. Soc*. 42, 115-123, 1994.

[2] ITU-T P.861 (1998), *Objective quality measurement of telephone-band (300-3400 Hz) speech codecs*.

[3] ITU-R BT.601-5 (1995), *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*.

APPENDIX II

CPqD

Video quality assessment using objective parameters based on image segmentation

Abstract

This appendix presents a methodology for video quality assessment using objective parameters based on image segmentation. Natural scenes are segmented into plane, edge and texture regions, and a set of objective parameters are assigned to each of these contexts. A perceptual-based model that predicts subjective ratings is defined by computing the relationship between objective measures and results of subjective assessment tests, applied to a set of natural scenes and MPEG-2 video codecs. In this model, the relationship between each objective parameter and the subjective impairment level is approximated by a logistic curve, resulting in an estimated impairment level for each parameter. The final result is achieved through a linear combination of estimated impairment levels, where the weight of each impairment level is proportional to its statistical reliability. The results presented in this appendix show that the use of region-based objective measurements provides more accurate predictions compared to predictions based on global parameters.

II.1 Introduction

Video quality assessment has become a crucial issue with the increasing use of digital video compression systems and the subsequent video services, such as primary and secondary distribution of digital TV, video-on-demand, videophone, videoconference, etc. Due to the flexibility of video coding standards, competing codecs do not provide the same picture quality. Therefore, methods for video quality assessment represent important tools to compare the video quality of competing codecs and to quantify their performance in a large number of applications.

The challenge in developing techniques to estimate the quality of video compression systems stems, in part, from the fact that compression algorithms introduce video impairments which are strongly dependent on the levels of details and motion in the scenes. Moreover, the visual perception of video impairments also depends on the details and motion of the scenes. Thus traditional evaluation methods, which are based on static test signals, are inadequate to quantify the performance of video compression systems.

This appendix presents a methodology for video quality assessment, when the video is processed by unidirectional transmission systems that use digital interfaces and, ideally, digital transport facilities. The method has been applied to assess video compression systems according to MPEG standard [1] and [2] but it could also be used to evaluate other types of systems, such as video codecs based on other analysis techniques (i.e. wavelets and prediction filters) and composite signal encoders/decoders.

Figure II.1 shows the configuration of the objective parameters computation process used for video quality estimation. The file format of the input and the output digital video signals is YCbCr4:2:2, as determined by ITU-R BT.601-5 [3].

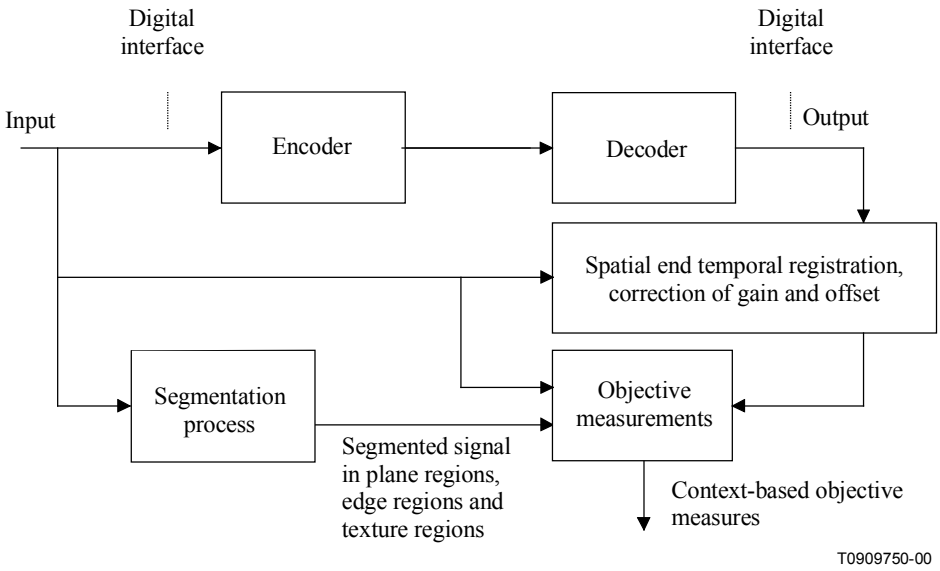


Figure II.1/J.144 – Objective parameters computation

In Figure II.1, each objective parameter is computed separately within the following contexts of the scenes: plane regions, edge regions and texture regions. This is one of the most important aspects of this methodology. A blocking distortion, for example, can be measured by an edge detector applied to the plane regions of the video scene, wherein the visual perception of this distortion is more noticeable. The computational complexity of the method is reduced by using low-complexity estimators and by constraining their computation within the correspondent contexts of the scenes. These contexts are defined by an image segmentation algorithm that is applied to the original natural scenes (i.e. the input test signal). This type of algorithm normally requires high computational complexity; however, it is executed only once. Note that the spatial and temporal registration between the input and output video signals and the correction of gain and offset are also required to compute the objective parameters correctly. The information about registration (or alignment) and calibration is addressed in [4].

The objective parameters are computed by direct comparison between original and impaired scenes. All estimators are applied to fields rather than frames of video in order to ensure the statistical reliability of the measures in scenes with a high level of motion.

A perceptual-based model that predicts subjective ratings is defined by computing the relationship between objective measures and results of subjective assessment tests applied to a set of natural scenes and MPEG-2 video codecs. These scene-dependent perceptual models are defined in two steps as follows:

1) The relationship between each parameter and the subjective impairment level is approximated by a logistic curve, resulting in an estimated impairment level for each parameter.

2) The final result is achieved by linearly combining the estimated impairment levels, where the weight of each impairment level is proportional to its statistical reliability.

The details of the just outlined methodology are presented in the next clauses. Clause II.2 gives a short description of the adopted set-up for the subjective evaluation tests. The methods to determine the objective parameters and to segment the natural scenes are described in clause II.3. Clause II.4 introduces the perceptual models for subjective quality estimation and reports the results that were obtained in this study. Clause II.4 also points out the advantages of using objective parameters based on image segmentation for subjective quality estimation and the dependence of the perceptual

models with the assessors category and with the viewing distance from the monitor (4H or 6H). The conclusions of this contribution are presented in clause II.5.

II.2 Subjective assessment tests

The image processing laboratory of the CPqD/TELEBRÁS (Brazilian Research and Development Center for Telecommunications) has a special room for subjective evaluation trials, according to ITU-R BT.500-7 [5]. This room was used to evaluate the performance of some manufactured and simulated MPEG-2 video codecs on a subset of the natural scenes suggested by ITU-R BT.802-1 [6]. The manufactured MPEG-2 codecs were provided by TV Globo (a Brazilian TV company). The scenes were also processed by the MPEG-2 coding software available at CPqD/TELEBRÁS.

A short description of the set-up utilized for subjective evaluation of the aforementioned MPEG-2 video codecs is given as follows.

II.2.1 Sessions of subjective evaluation

Table II.1 presents a summary of the conditions used for the subjective assessment tests.

Table II.1/J.144 – Conditions of the subjective assessment tests

Conditions for evaluation	According to item 2.1 of ITU-R BT.500-7 [5]
Source of signals	D1 VTR
Monitor	20" studio monitor with digital interface
Viewing distances	4H and 6H
Assessment method	DSIS (Double-Stimulus Impairment Scale) method with nine points in the interval from 1 to 5 [5]
Test sequences	5 scenes of conventional definition digital TV (see II.2.2)
Presentation duration	10 seconds (original signal) + 3 seconds (gray signal) + 10 seconds (signal under evaluation) + 5 seconds for vote, as suggested by Figure 3.a of ITU-R BT.500-7 [5]
Assessors	14 experts and 34 non-experts
Assessors per session	5
Sessions per assessor	2
Presentations per session	48
Assessed items	See II.2.3
Presentation of results	Mean and standard deviation of the impairment level regarding to the reference signal (original scene) Discarding of scores and assessors as suggested by ITU-R BT.500-7 [5]

II.2.2 Natural scenes

The subjective evaluation sessions utilized a set of five natural scenes (see Table II.2), which are defined as test sequences for conventional TV in ITU-R BT.802-1 [6]:

Table II.2/J.144 – Natural scenes used for subjective evaluation

Sequence name	Scene number in ITU-R 802-1
Flower Garden	15
Mobile and Calendar	30
Table Tennis	29
Diva with Noise	17
Kiel Harbour-4	26

II.2.3 Systems under test

In total, 26 systems were included in the sessions of subjective evaluation. These items are presented in Table II.3.

Table II.3/J.144 – Systems under test

Group	Type	Characteristics	Assessed items
1	Manufactured MP@ML MPEG-2 codec for CBR (constant bit rate) applications	Bit rates: 5, 10 and 15 Mbit/s N = 12 and M = 2	6
2	Simulated MP@ML MPEG-2 codec for CBR applications	Bit rates: 2.5, 5, 7.5, 10, 12 and 15 Mbit/s N = 12 and M = 1 and 2	12
3	Manufactured 422P@MPL MPEG-2 codec for CBR applications	Bit rate: 18 Mbit/s N = 2 and M = 2	1
4	Simulated MP@ML MPEG-2 codec for VBR (variable bit rate) applications using intra-frame coding only	Fix quantizer scale [2] in 4, 8, 16, 32 and 62	5
5	Composite signal conversion	NTSC and PAL-M	2

II.3 Objective measurements based on context

This clause describes the video material used for the objective evaluation (i.e. the material used for objective parameters computation – See II.3.1), suggests three image segmentation methods that can be used to divide the video material into plane, edge and texture regions (see II.3.2), and presents the objective parameters that have been adopted in this study (see II.3.3).

II.3.1 Video material used for objective evaluation

The video material used for objective evaluation consists of a 17-second long video sequence, composed of 10 clips of natural scenes and 2 clips of artificial test signals.

Five clips of natural scenes, 2 seconds long each, were selected from the natural scenes presented in II.2.2. The purpose of using 2-second clips instead of 10-second clips, as in the subjective tests described in clause II.2, was to reduce the computational complexity of the objective evaluation process. The choice of 2-second clips was based on the following criteria:

- The clip of 2 seconds of a given scene represents a critical segment of its 10 seconds material compared to the mean criticality of the scene. This criticality was defined as the number of bits per frame resulting from the coding process of a MP@ML MPEG-2 codec (N = 12 and M = 2) with variable bit rate and quantizer scale equals 16.

- The clip of 2 seconds of this scene also represents a critical segment from the subjective point of view, when the scene is processed by a MP@ML MPEG-2 codec (N = 12 and M = 2) at 5 Mbit/s.

The five other clips, 1 second long each, consist of scenes with low or no motion. These scenes have been used in the objective evaluation process, intercalating the previous 2 seconds long clips, in order to test the adaptive behaviour of the MPEG-2 video codecs (i.e. the behaviour regarding to rate and quality control, performance in regime and scene transition). They are also specified in ITU-R BT.802-1 [6]. Although this is not part of the scope of this contribution, it is important to say that the determination of the performance variation (dispersion of the signal-noise ratio) after each scene transition and in regime (difference of performance over I, P and B frames) has been used to characterize the dynamic behaviour of manufactured MPEG-2 video codecs.

The artificial test signals are (1 second long each):

- Narrow-band noise [4] – static and trichromatic video signal defined by noise with resolution about 1/25 of the Nyquist's limit and with approximately uniform histogram for each of components Y, Cb and Cr.
- Circular zone-plate [4] – static and trichromatic video signal defined by a sinusoidal pattern for the components Y, Cb and Cr, with constant horizontal and vertical frequencies along the same column and along the same line of a given field of video, respectively, and outward crescent frequencies from the centre of the image.

These artificial signals have been used to determine the following parameters:

- Displacement of active video.
- Active video area.
- Gain and offset.
- 2D frequency response.
- Displacement between chrominance and luminance (a vertical displacement between these components has been noticed very often in manufactured MP@ML MPEG-2 systems, due to the conversions YCbCr4:2:0 ⇔ YCbCr4:2:2, creating a halo of spurious chromaticity on the edges of the output signal).

The contiguous test material of 17 seconds is described in Table II.4:

Table II.4/J.144 – Test material for objective evaluation

Time Code (mm:ss:ff)	Scene	Short name	Temporal characteristic	Duration (seconds)
00:00:00	Narrow-Band Noise	Noise	static	1
00:01:00	Flower Garden	Garden	dynamic	2
00:03:00	Tree	Tree	static	1
00:04:00	Mobile and Calendar	Mobile	dynamic	2
00:06:00	Clown	Clown	static	1
00:07:00	Table Tennis	Tennis	dynamic	2
00:09:00	Balls of Wool	Balls	dynamic	1
00:10:00	Diva with Noise	Diva	dynamic	2
00:12:00	Boy with Toys	Boy	static	1
00:13:00	Kiel Harbour-4	Kiel	dynamic	2
00:15:00	Young Couple	Couple	static	1
00:16:00	Circular Zone-Plate	Zone Plate	static	1

II.3.2 Spatial segmentation

There were developed three algorithms for image segmentation [7]. The first is an image segmentation algorithm based on edge detection using recursive filtering (see II.3.2.1), the second is a fuzzy image segmentation algorithm based on spatial features (see II.3.2.2) and the third is an image segmentation algorithm based on watershed (see II.3.2.3). The results of the objective evaluation using these algorithms are discussed in II.4.3. The strategy in these segmentation algorithms is to classify the luminance component of each field of video into three mutually exclusive contexts: plane regions, edge regions and texture regions. These algorithms are shortly described as follows:

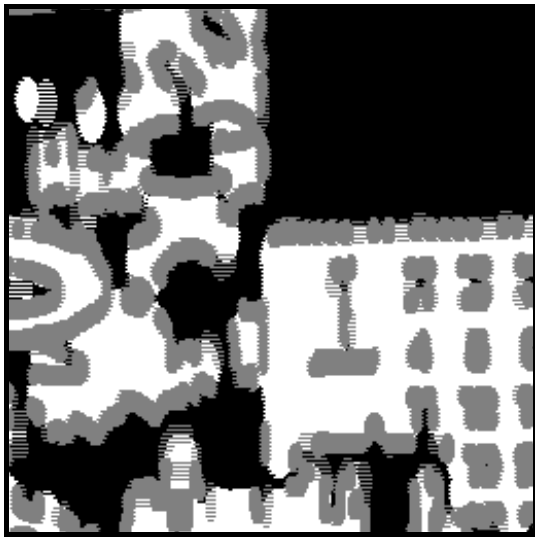
II.3.2.1 Algorithm I: Image segmentation based on edge detection using recursive filtering

This algorithm initially classifies each pixel, based on the brightness variance computed within a neighbourhood of the pixel, as belonging or not belonging to the plane regions of the image. The resulting binary image is then smoothed by a median filter [7]. The algorithm also applies to the original image an edge detector based on recursive filtering. The edges on the boundary of the plane regions are classified as belonging to the edge regions. The texture regions are the remaining regions of the image.

As an example, Figure II.2 shows part of the scene Mobile and Calendar. The result of segmentation by Algorithm I of this part can be seen in Figure II.3. Note that the plane regions are represented by white pixels, edge regions by gray pixels and texture regions by black pixels.

T0909760-00

Figure II.2/J.144 – Part of mobile and calendar



T0909770-00

Figure II.3/J.144 – Result of segmentation

II.3.2.2 Algorithm II: Fuzzy image segmentation based on spatial features

This algorithm is divided into two steps. In the first step, the algorithm assigns a membership function, defined in the interval [0, 1], to each one of the three contexts under classification. In the membership function of the plane regions, the membership value of a pixel is defined inversely proportional to the brightness variance computed within a neighbourhood of the pixel. The morphological gradient [8] applied to this function defines the membership function of the edge regions. The complement of the fuzzy union [9] between these two membership functions defines the membership function of the texture regions. In the second step, each pixel is classified as belonging to the context with highest value of membership among its three membership values determined in the previous step.

II.3.2.3 Algorithm III: Image segmentation algorithm based on watershed

This algorithm first simplifies the luminance component by increasing its homogeneous regions through the application of an edge-preserving smoothing filter [10]. Secondly, a watershed algorithm is applied to the morphological gradient of the simplified image. The watershed detects homogeneous regions, denoted catching basins, with specified minimum relative contrast. The plane regions are the catching basins with area greater than a threshold. The texture regions are given by the erosion of the complement of the plane regions. The edge regions are the remaining regions of this process.

II.3.3 Objective parameters

The objective parameters are obtained for each context (plane, edge and texture) and from the samples of luminance and chrominance of the input (Y_{ref}, Cb_{ref} and Cr_{ref}) and output (Y_{dec}, Cb_{dec} and Cr_{dec}) signals, after spatial and temporal registration and correction of gain and offset, as shown in Figure II.1. The measures and the underlying process to compute them are described as follows:

- MSE (Mean Square Error).
- PSD (Positive Sobel Difference).
- NSD (Negative Sobel Difference).
- ASD (Absolute Sobel Difference).

Let X(i,j) be the j-th pixel of the i-th line of the input signal, Z(i,j) be the j-th pixel of the i-th line of the output signal and the elements $X_m(i,j)$ and $Z_m(i,j)$ be the pixels of the input and output signals, respectively, after a median filtering.

The computation of MSE in a context R (plane, edge or texture) is defined by the mean value of:

$$SE(i,j) = [\ X(i,j) - Z(i,j)\ ]^2, \text{ where } (i,j) \in R.$$

The computation of PSD in a context R (plane, edge or texture) is defined by the mean value of:

$$PS(i,j) = \max [\ sobel(X_m(i,j)) - sobel(Z_m(i,j))\ ,\ 0\ ], \text{ where } (i,j) \in R.$$

The computation of NSD in a context R (plane, edge or texture) is defined by the mean value of:

$$NS(i,j) = -\max [\ sobel(Z_m(i,j)) - sobel(X_m(i,j))\ ,\ 0\ ], \text{ where } (i,j) \in R.$$

The computation of ASD in a context R (plane, edge or texture) is defined by the mean value of:

$$AS(i,j) = |\ sobel(X_m(i,j)) - sobel(Z_m(i,j))\ |, \text{ where } (i,j) \in R.$$

In other words, ASD = PSD + NSD.

The objective parameters utilized for subjective quality estimation refer to the mean value of the aforementioned measures computed over a set of 2N fields of the final portion (i.e. in regime condition) of each one of the five clips indicated in Table II.4 and that belong to the scenes submitted to subjective evaluation. The value N is a multiple of the interval between intra-frame coded images (type I), that is, it is a multiple of the GOP length [1, 2]. In order to satisfy this condition to all systems defined in Table II.3, it is chosen N = 12.

II.4 Subjective quality estimation

This clause describes how the subjective quality estimation models are defined for each scene. Clause II.4.1 describes a perceptual model to estimate the subjective impairment level based on a single parameter. The results of this approximation for each objective parameter are linearly combined to estimate the final subjective impairment level. This linear prediction model is presented in II.4.2. Clause II.4.3 presents and discusses the results of this study.

II.4.1 Subjective quality estimation based on a single parameter: Logistic approximation

For each scene, the relationship between each objective parameter D and the subjective result U is initially defined as follows.

A normalized impairment level between 0% and 100% is defined by [5] as:

$$d = (U_{\max} - U)/(U_{\max} - U_{\min}) \times 100\%$$

The relationship between d and each objective parameter D is approximated by the following logistic function [5]:

$$\underline{d} = \frac{1}{1 + \left(\frac{D_M}{D}\right)^G} \times 100\%$$

where the values D_M and G are computed in order to minimize the mean square error:

$$e = E\left[\{d - \underline{d}\}^2\right]$$

for each scene and each objective parameter separately. The statistical reliability of $\underline{d}$ is defined as 1/e.

II.4.2 Subjective quality estimation: Linear prediction in three steps

The estimation of the normalized impairment level d by a set of estimated impairment levels $\underline{d}$ (one per parameter as defined in II.4.1) is implemented in three steps of linear prediction as described below.

Step 1

First consider the following sets of estimated impairment levels selected for the luminance component:

- $\underline{d}^{MSE}$
- $\underline{d}^{PSD}$ and $\underline{d}^{NSD}$
- $\underline{d}^{ASD}$
- $\underline{d}^{MSE}$, $\underline{d}^{PSD}$ and $\underline{d}^{NSD}$
- $\underline{d}^{MSE}$ and $\underline{d}^{ASD}$

For a given scene and context of this scene (plane, edge or texture), the best set is the one with the least prediction error. By using this criterion to choose a set of estimated impairment levels for each context, this step linearly combines the impairment levels of each selected set and outputs three estimation values (one per context) denoted by: $\underline{d}_{YP}$, $\underline{d}_{YE}$ and $\underline{d}_{YT}$.

Similarly, the considered sets of estimated impairment levels for the chrominance components of the scene are:

- $\underline{d}^{MSE(Cb)}$ and $\underline{d}^{MSE(Cr)}$
- $\underline{d}^{ASD(Cb)}$ and $\underline{d}^{ASD(Cr)}$

and the three resulting estimation values (one per context) are denoted by: $\underline{d}_{CP}$, $\underline{d}_{CE}$ e $\underline{d}_{CT}$.

Step 2

The estimation values $\underline{d}_P$, $\underline{d}_E$ and $\underline{d}_T$ result from linear prediction based on the vectors ($\underline{d}_{YP}$, $\underline{d}_{CP}$), ($\underline{d}_{YE}$, $\underline{d}_{CE}$) and ($\underline{d}_{YT}$, $\underline{d}_{CT}$), respectively.

Step 3

The estimation values $\underline{d}_P$, $\underline{d}_E$ and $\underline{d}_T$ are combined by linear prediction to produce the estimated impairment level $\underline{d}$.

In all steps above, the predictors satisfy the following restrictions.

Let ($\underline{d}_1, \underline{d}_2, ..., \underline{d}_P$) be the input vector of the linear predictor. The output $\underline{d}_o$ is given by:

$$\underline{d}_o = \sum a_i \underline{d}_i$$

where the weights $\{a_i\}$ are computed in order to minimize the mean square error:

$$E\left[\{d - \underline{d}_o\}^2\right], \text{ such that}$$

$$\sum a_i = 1 \text{ and}$$

$$a_i / a_k = e_k / e_i$$

where the statistical reliability of $\underline{d}_i$ is $1/e_i$, as defined in II.4.1.

It has been observed that this type of prediction is more robust than the one obtained by optimum predictors, because it is less dependent on the training database. It reaches better results when applied to test databases, as exemplified in II.4.3.

II.4.3 Subjective quality estimation: Presentation and discussion of results

This clause is divided into three main topics. The results and prediction models obtained by the subjective quality estimation based on Algorithm I (the image segmentation algorithm previously described in II.3.2.1) are described in II.4.3.1. Clause II.4.3.1 also presents the dependence between the perceptual models and the assessors category (experts and non-experts) and between the perceptual models and the viewing distance from the monitor (4H and 6H). The variation of the estimation accuracy with the image segmentation algorithms is discussed in II.4.3.2. Clause II.4.3.3 points out the advantages of the proposed subjective estimation method compared to other methods that are based on global measurements or optimal prediction.

II.4.3.1 Results: Perceptual models and performance

Table II.5 presents the results of the subjective estimation method based on Algorithm I (see II.3.2.1) for segmenting the following scenes: Garden, Mobile, Tennis, Diva and Kiel, separately. In Table II.5:

- The weights of the linear prediction described in Step 2 of II.4.2 are equivalent to the relative subjective weights of luminance (Y) and chrominance (C) impairments in plane regions, edge regions and texture regions. The global mean value computed over all scenes is given at the last line of this table.
- The weights of the linear prediction described in Step 3 of II.4.2 are equivalent to the relative subjective weights of the degradation in plane regions (P), edge regions (E) and texture regions (T). The global mean value computed over all scenes is given at the last line of this table.

- The mean square error (MSE) and the mean absolute error (MAE) between the normalized impairment level d and the estimated impairment level $\underline{d}$, taking into account a normalization scale from 0% to 100%, are shown at the two last columns of this table. The error between the mean normalized impairment level and the mean estimated impairment level, computed over all scenes, is shown at the last line of these columns.

The results presented in Table II.5 refer to the perceptual models obtained from the subjective scores of the 34 non-expert assessors of Table II.1 and from the 26 assessed systems of Table II.3.

Table II.5/J.144 – Perceptual models and results: Non-expert assessors

	Step 2: Plane		Step 2: Edge		Step 2: Texture		Step 3			Error	
Scene	**Y(%)**	**C(%)**	**Y(%)**	**C(%)**	**Y(%)**	**C(%)**	**P(%)**	**E(%)**	**T(%)**	**MSE**	**MAE**
Garden	61	39	70	30	37	63	13	37	51	18.1	3.0
Mobile	74	26	75	25	63	37	83	7	9	24.2	3.6
Tennis	67	33	65	35	70	30	45	13	42	25.3	3.5
Diva	49	51	92	8	42	58	27	59	14	5.4	1.5
Kiel	62	38	66	34	40	60	32	39	29	22.7	3.6
Global	**63**	**37**	**73**	**27**	**50**	**50**	**40**	**31**	**29**	**6.2**	**1.8**

Tables II.6 and II.7 show the dependence between perceptual models and results for:

- non-expert and expert assessors;
- viewing distance (4H and 6H) from the monitor (each case with 50% of the total number of assessors).

Table II.6/J.144 – Perceptual models and results: Non-expert and expert assessors

	Non-expert assessors						Expert assessors					
	Region			Component		Error	Region			Component		Error
Scene	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**	**P(%)**	**E(%)**	**T(%)**	**Y(%)**	**C(%)**	**MSE**
Garden	13	37	51	52	48	18.1	12	53	34	51	49	23.5
Mobile	83	7	9	73	27	24.2	72	13	15	72	28	73.4
Tennis	45	13	42	68	32	25.3	47	12	41	70	30	48.1
Diva	27	59	14	73	27	5.4	22	42	36	55	45	21.2
Kiel	32	39	29	57	43	22.7	43	36	21	47	53	44.1
Global	**40**	**31**	**29**	**62**	**38**	**6.2**	**39**	**31**	**30**	**58**	**42**	**12.1**

Table II.7/J.144 – Perceptual models and results: 6H and 4H viewing distances

	6H viewing distance						4H viewing distance					
	Region			Component		Error	Region			Component		Error
Scene	P(%)	E(%)	T(%)	Y(%)	C(%)	MSE	P(%)	E(%)	T(%)	Y(%)	C(%)	MSE
Garden	15	40	45	51	49	20.0	9	35	55	47	53	16.8
Mobile	83	8	9	77	23	24.7	71	13	16	62	38	59.4
Tennis	47	15	38	60	40	40.9	42	10	48	74	26	21.4
Diva	41	36	22	64	36	15.5	21	46	33	54	46	12.5
Kiel	34	40	26	54	46	18.3	31	46	23	57	43	26.0
Global	**44**	**28**	**28**	**61**	**39**	**7.9**	**35**	**30**	**35**	**59**	**41**	**9.6**

The results presented in Tables II.5, II.6 and II.7 are commented below:

- The subjective quality estimation using objective parameters based on image segmentation, computed over the 26 systems described in II.2.3, resulted in a mean absolute error (MAE) of less than 4% for each individual scene and a global MAE of 1.8%, considering non-expert assessors.
- Comparing the perceptual models based on the opinion of expert and non-expert assessors, the weight of the impairments in chrominance is slightly greater in the models based on expert assessors.
- Comparing the perceptual models based on 4H and 6H viewing distances, the weight of the impairments in edge and texture regions is significantly greater in the models based on 4H viewing distance, as expected.

II.4.3.2 The variation of the estimation accuracy with the image segmentation algorithm

The results of the subjective quality estimation, based on Algorithms II and III (briefly described in II.3.2) and obtained from the scores of the 34 non-expert assessors, are shown in Table II.8. Comparing the results of this table with the results previously presented on the left side of Table II.6 (for Algorithm I), the estimation accuracy presented small variations for a given scene depending on the image segmentation algorithm. On the other hand, there was no relevant variation in the global estimation accuracy considering the three image segmentation algorithms. This suggests that even simpler image segmentation algorithms may provide satisfactory results.

Table II.8/J.144 – Perceptual models and results: Algorithms II and III

	Algorithm II						Algorithm III					
	Region			Component		Error	Region			Component		Error
Scene	P(%)	E(%)	T(%)	Y(%)	C(%)	MSE	P(%)	E(%)	T(%)	Y(%)	C(%)	MSE
Garden	9	32	59	53	47	18.0	10	44	46	53	47	15.8
Mobile	65	26	9	59	41	20.7	82	11	6	60	40	18.7
Tennis	54	27	19	70	30	28.5	68	21	11	72	28	31.3
Diva	25	50	24	75	25	7.1	50	31	19	60	40	7.4
Kiel	23	31	46	64	36	25.9	28	33	38	59	41	22.4
Global	**35**	**33**	**31**	**66**	**34**	**7.4**	**48**	**28**	**24**	**63**	**37**	**6.5**

II.4.3.3 The advantages of the adopted subjective quality estimation method

The example illustrated in Table II.9 focuses on two very important properties of any method for subjective quality estimation based on objective parameters: accuracy and robustness [11] to [14]. This example compares the adopted subjective quality estimation method, which relies on context-based objective measurements and the linear prediction method described in II.4.2, with the following methods:

- a method that relies on the same context-based objective measurements, but uses optimal prediction;
- a method that uses the linear prediction described in II.4.2, but adopts global measurements.

The example used the systems of Group 2 and the NTSC system of Group 5 as training database, and the systems of Group 1 and the PAL-M system of Group 5 as test database (see Table II.3). The objective parameters used in this example were the ones based on MSE and described in II.3.3. The results were obtained from the scores of the non-expert assessors. The input values of the table are mean square prediction errors. The last line of the table shows the mean value of this parameter computed over the set of scenes.

Table II.9/J.144 – Comparison: Robustness and accuracy

	Adopted method		**Optimum predictor**		**Global measurements**	
Scene	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
Garden	3.9	87.6	2.8	71.8	3.9	62.3
Mobile	30.1	48.6	10.5	82.1	179.1	162.5
Tennis	10.8	91.3	7.7	335.0	108.9	221.2
Diva	1.4	8.9	0.8	17.7	1.8	34.3
Kiel	22.4	9.3	20.5	13.4	30.6	27.7
Mean	**13.7**	**49.1**	**8.5**	**104.0**	**64.9**	**101.6**

The advantage of computing objective parameters based on context becomes clear when the procedure described in II.4.1 and II.4.2 is also applied to global measurements. Note that the use of context-based measurements can significantly improve the estimation results in all scenes (with the exception of Flower Garden). Possibly, this indicates that the image segmentation process for Flower Garden needs to be refined.

The example also shows that the prediction process described in II.4.2 is more robust (i.e. it is less dependent on the training database) when it is compared to the optimum predictor, improving the prediction results on the test database.

II.5 Conclusions

This appendix presents a methodology for subjective quality estimation using objective parameters based on image segmentation. The objective parameters are computed within plane regions, edge regions and texture regions resulting from the image segmentation process.

The results presented in this appendix show that the use of context-based objective parameters compared to global parameters leads to more accurate predictions. This aspect is reinforced by the use of the perceptual model based on the linear prediction method described in II.4.2. This method has led to more robust prediction results when it is compared to the optimal prediction.

These results can be further improved if:

- the temporal information is included in the image segmentation process (e.g. edge regions could be further classified into edge regions with low motion and edge regions with high motion);
- the plane, edge and texture regions of chrominance are also considered in the image segmentation process, since Algorithms I, II and III were used to segment the luminance component only.

Therefore, we suggest the inclusion of the linear prediction method presented in this appendix and of context-based objective measurements in new ITU Recommendations, which are related to objective evaluation of video quality.

II.6 References

[1] ISO/IEC 11172-1:1993, *Information technology – Coding of moving pictures and associated audio for digital storage media up to about 1.5 Mbit/s – Part 1: Systems.*

[2] ITU-T H.262 (2000), *Information technology – Generic coding of moving pictures and associated audio information: Video*.

[3] ITU-R BT.601-5 (1995), *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*.

[4] ANSI T1.801.03 (1996), *Digital transport of one-way video signals – Parameters for objective performance assessment*.

[5] ITU-R BT.500-7 (1995), *Methodology for the subjective assessment of the quality of television pictures*.

[6] ITU-R BT.802-1 (1994), *Test pictures and sequences for subjective assessments of digital codecs conveying signals produced according to ITU-R Recommendation BT.601*.

[7] GONZALEZ, WINTZ (P.), Digital Image Processing, *Addison Wesley*, 1987.

[8] DOUGHERTY: An Introduction to Morphological Image Processing, *SPIE Optical Engineering Press*, Bellingham, WA, Vol. TT9, 1992.

[9] KAUFMANN (A.): Introduction to The Theory of Fuzzy Subsets, *Academic Press*, New York, NY, Vol. 1, 1975.

[10] BARRERA (J.), BANON (J.F.), LOTUFO (R.A.): Mathematical Morphology Toolbox for the Khoros System, *Conference on Image Algebra and Morphological Image Processing, V International Symposium on Optics, Imaging, and Instrumentation, SPIE's Annual Meeting*, San Diego, USA, 24-29 July 1994.

[11] ITU-T Contribution COM 12-66, *Selections from the draft American National Standard – Digital transport of one-way signals – Parameters for objective performance assessment*, USA, January 1996.

[12] ITU-T Study Group 12 Delayed Contribution D021, *Objective and subjective measures of MPEG video quality: summary of experimental results*, USA, April 1997.

[13] ITU-T Study Group 12 Delayed Contribution D101, *A Two-Stage Objective Model for Video Quality Evaluation*, Bellcore, May 1996.

[14] ANSI T1A1 Contribution Number T1A1.5/96-121, *Objective and subjective measures of MPEG video quality*, GTE Labs., NTIA/ITS, October 1996.

APPENDIX III

Tektronix/Sarnoff

Introduction

New digital television services create a demand for monitoring Quality of Service with measurement tools that are very different from their analogue counterparts. A key requirement is that Objective picture quality measurements should correlate closely to subjective quality assessment.

This appendix describes a Human Vision Model based measurement tool that can be used within a digital television system. A practical implementation provides results that demonstrate high correlation with subjective assessments made in accordance with ITU-R BT.500-7.

Specific issues addressed include:

- measurement of PQR objective picture quality rating within digital video transmission networks;
- requirements for pre-processing of video prior to analysis;
- details of the algorithm used for analysis;
- results of tests showing correlation between objective measurements and subjective picture quality assessments.

A solution to measurement of perceptual video quality within digital video systems is described and is already incorporated in a commercial implementation.

III.1 PQR objective picture quality rating in operational environments

It is broadly accepted that Objective picture quality measurements may be made more accurately when knowledge of the reference video is available. In the generic diagram shown in Figure III.1, program video enters the transmission system (Reference Video) and is transmitted through the system and monitored at the output (Processed Video). Analysis of the differences between the Processed and Reference Video with a Human Vision Model provides an accurate measure of the PQR Objective picture quality rating. (Details of the algorithm used within the jointly-developed Sarnoff/Tektronix Human Vision Model to provide PQR Objective picture quality rating are included in clause III.4.)

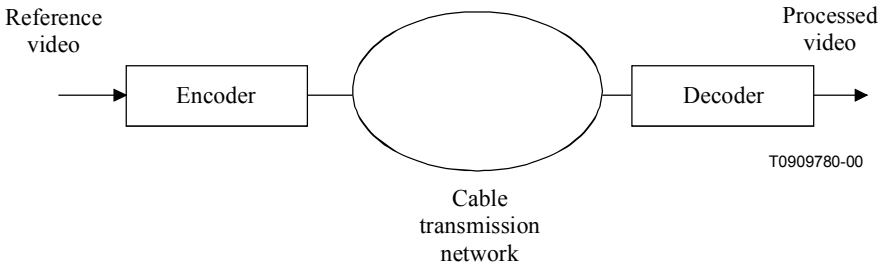


Figure III.1/J.144 – Generic diagram of a compressed video transmission system

An essential prerequisite for the Human Vision Model analysis is normalization of the processed video. The encoding and decoding may cause horizontal and vertical picture shifts and cropping, and may also cause luminance and chrominance gain and level changes. They must be normalized prior to application of the Human Vision Model. (Details of the normalizing process are included in clause III.2.)

Certain transmission systems may require extension of the generic diagram to reflect concatenated codecs and/or use of PAL coding and decoding, but the principles remain the same, and the process of PQR Objective picture quality rating remains valid.

In a laboratory environment, video test sequences may be used in place of live video material. They provide a repeatable source of video and facilitate common measurements between differing laboratories. Inclusion of a wide range of standard programme material within test video sequences ensures optimum validity for live programme material.

In an operational environment, the video test sequences may be replaced by live program material. Reference video and Processed video from a decoder placed at the transmission source as in Figure III.2 provide a measure of the PQR Objective picture quality rating for the operational system.

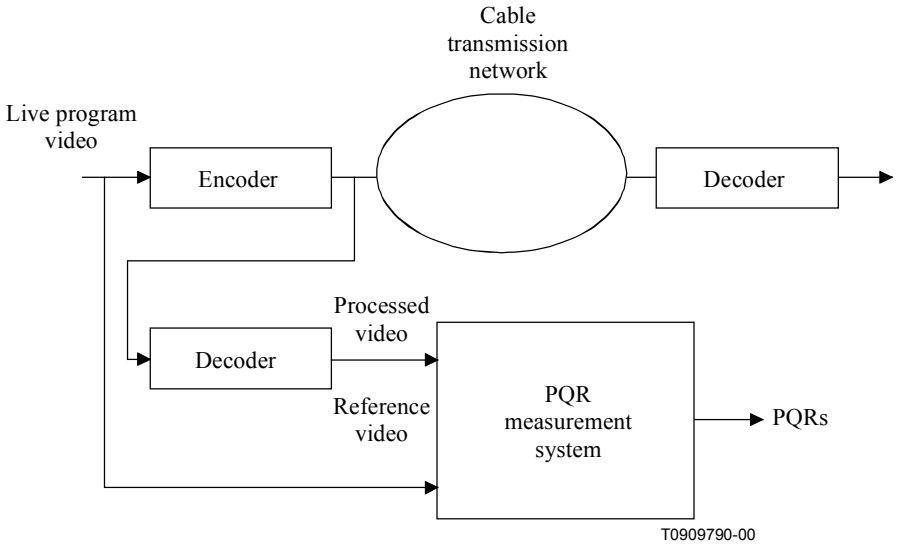


Figure III.2/J.144 – PQR objective picture quality rating with reference available

Lack of access to the reference video may restrict the use of continuous live material. An obvious example is monitoring of satellite feeds at a cable head-end. In this situation, a common video sequence such as a station logo may be chosen as the Reference Video. This could be provided to the Cable Operator and stored locally as a reference for comparison against the processed video. See Figure III.3.

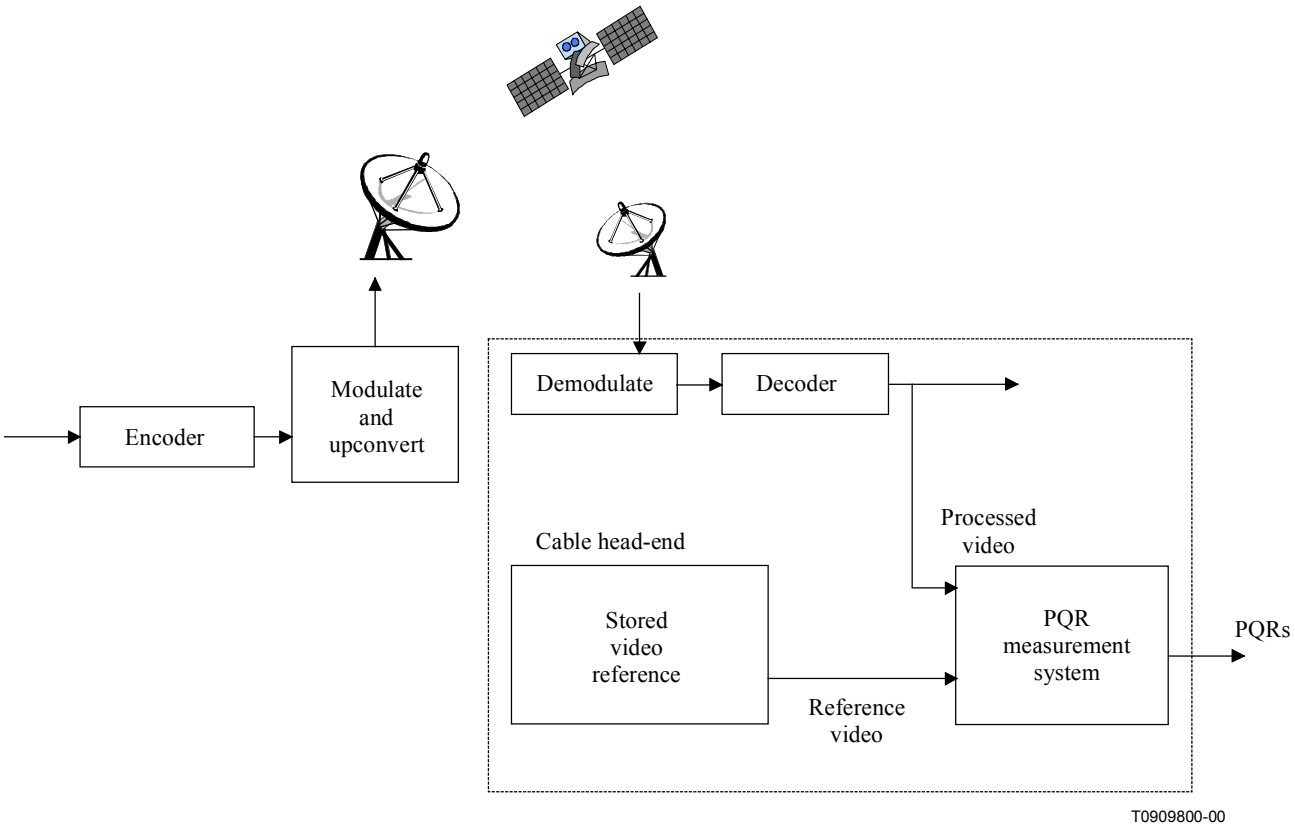


Figure III.3/J.144 – PQR objective picture quality rating at a remote head-end with stored reference

III.2 Pre-processing of video – Normalization

Application of the PQR Objective picture quality rating method to any video system requires normalization of the processed video. Normalization means that time-invariant systematic changes in the video from reference input to processed video output are removed prior to performing the human visual system (HVS) based measurement. As the most sensitive and accurate objective picture quality measurement method, the PQR method is based on HVS filters that compare reference and processed pictures on what is effectively a pixel-by-pixel basis. Separation of the measurement into two parts, normalization and PQR-calculation, is necessary to obtain the most meaningful results.

Parameters to be adjusted by the normalization process are horizontal and vertical picture shifts; luminance and colour gain changes; luminance and colour DC level changes; and component or luminance to colour channel-to-channel delay offset. Because these changes could produce changes in perceived picture quality they shall be reported as part of the results of the measurement method. It is necessary to separate these changes from the PQR calculation for two reasons. The main reason is to provide the most accurate PQR value. Second, such normalization corresponds closely with typical system operation for the gain and DC level parameters where appropriate adjustments are generally available and routinely made. Small values of picture shift, horizontally or vertically, are generally not considered to change perceived picture quality; however, their presence is indeed a picture error and will produce significant problems in multi-generation applications. Time varying changes in the pictures that are due to video content and the compression system are measured by the PQR-calculation.

The idea of normalization prior to making a picture quality assessment is also to be required in subjective measurement standards as reflection of typical system operation. The following statement is to be included in ITU-T P.911: Subjective Audiovisual Quality Assessment Method for Multimedia Applications and ITU-T P.910: Subjective Video Quality Assessment Methods for

Multimedia Applications. "Operational parameters, such as signal level, for the test sequences shall match those of the alignment signals used to verify the viewing [and listening] conditions. Any operational adjustments performed so that source or processed sequences meet this requirement should be reported".

Figure III.4 shows the PQR measurement system operation with respect to normalization. Processed video is normalized on a field-by-field basis by comparison with the reference video or by measurement of calibrated test signals embedded in the reference sequence. Only time-invariant static changes in video are removed, dynamic changes due to the compression and decompression processes are measured as part of the PQR calculation. Normalization of the processed video prior to PQR calculations shall meet the tolerances shown in Table III.1.

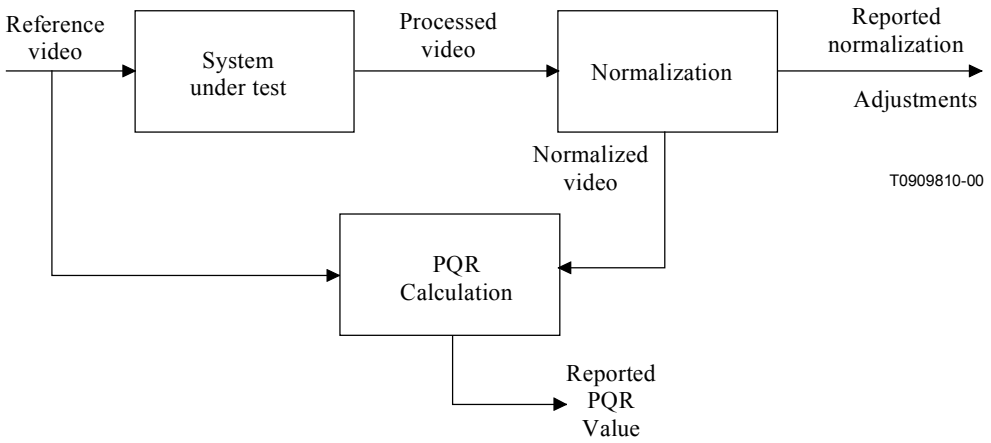


Figure III.4/J.144 – PQR measurement system operation

Table III.1/J.144 – Normalization parameters and tolerance

Parameter	Normalization tolerance
Luminance gain	< 0.2 dB
Colour (difference) gain	< 0.2 dB
Luminance DC level	< 0.5% of signal max
Colour (difference) DC level	< 0.5% of signal max
Channel-to-channel delay offset	< 2 ns
Horizontal pixel shift	< 0.1 pixel
Vertical line shift	< 0.1 line

III.3 System overview

PQR objective picture quality ratings are one of the key results provided by the PQR measurement system referred to earlier. The following provides a description of the Human Vision Model used within the PQR measurement system.

The Sarnoff/Tektronix Human Vision Model is a method of predicting the perceptual ratings that human subjects will assign to a degraded colour-image sequence relative to its non-degraded counterpart. The model takes in two image sequences and produces several difference estimates, including a single metric of perceptual differences between the sequences. These differences are quantified in units of the modelled human just-noticeable difference (JND). A version of the model that applies only to static, achromatic images is described by Lubin (1993, 1995).

The Human Vision Model can be useful in a general context (see Figure III.5). An input video sequence passes through two different channels on the way to a human observer (not shown in the figure). One channel is uncorrupted (the reference channel), and the other distorts the image in some way (the channel under test). The distortion, a side effect of some measure taken for economy, can occur at an encoder prior to transmission, in the transmission channel itself, or in the decoding process. In Figure III.5, the box called "system under test" refers schematically to any of these alternatives. Ordinarily, evaluation of the subjective quality of the test image relative to the reference sequence would involve the human observer and a real display device. This evaluation would be facilitated by replacing the display and observer by the Human Vision Model, which compares the test and reference sequences to produce a sequence of JND maps instead of the subjective comparison.

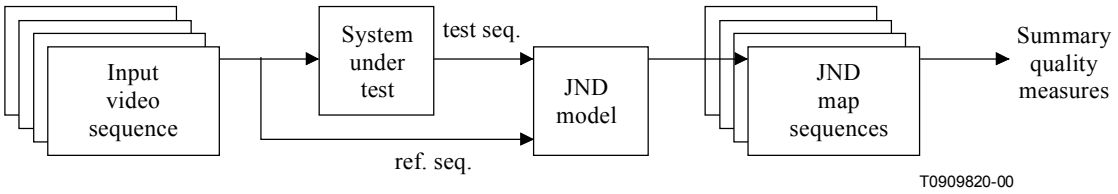


Figure III.5/J.144 – Human vision model in system evaluation

Figure III.6 shows an overview of the algorithm. The inputs are two image sequences of arbitrary length. For each field of each input sequence, there are three data sets, labelled Y', C_b', and C_r' at the top of Figure III.6 derived, e.g. from a D1 tape. Y, C_b, C_r data are then transformed to R', G', and B' electron-gun voltages that give rise to the displayed pixel values. In the model, R', G', B' voltages undergo further processing to transform them to a luminance and two chromatic images that are passed to subsequent stages.

The purpose of the front-end processing is to transform video input signals to light outputs, and then to transform these light outputs to psychophysically defined quantities that separately characterize luma and chroma.

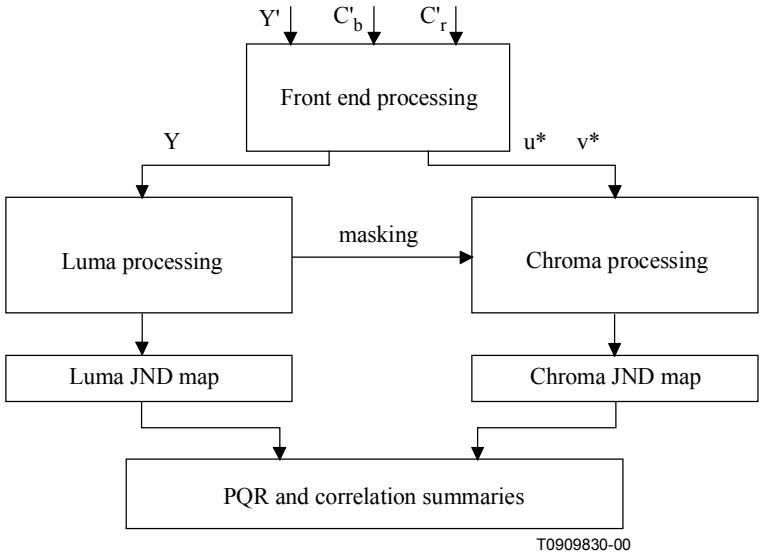


Figure III.6/J.144 – Sarnoff/Tektronix human vision model flow chart

A luma-processing stage accepts two images (test and reference) of luminances Y, expressed as fractions of the maximum luminance of the display. From these inputs, the luma-processing stage generates a luma JND map. This map is an image whose gray levels are proportional to the number of JNDs between the test and reference image at the corresponding pixel location.

Similar processing, based on the CIE L*u*v* uniform-colour space, occurs for each of the chroma images u* and v*. Outputs of u* and v* processing are combined to produce the chroma JND map. Both chroma and luma processing are influenced by inputs from the luma channel called *masking*, which render perceived differences more or less visible depending on the structure of the luma images.

Luma, chroma and combined luma-chroma JND maps are each available as output, together with a small number of summary measures derived from these maps. Single PQR value summaries model an observer's overall rating of distortions in a test sequence. JND maps give a more detailed view of the location and severity of artifacts.

It should be noted that two basic assumptions underlie the model presented here:

a) Each pixel is square and subtends .03 degrees of viewing angle. This number was derived from a screen height of 480 pixels, and a viewing distance of four screen heights (the closest viewing distance prescribed by the ITU-R BT.500). When the model is compared with human perception at longer viewing distances than four screen heights, the model overestimates the human's sensitivity to spatial details. In the absence of hard constraints on viewing distance, the model is chosen to be as sensitive as possible within the frame of ITU-R BT.500.

b) The model applies to screen luminances of .01 to 100 ft-L (for which overall sensitivity was calibrated), but with greatest accuracy at about 20 ft-L (for which all spatio-temporal frequencies were calibrated). It is assumed that changing luminance incurs proportional sensitivity changes at all spatio-temporal frequencies, and this assumption is less important near 20 ft-L, where more calibration took place.

The processing shown in certain of the boxes in Figure III.6 is described in more detail below, keyed to Figures III.7, III.8 and III.9.

III.4 Algorithm overview

III.4.1 Front end processing

The stack of four fields labelled Y', C_b', C_r' at the top of Figure III.7 indicates a set of four consecutive fields from either a test or reference image sequence. The first stage of processing transforms Y', C_b', C_r' data, to R', G', B' gun voltages.

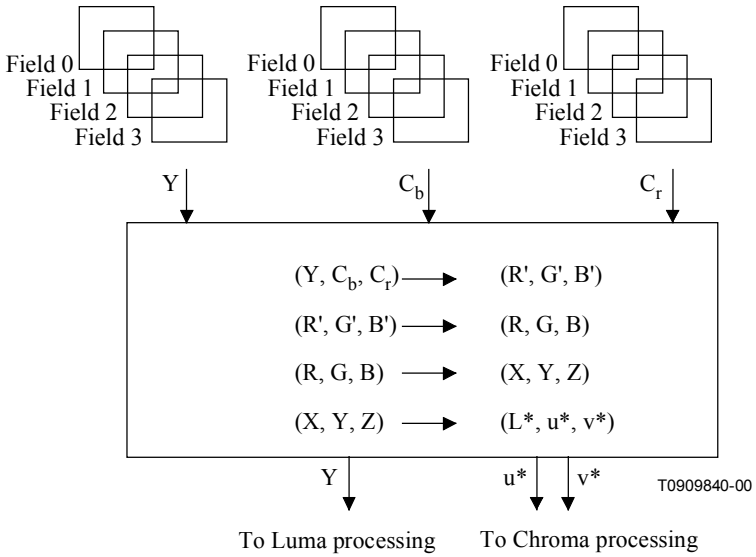


Figure III.7/J.144 – Front end processing

The second stage of processing, applied to each R', G', B' image, is a point-non-linearity. This stage models the transfer from R', G', B' gun voltages to model-intensities (R, G, B) of the display (fractions of maximum luminance). The non-linearity also performs clipping at low luminances in each plane by the display.

Following the non-linearity, one of two processing options is available: half-height and full-height. For interlaced scans, half-height images[1] are processed as given, without blank inter-lines. Full-height modelling is available for progressive scans (in which a field contains one frame, i.e. a single image rather than two interlaced fields).

Then, the vector (R, G, B) at each pixel in the field is subjected to a linear transformation (which depends on the display phosphors) to CIE 1931 tristimulus coordinates (X, Y, Z). The luminance component Y of this vector is passed to luma processing.

To ensure (at each pixel) approximate perceptual uniformity of the colour space to isoluminant colour differences, the individual pixels are mapped into CIELUV, an international-standard uniform-colour space (see Wyszecki and Stiles, 1982). The chroma components u*, v* of this space are passed to the chroma processing steps in the model[2].

III.4.2 Luma processing

As shown in Figure III.8, each luma value is first subjected to a compressive non-linearity. Then, each luma field is filtered and down-sampled in a four-level Gaussian pyramid, in order to model the psychophysically and physiologically observed decomposition of incoming visual signals into different spatial-frequency bands. After this decomposition, the bulk of subsequent processing by the model consists of similar operations (e.g. oriented filtering) performed at each pyramid level.

---

[1] Rows in a half-height image correspond to one field, i.e. to either the even or odd lines of a frame.

[2] The luminance channel L* from CIELUV is not used in luma processing, but instead is replaced by a visual non-linearity for which the vision model has been calibrated over a range of luminance values. L* is used in chroma processing, however, to create a chroma metric that is approximately uniform and familiar to display engineers.

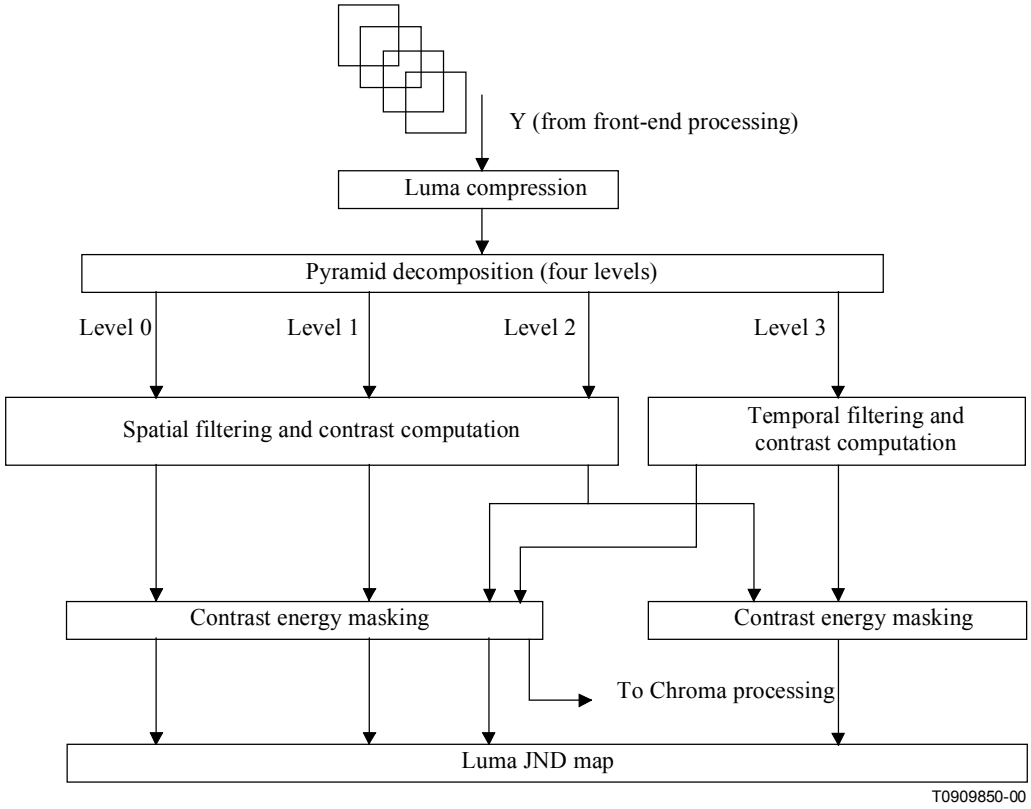


Figure III.8/J.144 – Luma processing overview

After this pyramid-making process, the lowest-resolution pyramid image is subjected to temporal filtering and contrast computation, and the other three levels are subjected to spatial filtering and contrast computation. In each case the contrast is a local difference of pixel values divided by a local sum, appropriately scaled. Initially, this establishes the definition of 1 JND, which is passed on to subsequent stages of the model[3]. (Calibration iteratively revises the 1-JND interpretation at intermediate model stages.) The absolute value of the contrast response is passed to the following stage, and the algebraic sign is preserved for reattachment just prior to image comparison (JND map computation).

The next stage (contrast masking) is a gain-setting operation in which each contrast response is divided by a function of all the contrast responses. This combined attenuation of each response by other local responses is included to model visual "masking" effects such as the decrease in sensitivity to distortions in "busy" image areas. At this stage in the model, temporal structure (flicker) is made to mask spatial differences, and spatial structure is also made to mask temporal differences. Luma masking is also applied on the chroma side, as discussed below.

The masked contrast responses (together with the contrast signs) are used to produce the Luma JND map. This is done by:

- separating each image into positive and negative components (half-wave rectification);
- performing local pooling (averaging and downsampling, to model the local spatial summation observed in psychophysical experiments);
- evaluating the absolute image differences channel by channel;

[3] The association of a constant contrast with 1 JND is an implementation of what is known as Weber's law for vision.

- up-sampling to the same resolution (which will be half the resolution of the original image due to the pooling stage);
- evaluating the Minkowski Q-norm over all channels.

III.4.3 Chroma processing

Chroma processing parallels luma processing in several ways. Intra-image differences of chroma (u* and v*) of the CIELUV space are used to define the detection thresholds for the chroma model, in analogy to the way contrast (and Weber's law) is used to define the detection threshold in the luminance model. Also, in analogy with the luminance model, the chromatic "contrasts" defined by u* and v* differences are subjected to a masking step. A transducer non-linearity makes the discrimination of a contrast increment between one image and another depend on the contrast response that is common to both images.

Figure III.9 shows that, as in luma processing, each chroma component u*, v* is subjected to pyramid decomposition. However, whereas luma processing needs only four pyramid levels, chroma processing is given seven levels. This captures the empirical fact that chromatic channels are sensitive to far lower spatial frequencies than luma channels (Mullen, 1985). Also, it takes into account the intuitive fact that colour differences can be observed in large, uniform regions.

To reflect the inherent insensitivity of the chroma channels to flicker, temporal processing is accomplished by averaging over four image fields.

Then, spatial filtering by a Laplacian kernel is performed in u* and v*. This operation produces a colour difference in u*, v*, which (by definition of the uniform colour space) is metrically connected to just-noticeable colour differences. A value of one at this stage is taken to mean a single JND has been achieved, in analogy to the role of Weber's-law-based contrast in the luma channel. (As in the case of luma, the 1-JND chroma unit must undergo reinterpretation during calibration.)

This colour difference value is weighted, absolute-valued, and passed (with the contrast algebraic sign) to the contrast-masking stage. The masking stage performs the same function as it did in the luma model. It is somewhat simpler, since it receives input only from the luma channels and from the chroma channel whose difference is being evaluated. Finally, the masked contrast responses are processed exactly as in the luma model (see the last paragraph of III.4.2).

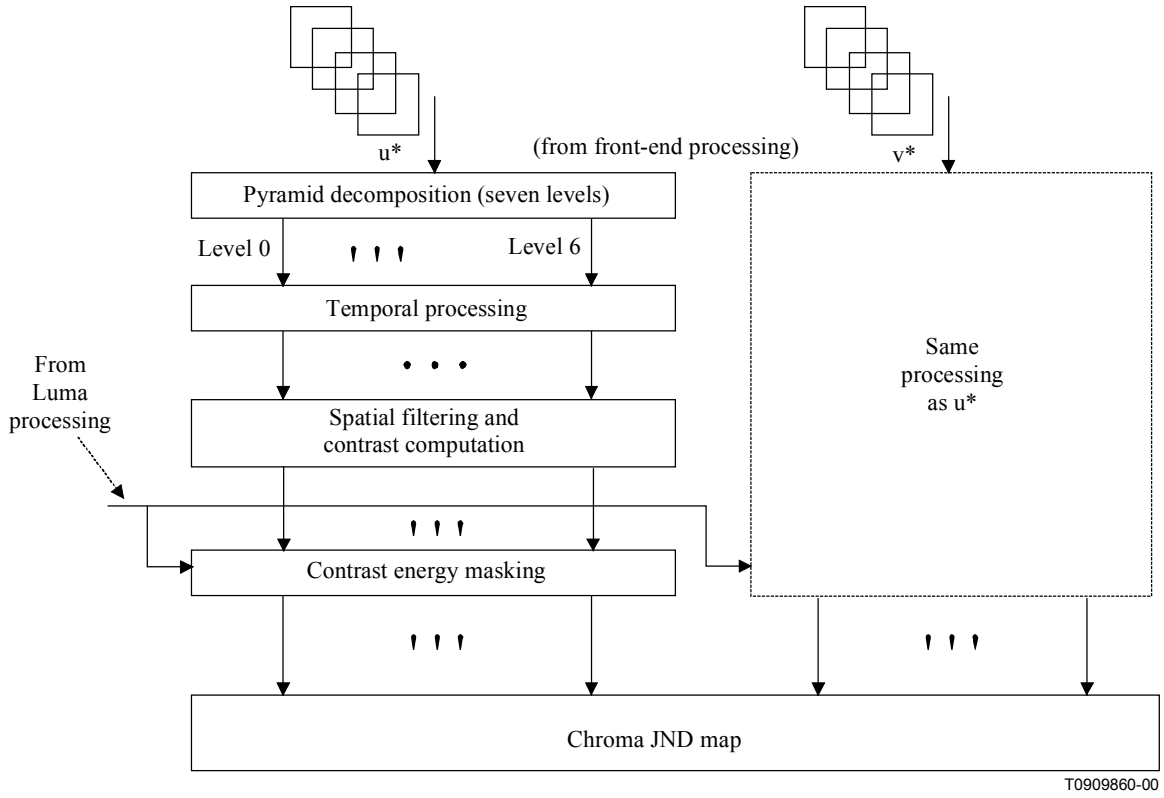


Figure III.9/J.144 – Chroma processing overview

III.4.4 Output summaries

For each field in the video-sequence comparison, the luma and chroma JND maps are first combined to give a total-JND map. This total-JND map is computed as a linear combination of the sum and the maximum of the luma and chroma map values, pixel-by-pixel.

Then, each of the three JND maps (luma, chroma, and combined luma-chroma) is reduced to a single-number summary, called a PQR (Picture Quality Rating) value. Single number summaries are computed by the Minkowski Q-norm. With this approach, each JND-map pixel value is raised to the Qth power. The PQR is then computed as the Qth root of a normalized sum of all Qth power pixel values.

Next, three single performance measures for many fields of a video sequence (one for luma, one for chroma, and one for combined luma-chroma) are computed. PQR values for each field in a sequence are reduced to one Picture Quality Rating for the entire sequence, again by a Minkowski Q-norm.

III.5 Correlation with subjective results

III.5.1 Overview

The IRT (Institut für Rundfunktechnik GmbH, Munich, Germany) and Tektronix recently completed the initial phase of an investigation into the performance of an objective picture quality rating (PQR) method based on the jointly-developed Sarnoff/Tektronix Human Vision Model. This clause provides a brief summary of the results of a blind test comparing the PQR Picture Quality Metric and the subjective Mean Opinion Scores (MOS) of viewers. The data set of 60 video scenes used in the experiment was generated by IRT from five different video sequences passed through two different MPEG-2 encoders at compressed rates of 2, 3, 4.5, 7, and 10 Mbit/s. The MOS scores were determined by IRT and the objective PQR assessments were determined by Tektronix. The subjective scoring procedure used panels of 25 assessors and followed strict ITU-R BT.500-7

(DSCQS method) procedures. The objective PQR scores were computed by Tektronix with the Sarnoff/Tektronix Human Vision Model based on Just-Noticeable Difference principles. No model parameters were adjusted to fit the IRT data set. To avoid possible biases in the experiment, the subjective and objective ratings were exchanged by Tektronix and IRT only after each group had completed their scoring. Given the absence of any adjustments to the model parameters, which are based on human vision science, the agreement between subjective and objective results displays a strong correlation of 0.88. Correlation over typical broadcast quality is 0.91. The results are shown in Figure III.12, and are promising for the future use of objective methods in the characterization and monitoring of video picture quality.

III.5.2 Video test set and processing

The video test scenes were supplied by IRT to Tektronix in SMPTE 125M 422-625/50 Hz format (i.e. PAL D1 tape format). Each scene is of 9 seconds duration. In the following, HRC stands for "Hypothetical Reference Circuit" (as defined by ANSI T1A1.5). Before the video was passed through the HRCs, Tektronix added a barcode near the top of each video frame. This code is used for determining horizontal and vertical pixel misalignment, frame count and other factors. The stripe was covered for the subjective tests, but the results of a test with a small control group and visible stripe showed that the stripe had little effect on viewer evaluations. After the alignment stripes were added, the sequences were passed through the HRCs by IRT. Two video coders (IRT[4] and Thomson) were employed at bit rates of 2.0, 3.0, 4.5, 7.0 and 10.0 Mbit/s. Although commercial broadcast systems are unlikely to operate below 3 Mbit/s, the 2.0 Mbit/s scenes were included to explore performance beyond normal limits. A final set of HRCs consisted of following a PAL conversion stage with the same two coders running at 3 Mbit/s. It is expected that the PAL conversion in particular would likely introduce some sub-pixel misalignment. The original sequences and their processing into the test scenes are summarized below.

Original sequences	HRCs	Bit rates Mbit/s
1. Barcelona	1 IRT Coder	2.0
2. Mobile and Calendar	2 "	3.0
3. NDR	3 "	4.5
4. Football (Soccer)	4 "	7.0
5. Flower Garden	5 "	10.0
	6 Thomson Coder	2.0
	7 "	3.0
	8 "	4.5
	9 "	7.0
	10 "	10.0
	11 PAL + MPEG (Thomson)	3.0
	12 PAL + MPEG (IRT)	3.0
	13 Reference – no compression	

Total Test Set of **60 Scenes** = (**5** sequences) × [(**2** encoders) × (**5** bit rates) + **2** PAL]

Barcelona: Colourful patterned extravaganza parade formation on a large playing field (see Figure III.13). The camera is slowly zooming out and the motion is low. The background stands contribute fine detail. The sequence is colourful, low motion, fine detail.

---

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

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

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

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

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

In Figure III.13 a typical frame image for each of the above sequences is shown.

III.5.3 Subjective evaluation

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

The presentation structure consisted of the following phase lengths illustrated in Figure III.10.

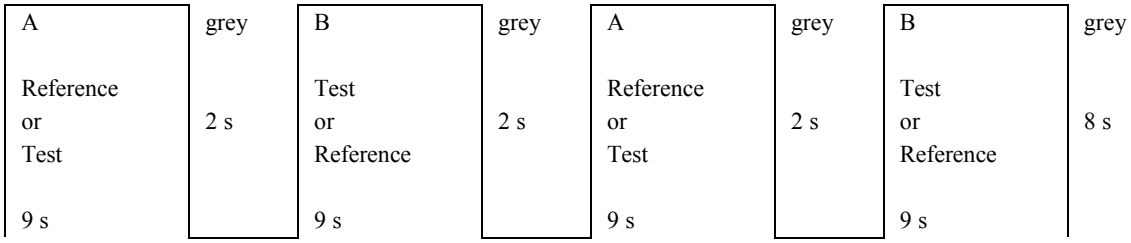


Figure III.10/J.144 – Presentation order for DSCQS method

A was the reference and B the HRC or vice versa, varying from test to test. The order was unknown to the assessors. The overall length of a test was 50 seconds.

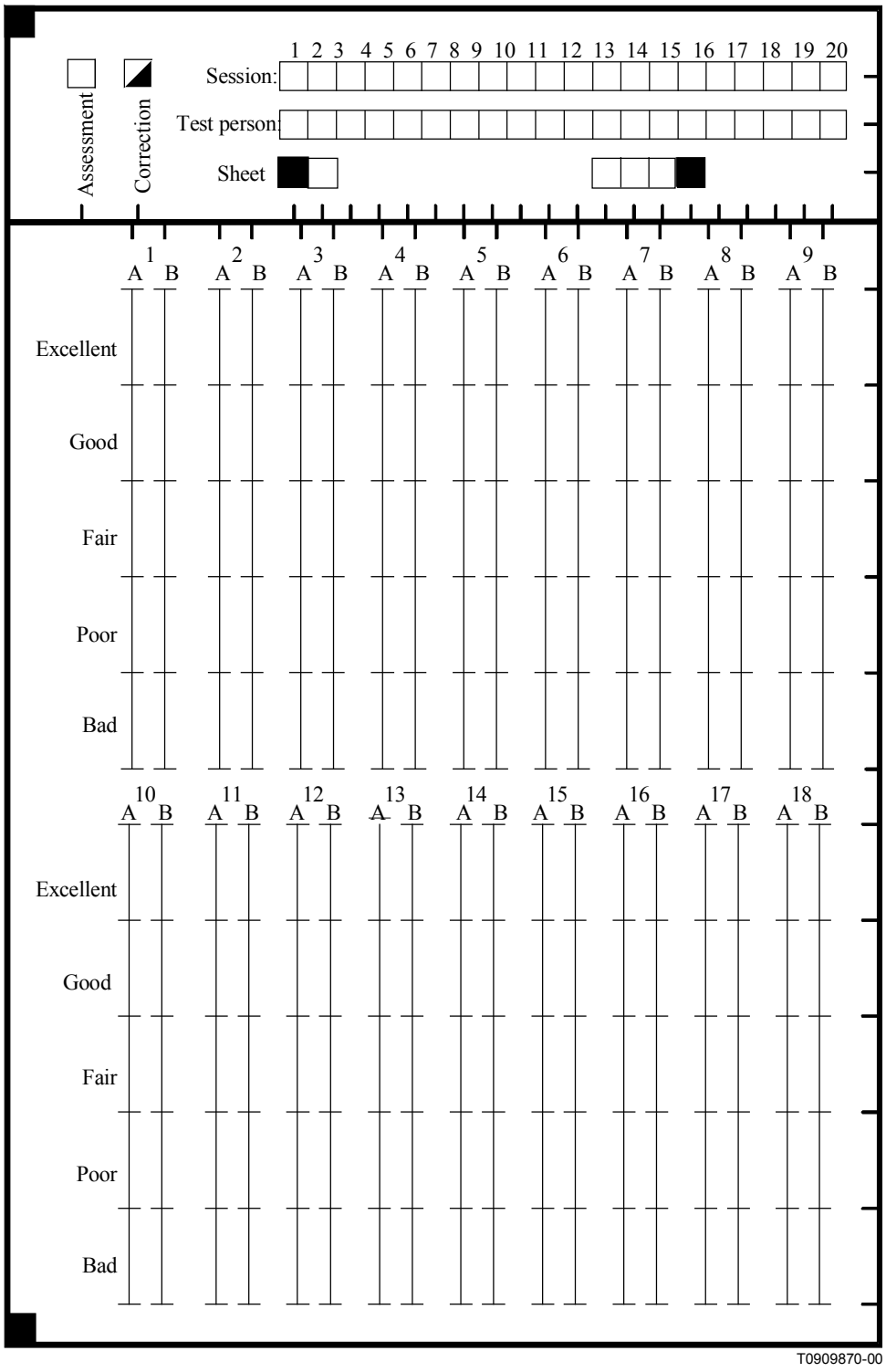


Figure III.11/J.144 – Test sheet used for the assessment of the test sequences

For the rating of the test sequences, a test sheet of the following type as shown in Figure III.11 was used.

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

In addition to the real test, examples and training sequences were shown. Four examples were shown at the beginning of the first session. They demonstrated the test method and spanned the quality range to be expected. The viewers were told not to assess the sequences because they were only examples. The examples are listed in Table III.2 below.

Table III.2/J.144 – Example sequences

Number	Test sequence	Coder	Bit rate Mbit/s
1	Zoom on a street	IRT	3
2	Barcelona 2	Thomson	4
3	Zoom on a street	IRT	10
4	Barcelona 2	Thomson	2

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

The training sequences had to be assessed by the subjects who did not know that the results were not evaluated. The training sequences are listed in Table III.3.

Table III.3/J.144 – Training sequences

Number	Test sequence	Coder	Bit rate Mbit/s
1	Renata	Thomson	2
2	Table Tennis	IRT	10
3	Renata	Thomson	4
4	Table Tennis	IRT	2
5	Renata	Thomson	10
6	Table Tennis	IRT	4

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

The test sessions were structured in the following way:

Session 1: examples (4) – training sequences (6) – real tests (31)

Session 2: training sequences (6) – real tests (34)

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

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

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

III.5.4 Objective picture quality assessment

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

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

III.5.5 Comparison of subjective and objective assessments

Figure III.12 displays the subjective MOSs determined by IRT and the objective PQRs estimated by Tektronix. The vertical error bars display the 95% confidence intervals for the spread in subjective viewer ratings. The relationship between subjective and objective assessments is well behaved and monotonic with a strong correlation of 0.88. From the rightward curvature in the relationship, it can be seen that there is a compression in viewer's picture quality assessment as quality degrades towards very poor. This effect is well known in the field of subjective testing, and is consistent with the compression effects found in other areas of human perception such as loudness and brightness. The group of three points in the upper right hand corner contains scenes where the encoder either failed catastrophically in regions of the scene or the quality was very poor. If these points are excluded then the correlation coefficient increases to 0.91. Given that the objective quality ratings did not require any fitting or optimization of parameters to the test data set, the results are quite encouraging that objective methods will contribute to reducing the time, expense, and possible biases associated with characterization and monitoring of video.

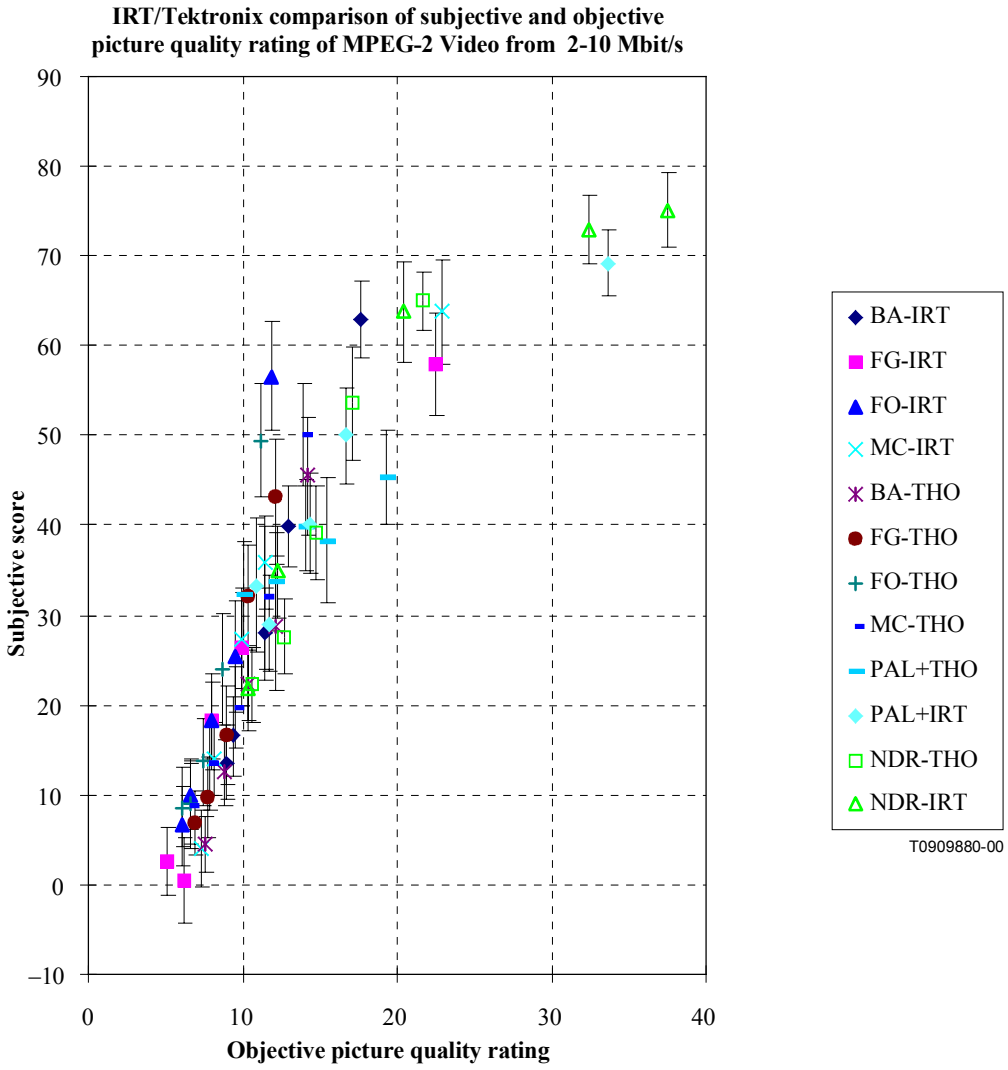
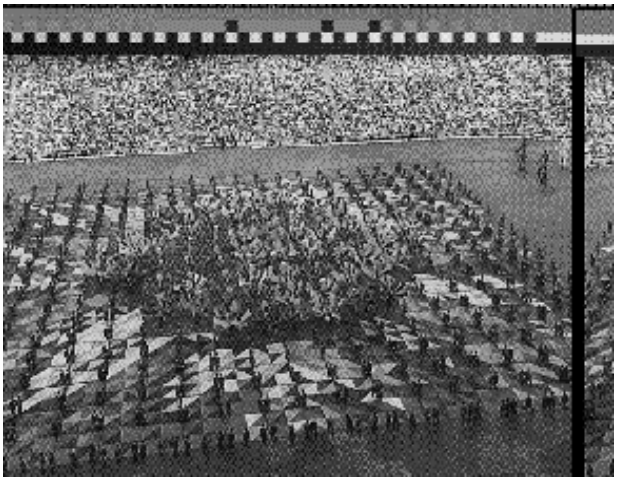
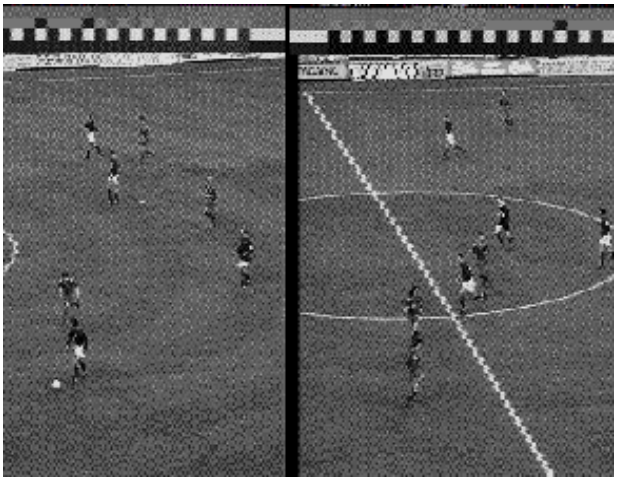


Figure III.12/J.144 – Comparison of IRT subjective mean opinion scores (MOSs) and Tektronix objective picture quality rating (PQR) for 60 2-10 Mbit/s MPEG-2 and PAL test scenes

The 95% confidence intervals for subjective scores are indicated by vertical bars. Correlation between objective and subjective ratings is 0.88 for the complete data set, and viewer compression in quality rating is apparent for upper right poorest quality scenes. The correlation is 0.91 if upper rightmost data scenes of poorest quality are excluded.

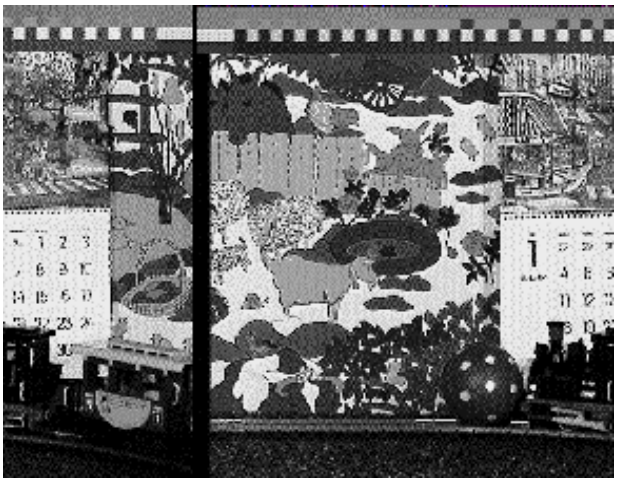


Barcelona



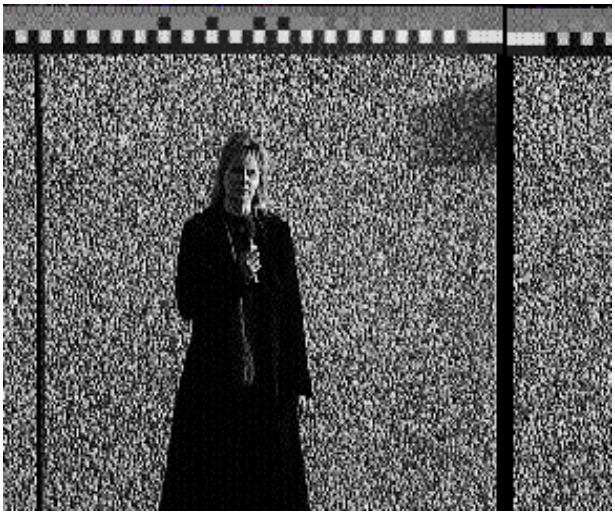
Football

Flower garden



Mobile and calendar

T0909890-00



NDR

Figure III.13/J.144 – Typical frame images of video test sequences

III.6 References

– LUBIN (J.): The use of psychophysical data and models in the analysis of display system performance, in A.B. Watson (ed.), *Digital Images and Human Vision*, *MIT Press*, pp. 163-178, 1993.

– LUBIN (J.): A visual system discrimination model for imaging system design and evaluation, in E. Peli (ed.), *Visual Models for Target Detection and Recognition*, *World Scientific Publishers*, 1995.

– MULLEN (K.T.): The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings, *J. Physiol.* 359, 381-400, 1985.

– WYSZECKI (G.), STILES (W.S.): Color Science, 2nd ed., *Wiley*, 1982.

APPENDIX IV

NHK/Mitsubishi Electric Corp.

Abstract

A digital compressed picture quality assessment system has been developed, in which picture degradation is calculated in real-time taking account of human visual perception. In this system, noise sensitivity in spatio-temporal frequencies is considered in relation to picture brightness. This approach has improved the accuracy of picture quality assessment for many types of degradation.

IV.1 Method of evaluating quality deterioration objectively

The model emulates human-visual characteristics using 3D (spatio-temporal) filters, which are applied to differences between source and processed signals. For the filter implementation, block-type frequency analysis methods like DCT are not used to avoid potential mutual effects between coding and assessment systems. The filter characteristics are varied based on the luminance level. The output quality score is calculated as a sum of weighted measures from the filters. The system is aimed at assessing picture quality in terms of fineness and repeatability, by exactly reflecting visual functions in the assessment system. In the following, these human visual characteristics are described, followed by explanations of hardware implementation.

IV.2 Human visual characteristics

IV.2.1 Spatial frequency response of visibility

The spatio-temporal frequency response of human visibility as shown in Figure IV.1 has been measured by J.G. Robson [1] and others. The spatial frequency response of visibility displays a sectional feature of perpendicularity to the temporal frequency response, achieving rotational symmetry with the optical axes in centre.

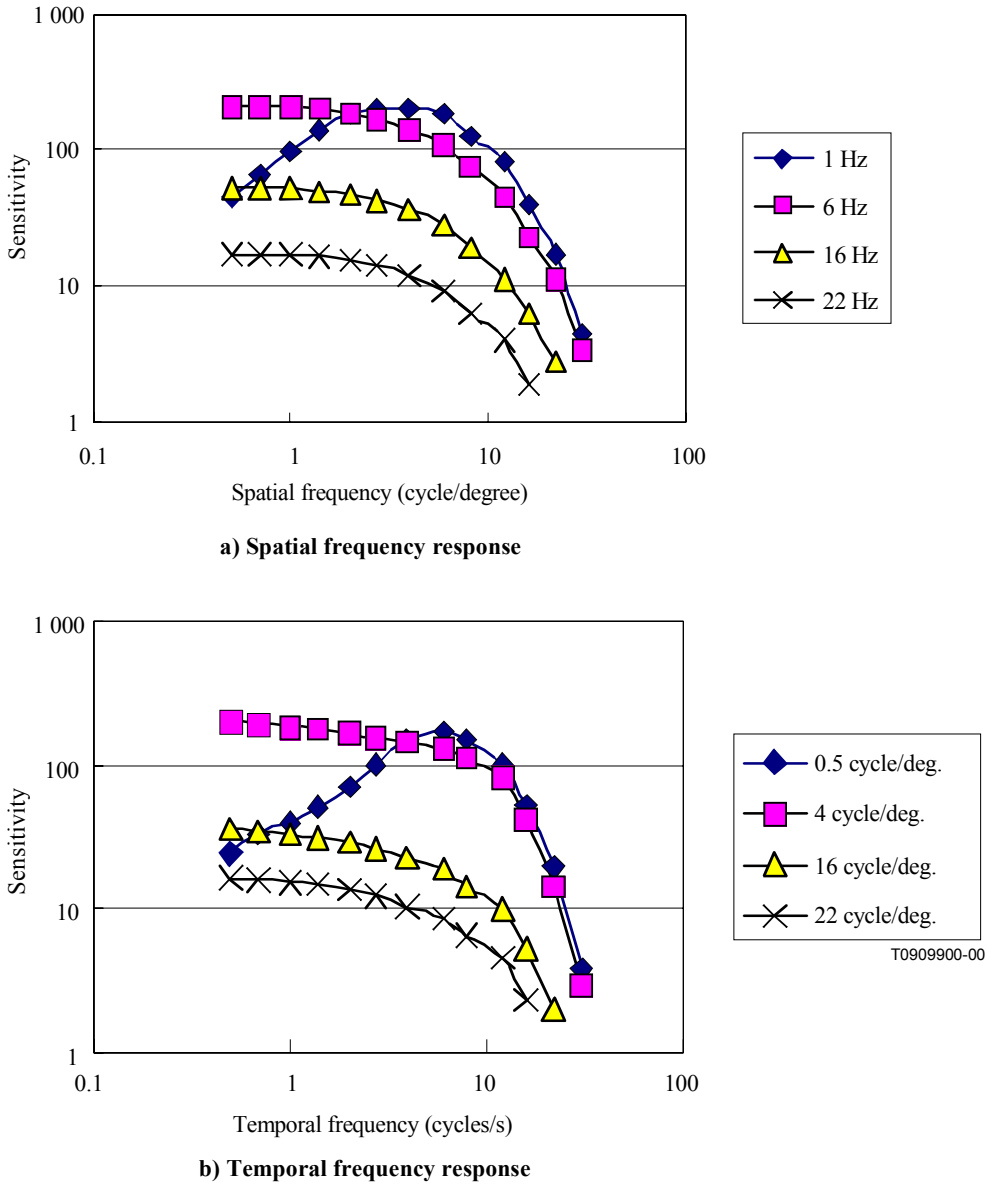


Figure IV.1/J.144 – Spatio-temporal frequency response of visibility

IV.2.2 Frequency response of visibility depending on picture brightness

As regards brightness dependency on the frequency response of visibility, the measurements by Kelly [2] and others show that not only the spatial frequency response but also the temporal frequency response depends on picture brightness. Figure IV.2 indicates dependency of the spatial frequency response of visibility on picture brightness in the case of an almost still picture with a temporal frequency of less than 4 Hz for visual sensitivity. "td" is a unit for the luminance of an eyeground image.

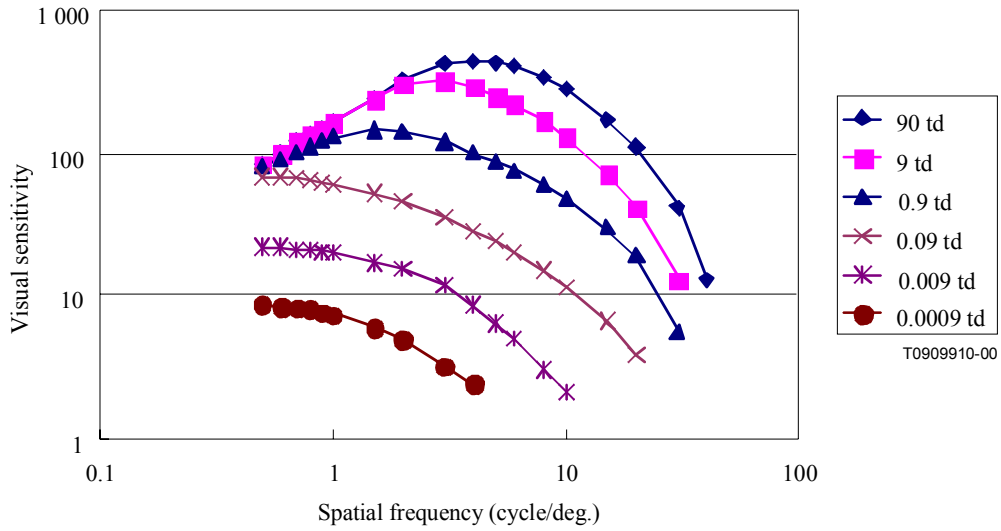


Figure IV.2/J.144 – Dependency of spatial frequency response of visibility on brightness

Figure IV.3 shows dependency of the temporal frequency response of visibility on brightness in the case of a uniform image. Human eyes are typically sensitive to flicker of about 10 Hz when picture brightness is high. When it is very low, flicker is largely invisible.

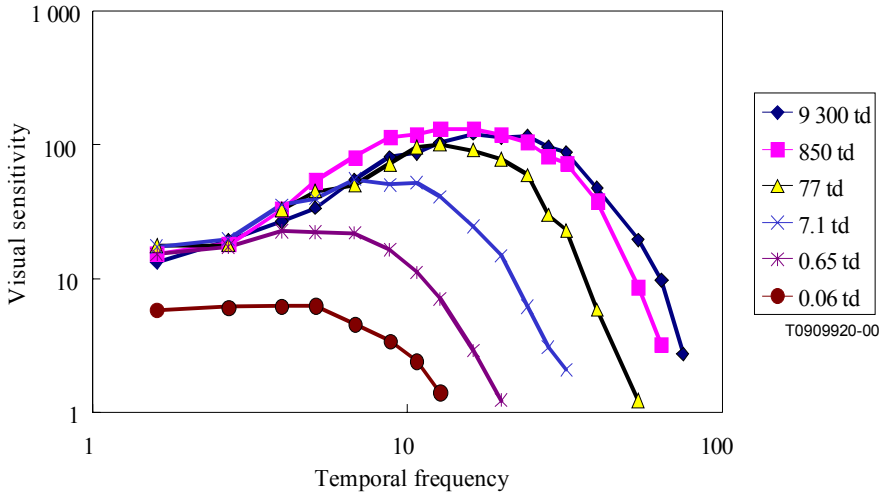


Figure IV.3/J.144 – Dependency of temporal frequency response of visibility on brightness

IV.2.3 Visual sensitivity depending on brightness

Figure IV.4 shows the perception limits of random noise on the TV monitor [3] at different brightness levels. It is found that there is dependency of visual sensitivity on brightness.

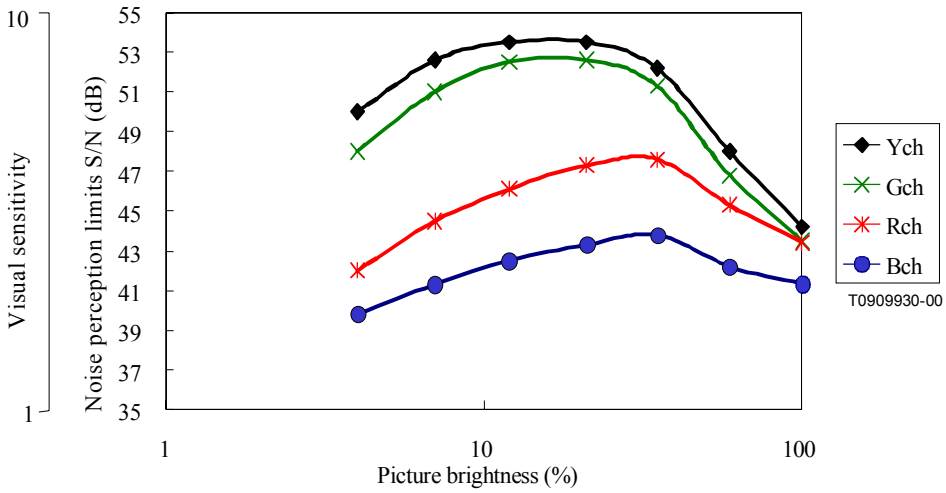


Figure IV.4/J.144 – Perception limits of random noise on the TV monitor

IV.3 Realization of visual functions by digital filter

IV.3.1 Structure of the assessment system

Figure IV.5 shows the structure of the assessment system. First, difference signals from the original and test sequence pictures are produced and then fed to the brightness-adaptive 3D digital filter with the same 3D frequency response of visibility and brightness dependency. Next, the filtered difference signals are compared with visual perception in each pixel. As a result, a numerical expression of the distortion beyond the perception limits of the human eye is obtained.

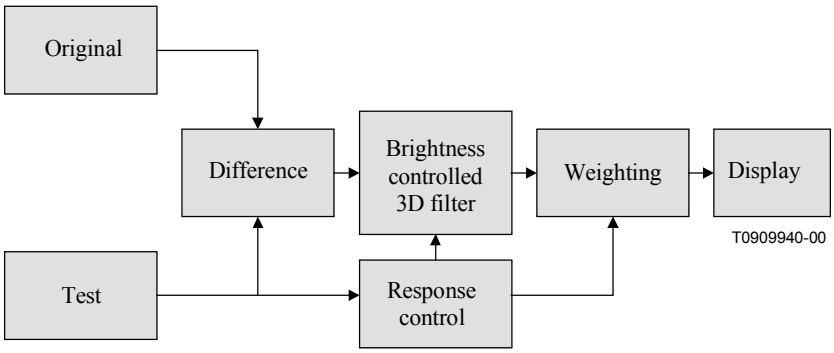


Figure IV.5/J.144 – Structure of an experimental assessment system

IV.3.2 Brightness-adaptive 3D digital filter

Figure IV.6 shows the composition of 3D digital filters, with the frequency response and sensitivity changing according to brightness. By combining the spatial filters and the temporal filters according to picture brightness, the frequency response of human visibility is emulated.

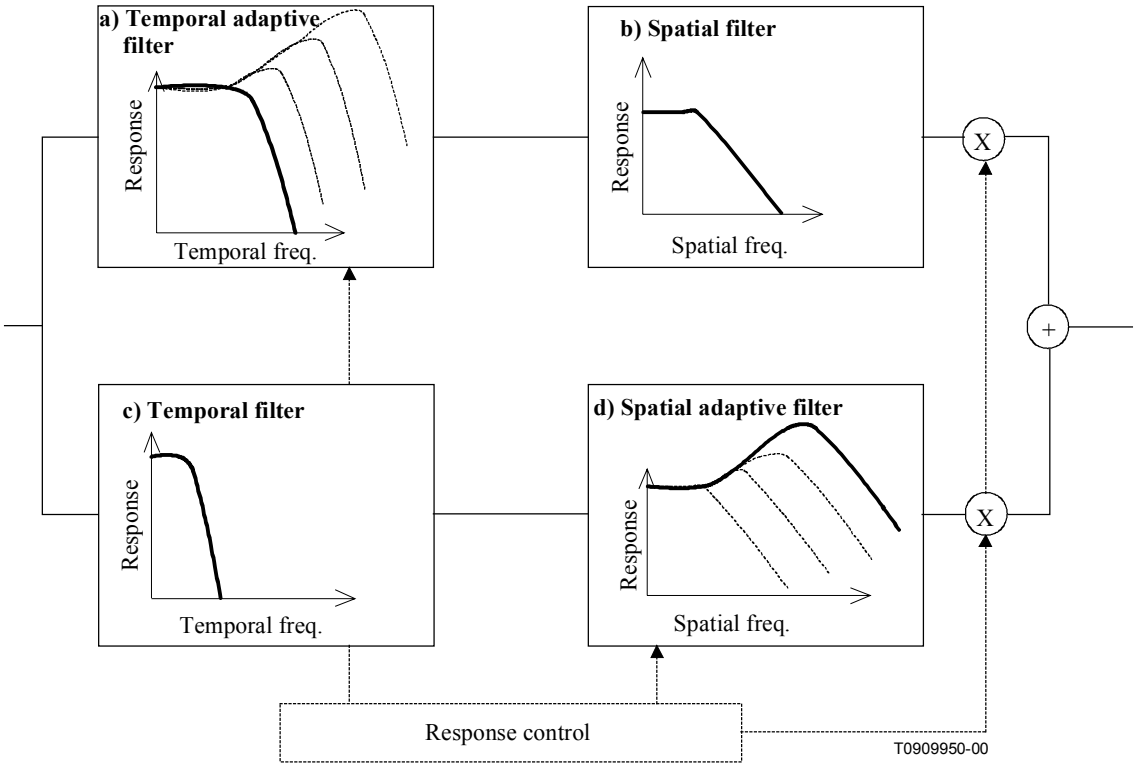


Figure IV.6/J.144 – Composition of 3D digital filters, with the frequency response and sensitivity changing according to brightness

IV.3.3 Adaptive spatial filter depending on picture brightness

Figure IV.7 shows the adaptive spatial filter d) in Figure IV.6, which is obtained by switching spatial filters having different characteristics.

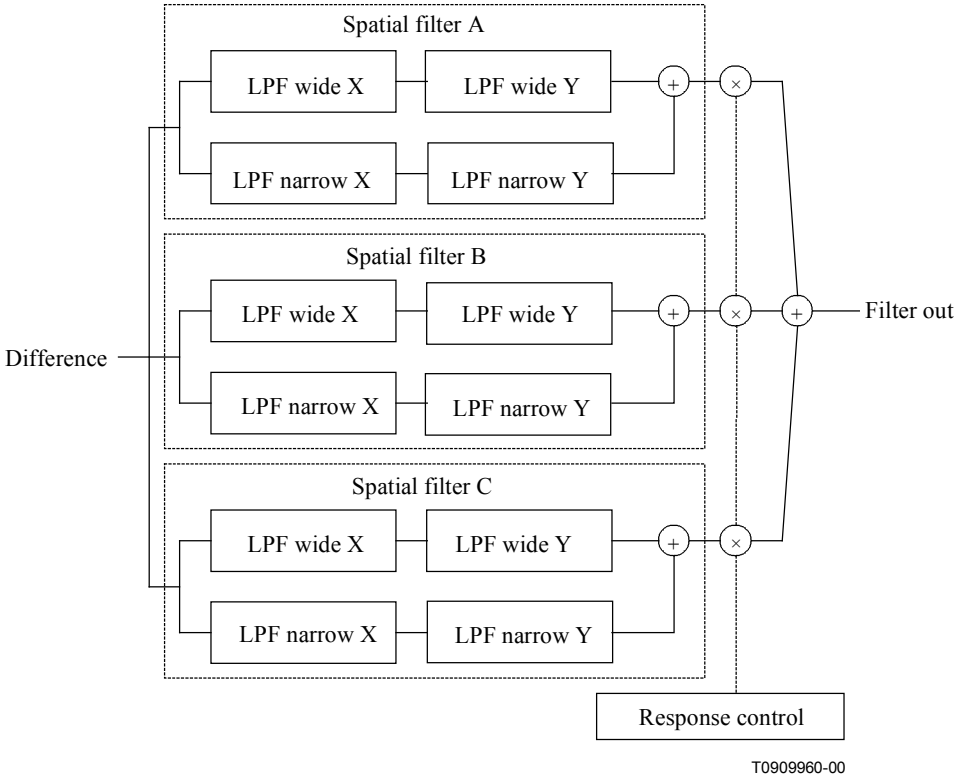


Figure IV.7/J.144 – Adaptive spatial filter obtained by switching spatial filters

IV.3.4 Volcano-shaped spatial frequency response

It follows that visual functions possess the characteristics of a band-pass spatial filter the horizontal and vertical axes. Representing these characteristics by a 3D digital filter, we obtain the volcano-shaped feature shown in the contour graph of Figure IV.8. This feature represents the response of human eyes to the effect that deterioration is conspicuous on the edges of the picture.

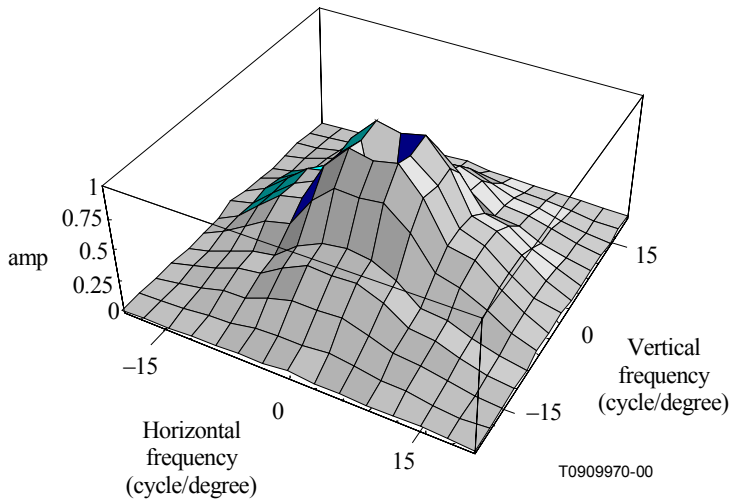


Figure IV.8/J.144 – Volcano-shaped spatial filter

IV.4 Example of assessment by the picture quality assessment system

Figure IV.9 illustrates the relationship between subjective assessment scores obtained by 20 video experts in accordance with ITU-R BT.500 and objective assessment scores that we obtained using our new assessment system. We made our assessment using component and composite images for the test sequence, and component images for the original sequence.

Regarding not only compression distortion, but also quality deterioration, including composite/component conversion and bandwidth limits, we can see that the picture quality (PQ) by the objective assessment system agrees well with the double stimulus continuous quality scale (DSCQS) of subjective evaluation.

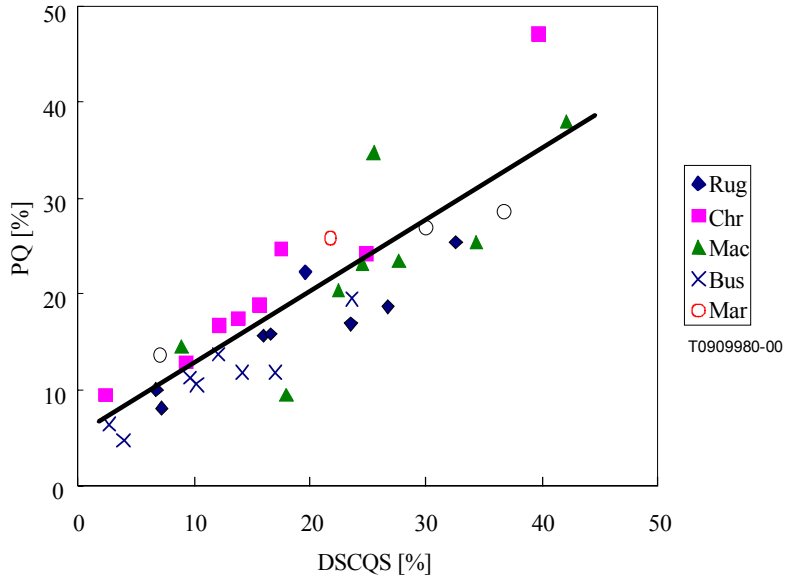


Figure IV.9/J.144 – Relationship between subjective and the objective assessment results

For reference, a relationship between RMS errors of the processed pictures and subjective scores is shown in Figure IV.10. In contrast to Figure IV.9, this graph shows lower correlation.

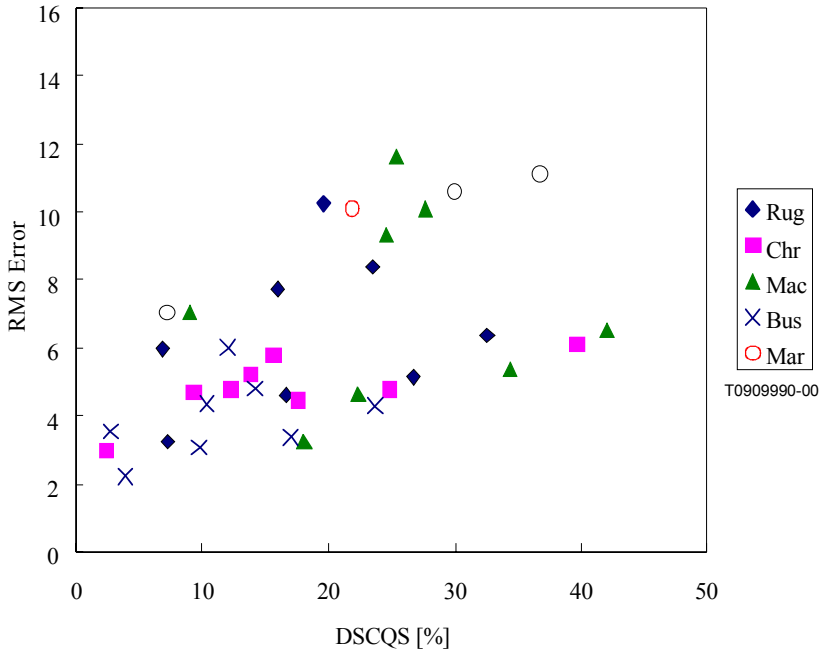


Figure IV.10/J.144 – Relationship between RMS error and subjective scores

IV.5 Real-time picture quality assessment system

Figure IV.11 is the external appearance of the assessment system. The system features:

1) real-time measurement;

2) automatic adjustment of CODEC delay and synchronization phase shift; and

3) ease of measurement because the system has original images built in.

The system has been improved in terms of accuracy by the faithful representation of human-style visibility functions. Typically, the human eye can see detail when it is bright, but only indistinctly when it is dark. Obviously, the spatial frequency response varies according to brightness. Moreover, our eyes can see flicker on the screen well when it is bright. When it gets dark, the temporal frequency response changes with the persistence of vision. Using the system, we have represented to the best of our ability human-style visibility functions that change substantially according to the level of brightness.

The assessment system has made it possible to measure with high correlation to subjective evaluation, irrespective of the type of video signal, and with good representation.

Figure IV.11/J.144 – External appearance of the assessment system

IV.6 References

[1] ROBSON (J.G.): Spatial and Temporal Contrast-Sensitivity Functions of the Visual System, *J.Opt. Soc. Am.*, pp. 1141-1142, August 1966.

[2] KELLY (D.H.): Visual Responses to Time-Dependent Stimuli. I. Amplitude Sensitivity Measurements, *J. of the Opt. Soc. of Am.*, Vol. 51, No. 4, pp. 422-429, April 1961.

[3] NISHIDA (Y.), KOIKE (J.), OHTAKE (H.), ABE (M.), YOSHIKAWA (S.): Design Concept for a Low-Noise CCD Image Sensor Based on Subjective Evaluation, *IEEE Trans. ED.*, Vol. 36, No. 2, 1989.

APPENDIX V

KDD
Objective video quality assessment scheme and performance evaluation

V.1 Scope

Recently, digital television broadcasting and transmission services are beginning to come into practical use. These services use video codecs (video signal encoding devices) based on MPEG-2, an international standard method for compression of digital video signals. Video codecs are comprised of encoders, which perform the compression, and decoders, which reconstruct the compressed video data. These devices work by removing redundant information from the enormous volume of information contained in video signals. This makes it possible to transmit the information efficiently using only a limited amount of bandwidth.

There is always some amount of degradation in the quality of video that has been compressed and transmitted using a video codec. The amount of degradation depends on the contents of the picture. Generally there is more distortion in fast-moving scenes, like those in a sports broadcast. There are also variations in the quality of the output produced by different codecs. MPEG-2 is an international standard, but the quality of specific types of compressed video still depends to a certain extent on the manufacturer's implementation.

For its television transmission especially in TV1, TV2 and TV3 (Contribution, Primary and Secondary distribution) [1], it is required to strive to achieve consistently high quality by constantly monitoring the quality of the transmitted pictures.

In conventional analogue FM transmission, there is little degradation in the picture due to the contents or to analogue modulation, so quality is stable. But in the transmission of compressed

digital video, the quality of the picture varies as described above according to the nature of the contents and the codec employed, and checking the quality of this kind of video is expected to be a very complex operation.

Hence, it is considered necessary to standardize a scheme to evaluate the picture quality of MPEG-2 based video codecs mainly used in TV1, TV2 and TV3. In these classes, the following functions are considered to be necessary:

- Generic assessment for various types of video contents Analogue/Digital Composite/Component video formats are supported.
- Real time assessment Precise temporal and spatial alignment between an original and a codec out signal.
- Sensitive and accurate assessment to subtle and complex distortions.

Considering the above, we are proposing a new evaluation scheme and its implementation based on the characteristics of human visual perception, enabling very precise measurements of video quality in [2]. In this appendix, we report verification results of this scheme.

V.2 Objective video quality assessment scheme

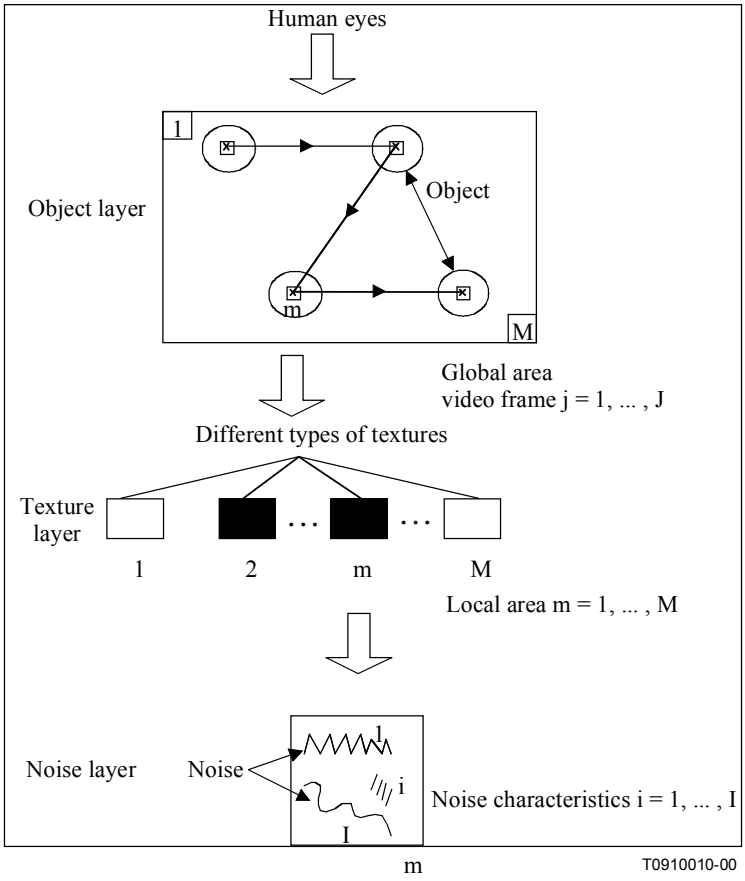


Figure V.1/J.144 – Three-layered model for video signal

Figure V.1 shows the three-layered picture quality assessment model as seen by the human eye. Generally, the human eye cannot watch a whole frame at a glance, but watch only a local spot area in a frame, which is around the gaze point of the human eyes, and recognizes the texture and also quality of the area depending on the degrees and characteristics of noise mixed in this texture. The whole frame is understood by moving the gaze point among objects, which are picture components

of the frame and picture quality assessment is also conducted for the whole frame at the same time. In this process, picture quality is determined by the noise over a frame. Therefore, to perform objective measurement of subjective picture quality, the macro-to-micro three-layered picture structures (object, texture and noise layers) are used, and a bottom-up noise weighting scheme is proposed which uses a particular weighting function at each layer taking into account human visual perception (Figure V.2).

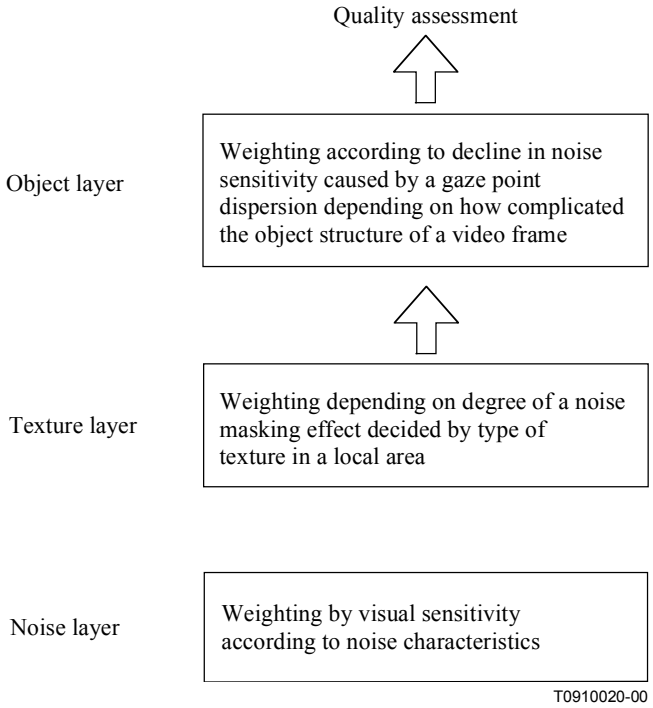


Figure V.2/J.144 – Three-layered bottom-up noise weighting

First, at the noise layer, common noise in a video compression process such as high frequency noise, low frequency noise, chroma noise, jerkiness, flicker and so on are weighted depending on their degrees and characteristics. For this weighting, it is useful to perform a frequency conversion to classify these noises. Second, at the texture layer, local spot areas are classified into several groups by their texture types. These groups include for example, "detail texture" such as a forest, trees and a stadium in which noises are strongly masked, and "flat texture" such as a human skin and a sky in which noises are easily recognized. Consequently, noises are weighted more or less according to their texture types. Finally, at the object layer, the dispersion degree of the gaze point is predicted by measuring how complicated the structure is of objects in the video frame. Then, noises in the whole frame are weighted corresponding to a decline in noise sensitivity caused by this dispersion.

To obtain mathematical expressions for these weighting processes, we make the following definitions:

$P(j,m,i)$: Power of a noise i in a local area m of a frame j

h_i: Weighting function for a noise i

$C(j,m)$: Texture in a local area (j,m)

t_c: Noise weighting function in a texture C

$G(j)$: Parameter indicating how complicated the structure is of objects of a frame j

q_G: Noise weighting function depending on dispersion degree of a gaze point

Following these definitions, noises are summed up in order from the low layer to the high layer.

In the noise layer, by summing up noise which is weighted by h_i corresponding to noise characteristics in a local area (j,m), we calculate $WMSE_{NL}$ as:

$$WMSE_{NL}(j,m) = \frac{1}{I}\sum_{i=1}^{I} h_i \cdot P(j,m,i) \quad \text{(V-1)}$$

Next, at the texture layer, by summing up $WMSE_{NL}(j,m)$ over the whole frame (m = 1, ..., M) being weighted by t_c corresponding to a texture C(j,m) in a local area (j,m), we calculate $WMSE_{TL}(j)$ as:

$$WMSE_{NL}(j) = \frac{1}{M}\sum_{m=1}^{M} t_c(j,m) \cdot WMSE_{NL}(j,m) \quad \text{(V-2)}$$

Finally, at the object layer, by taking an average value of $WMSE_{TL}$ over frames j = 1, ..., J being weighted by G(j) corresponding to the dispersion degree of the gaze point, we calculate $WMSE_{OL}$ as:

$$WMSE_{OL} = \frac{1}{J}\sum_{j=1}^{J} q_G(j) \cdot WMSE_{TL}(j) \quad \text{(V-3)}$$

We further convert this $WMSE_{OL}$ to WSNR and calculate the DSCQS (Double-stimulus continuous quality-scale method) (0-100%) defined in ITU-R BT.500-7 as:

$$WSMR(dB) = 10 \log_{10} \frac{255^2}{WMSE} \quad \text{(V-4)}$$

$$D(\%) = f(WSNR) \quad \text{(V-5)}$$

V.3 Implementation

The system is made up of two parts: a synchronization module, which enables precise comparison between the reconstructed video and the original video, and a calculation module, which determines video quality with reference to characteristics of human visual perception. Figure V.3 shows the configuration of the system, and Table V.1 describes principal parameters. As Table V.1 shows, both composite (NTSC)/component signals with full samplings are supported.

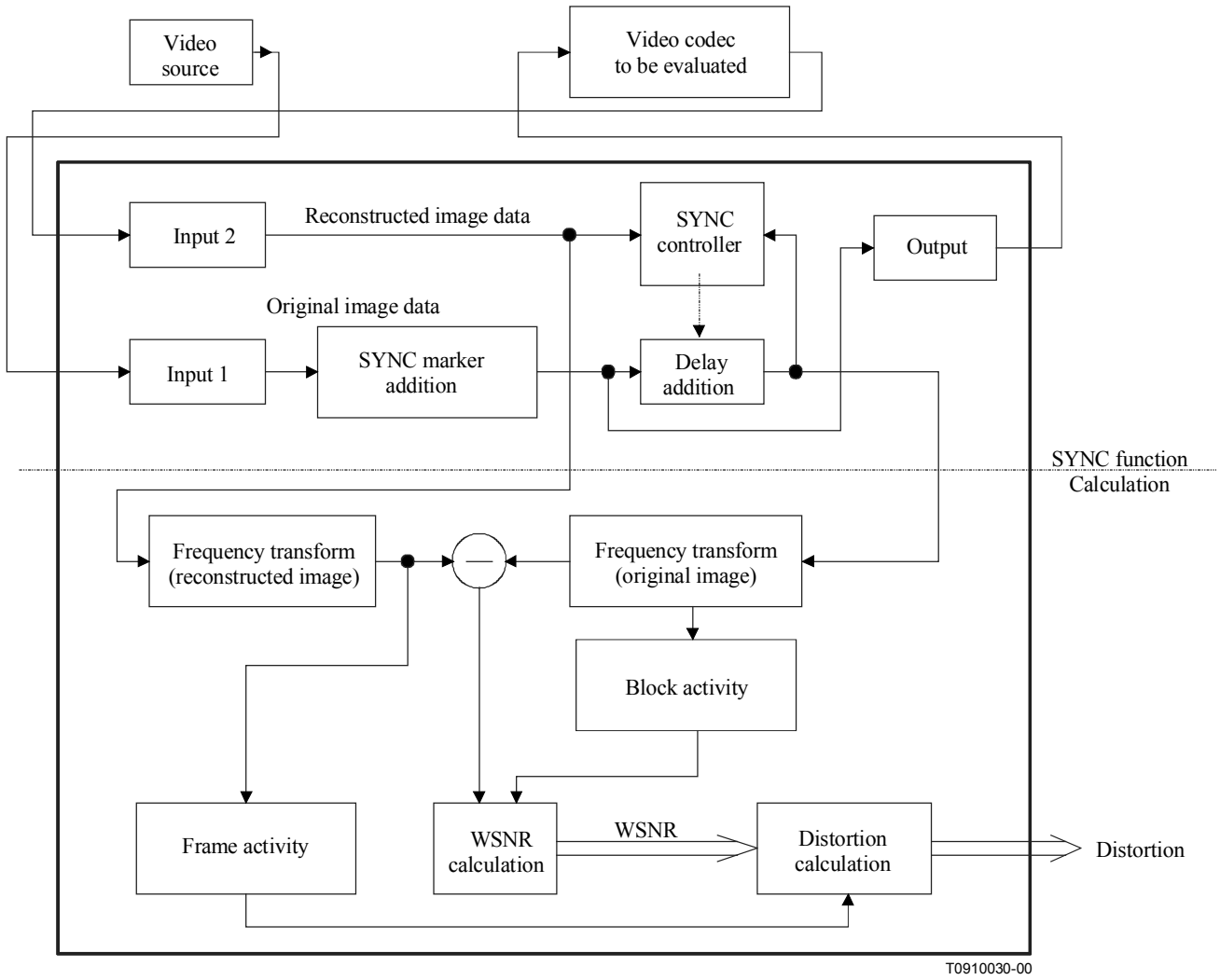


Figure V.3/J.144 – System configuration

V.3.1 Synchronization module

Television signals from the original video source are read into the system through input module 1 and marked with a synchronization marker that varies with each frame. Then the frames with markers are sent to the delay module, where they are stored in memory. At the same time, the frames are sent via the output module to the video codec that is to be evaluated. The video codec compresses the frames, which are read into the system again through input module 2 and compared with the marked frames stored in the delay of the video codec being evaluated. Finally, the synchronization module performs temporal (frame delay) and spatial (line and pixel shift) alignment precisely, so that the amount of quality degradation described below will be as close as possible to subjective assessment by human viewers.

These operations provide the synchronization needed for the evaluation, and the markers used in these operations are designed so as to work well even through the severely signal-distorted process such as high compression, Y/C separation and filterings in a video codec.

V.3.2 Calculation module

Unlike human vision, calculation of the quality of the picture takes a bottom-up approach, building up the whole from the various parts. First, in order to evaluate the effect of variations in sensitivity due to the spatial frequencies of noise, a difference value (noise is obtained for the frequency components of the original picture and the reconstructed image). This value is input into the WSNR

(Weighted Signal-to-Noise Ratio) module, which assigns different sensitivity weights for each frequency region. At the same time, it obtains a value (the block activity) that indicates whether each block in the picture is flat or busy. The noise masking effect is also applied to obtain an overall WSNR.

Finally, a value to indicate the size of the objects making up the picture is obtained (the frame activity). This enables the system to estimate the degree to which sensitivity to noise decreases due to dispersion of the amount of degradation in quality is obtained by applying the decrease in sensitivity to noise to the WSNR.

Table V.1/J.144 – Principal parameters

Applicable video signal format	NTSC composite signal 525/60 component signal D1 serial digital
Sampling frequency (Analogue input)	14.318 MHz (NTSC) 13.5 MHz (Component Y) 6.75 MHz (Component C)
Applicable codec	MPEG-1, 2 based codec Composite codec, etc.
Effective evaluation area	768 pixels~480 lines (NTSC) 720 pixels~480 lines (Component Y) 360 pixels~480 lines (Component C)
Signal analysis	Hadamard transform (NTSC) Discrete cosine transform (Component) Alternative: Fourier transform
Noise Weighting	Spatial frequency visual sensitivity Noise masking effect Gaze point scattering
Evaluation result	Picture quality assessment (Distortion, %) WSNR (dB) SNR (dB)
Control signal interface	RS-232C

V.4 Verification results

We compared the evaluation results of a proposed scheme with subjective assessment test results which have been already graded following ITU-R BT.500-7. Assessment targets are MPEG-2 SP@ML with 5 Mbit/s, 7 Mbit/s and 10 Mbit/s applied for ITU-R BT.601, 4:2:2 component TV test signals. These are 17 data including Mobile, Flower garden, Cheer leaders etc. Therefore, we have in total 17 data × 3 bit rate = 51 samples (Table V.2).

For these samples, we conducted the subjective assessment test on two different days (23 and 24 March 1995) with the same conditions and viewers. The "triangle" of the objective assessment and two days subjective assessment results are shown in Figure V.4.

Table V.2/J.144 – Test data list

1	Susie
2	Popple
3	Table tennis
4	Mobile & Calendar
5	Autumn leaves
6	Football
7	Storm
8	Cheer leaders
9	Cast
10	Cruising
11	Bicycle
12	Horse riding
13	Summer flowers
14	Ferris wheel
15	Flower garden
16	Kiel Harbor 4
17	Balls of wool

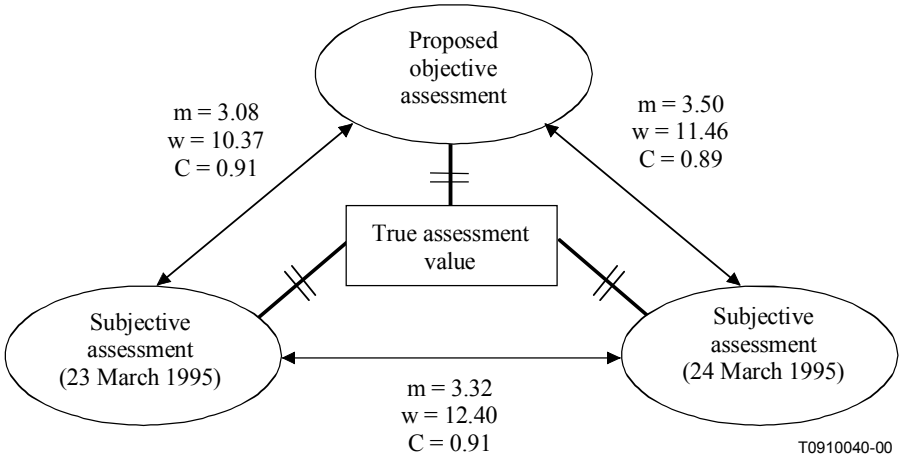


Figure V.4/J.144 – Comparisons with subjective assessment tests

Figure V.4 proves that assessment accuracy expressed by rmse, rwse and correlation of three assessment results are nearly equal from the triangle centre, which is the true assessment value. In addition, Figure V.5 shows distributions of 51 samples among an objective and two subjective assessments. Samples in the three graphs are randomly distributed but the subtle difference in each distribution can be seen. In distribution of 23rd and 24th subjective comparison, it is uniformly random but inequality in distributions can be seen in subjective and objective assessment comparisons depending on score range. That is, both graphs of 23rd and 24th vs objective scheme

give sample plots with higher correlation at 20% – 40% but less correlation at 10% – 20%. Further study will be needed to eliminate this.

By this fact, it is concluded to be feasible to use the proposed scheme in addition to ITU-R BT.500-7.

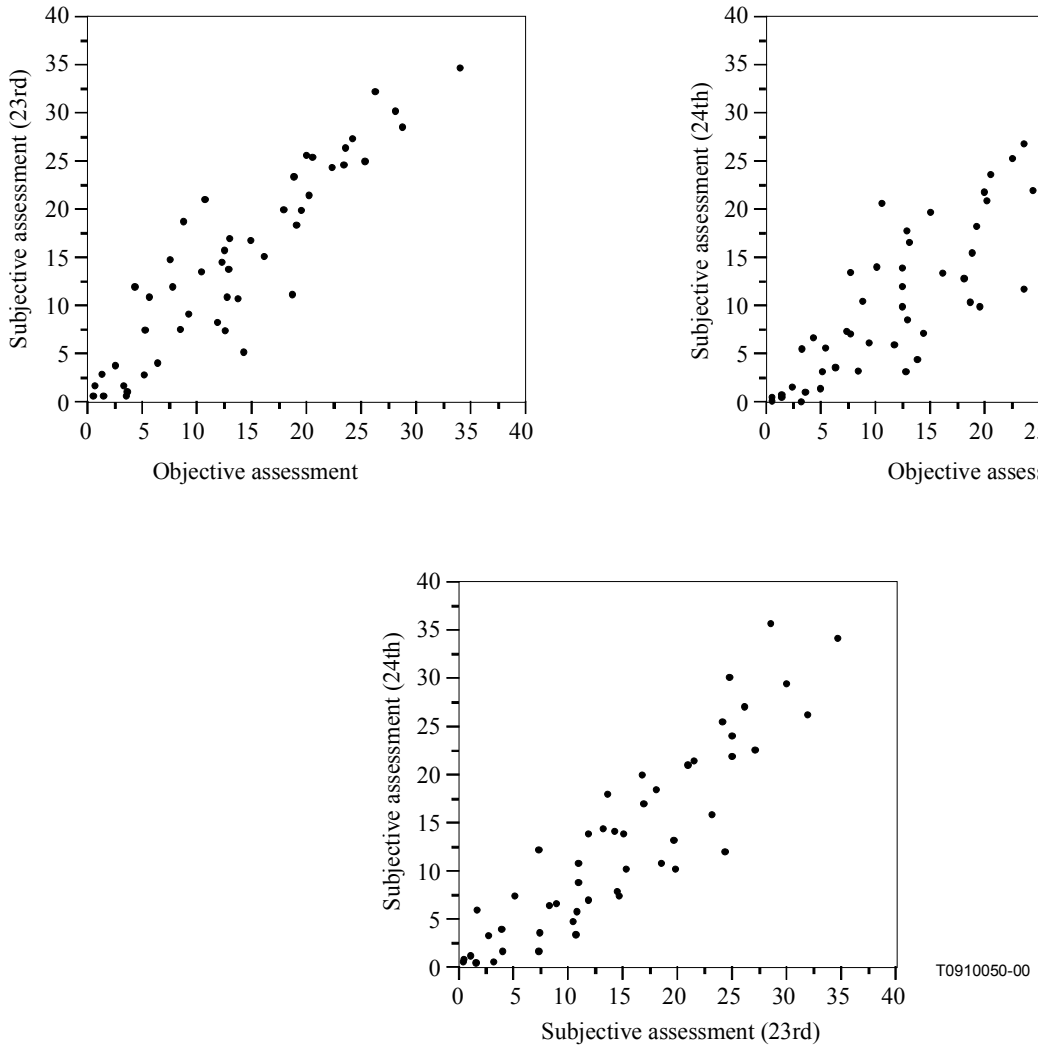


Figure V.5/J.144 – Comparisons among an objective and two subjective assessments

V.5 References

[1] *2nd version of Table defining video quality classes*, Expert meeting on subjective and objective video quality assessment, Turin, 14-16 October 1997.

[2] *Progress report on development of digital compressed picture quality assessment system in Japan*, SG 9 Document D15 Geneva, 21-25 April 1997.

APPENDIX VI

EPFL

The perceptual distortion metric (PDM) submitted by EPFL is based on a spatio-temporal model of the human visual system. It consists of four stages, through which both the reference and the processed sequences pass. The first converts the input to an opponent-colours space. The second stage implements a spatio-temporal perceptual decomposition into separate visual channels of different temporal frequency, spatial frequency and orientation. The third stage models effects of pattern masking by simulating excitatory and inhibitory mechanisms according to a model of contrast gain control. The fourth and final stage of the metric serves as pooling and detection stage and computes a distortion measure from the difference between the sensor outputs of the reference and the processed sequence.

APPENDIX VII

NASA

VII.1 Introduction

The emerging infrastructure for digital video requires a critical component: a reliable means for automatically measuring visual quality. Such a means is essential for evaluation of codecs, for monitoring broadcast transmissions, and for ensuring the most efficient compression of sources and utilization of communication bandwidths. This appendix describes a new video quality metric, called DVQ (Digital Video Quality), that can be used for automatically measuring visual quality.

VII.2 The DVQ metric

All video quality metrics are inherently models of human vision. The DVQ metric is an attempt to incorporate many aspects of human visual sensitivity in a simple image processing algorithm. Simplicity is an important goal, since one would like the metric to run in real-time and require only modest computational resources. One of the most complex and time-consuming elements of other proposed metrics are the spatial filtering operations employed to implement the multiple, bandpass spatial filters that are characteristic of human vision. We accelerate this step by using the Discrete Cosine Transform (DCT) for this decomposition into spatial channels. This provides a powerful advantage since efficient hardware and software are available for this transformation, and because in many applications the transform may have already been done as part of the compression process.

Figure VII.1 is an overview of the processing steps of the DVQ metric. These steps are described in greater detail elsewhere [1] to [3], here we provide only a brief review. The input to the metric is a pair of colour image sequences: reference (R) and test (T). The first step consists of various sampling, cropping, and colour transformations that serve to restrict processing to a region of interest and to express the sequences in a perceptual colour space. This stage also deals with de-interlacing and de-gamma-correcting the input video. The sequences are then subjected to a blocking (BLK) and a Discrete Cosine Transform (DCT), and the results are then transformed to local contrast (LC). Local contrast is the ratio of DCT amplitude to DC amplitude for the corresponding block. The next step is a temporal filtering operation (TF) which implements the temporal part of the contrast sensitivity function. This is accomplished through a suitable recursive discrete second order filter. The results are then converted to just-noticeable differences by dividing each DCT coefficient by its respective visual threshold. This implements the spatial part of the contrast sensitivity function (CSF). At the next stage the two sequences are subtracted. The difference sequence is then subjected to a contrast masking operation (CM), which also depends upon the reference sequence. Finally the masked differences may be pooled in various ways to illustrate the perceptual error over various dimensions (POOL), and the pooled error may be converted to visual quality (VQ).

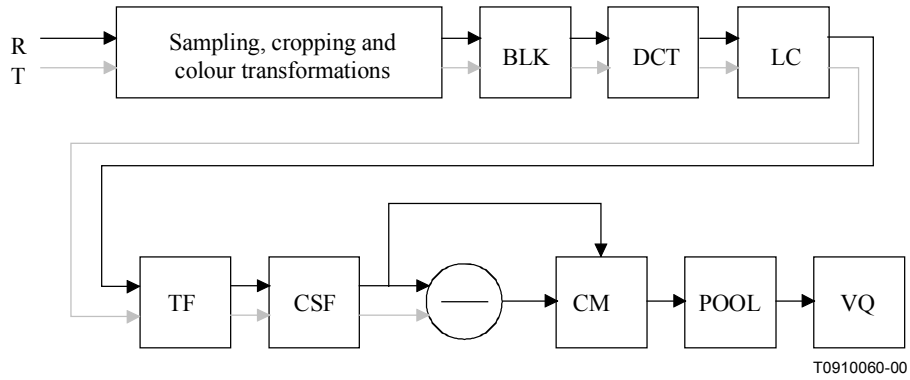


Figure VII.1/J.144 – Overview of DVQ processing steps

The parameters of the metric have been estimated from psychophysical data, both from the existing literature and from measurements of visibility of dynamic DCT quantization error.

VII.2.1 Input

The input to the metric is a pair of colour image sequences. The dimensions of this input are {s,f,c,y,x}, where s = sequence (2), f = frames, c = colour (3), y = rows, and x = columns. The first of the two sequences is the reference, the second is the test. Typically the test will differ from the reference in the presence of compression artifacts. The input colour space must be defined in sufficient detail that it can be transformed into CIE coordinates, for example by specifying the gamma and chromaticity coordinates of each primary. Two common examples used in this appendix are a linear (gamma = 1) RGB space, and YCbCr with gamma = 2.2.

VII.2.2 Colour transformations

The first step in the process is the conversion of both image sequences to the YOZ colour space. This is a colour space we have previously used in modelling perceptual errors in still image compression. The three components of this space are Y (CIE luminance in candelas/m^2), O, a colour-opponent channel given by $O = \{X = 0.47;\ Y = -0.37;\ Z = -0.1\}$ and a blue channel given by the CIE Z coordinate. Transformation to the YOZ space typically involves:

1) a gamma transformation, followed by;

2) a linear colour transformation.

These operations do not alter the dimensionality of the input.

VII.2.3 Blocked DCT

At this point a blocked DCT is applied to each frame in each colour channel. The dimensions of the result are {s, f, c, by, bx, v, u}, where by and bx are the number of blocks in vertical and horizontal directions, and where now v = u = 8.

VII.2.4 Local contrast

The DCT coefficients are converted to units of local contrast in the following way. First we extract the DC coefficients from all blocks. These are then time filtered, using a first-order, low-pass, IIR filter with a gain of 1 and a time constant of τ_1. The DCT coefficients are then divided by the filtered DC coefficients on a block-by-block basis. The Y and Z blocks are divided by Y and Z DC coefficients; the O is divided by the Y DC. In each case, a very small constant is added to the divisor to prevent division by zero. Finally, the quotients are adjusted by the relative magnitudes of their coefficients corresponding to a unit contrast basis function. These operations convert each DCT

coefficient to a number between −1 and 1, that expresses the amplitude of the corresponding basis function as a fraction of the average luminance in that block.

The DC coefficients themselves are converted in a similar fashion: the mean DC over the entire frame is subtracted, and the result is divided by that mean.

VII.2.5 Temporal filtering

Both sequences are then subjected to temporal filtering. The temporal filter is a second-order IIR filter, as described above in the fit of the dynamic DCT noise data. Use of an IIR filter minimizes the number of frames of data that must be retained in memory. For even greater simplicity, a first order filter may be used.

VII.2.6 JND conversion

The DCT coefficients, now expressed in local contrast form, are now converted to just-noticeable-differences (JNDs) by dividing their respective spatial thresholds. These thresholds are first multiplied by a spatial summation factor *s*, whose purpose and estimation are described below. The thresholds for the two colour channels are either derived from the luminance thresholds[3] or based on additional chromatic thresholds. After conversion to JNDs, the coefficients of the two sequences are subtracted to produce a *difference sequence*.

VII.2.7 Contrast masking

Contrast masking is accomplished by first constructing a *masking sequence*. This begins as the reference sequence, after JND conversion. This sequence is rectified, and then time-filtered by a first-order, low-pass, discrete IIR filter, with a gain of g_1 and a time constant of τ_2. These values are then raised to a power *m*, any values less than 1 are replaced by 1, and the result is used to divide the difference sequence. This process mimics the traditional contrast masking result in which contrasts below threshold have no masking effect, and that above threshold the effect rises as the mth power of mask contrast in JNDs.

VII.2.8 Minkowski pooling

The dimensions of the result at this point are {*f, c, by, bx, v, u*}, where, to remind, *f* is frames, *c* is colour channels, *by* and *bx* are the number of blocks in vertical and horizontal directions, and where $v = u$ are the vertical and horizontal frequencies. These elementary errors may then be combined over various dimensions, or all dimensions, to yield summary measures of visual error. This summation is done using a Minkowski metric:

$$J_x = M\left(j_{f,c,by,bx,y,x}, \beta\right) = \left(\sum_x \left| j_{f,c,by,bx,y,x} \right|^{\beta} \right)^{\frac{1}{\beta}} \qquad \text{(VII-1)}$$

In this equation we have indicated summation over all six dimensions, but any subset of these dimensions may be considered as well. A virtue of the Minkowski formulation is that it may be nested. For example, we may first sum over only the colour dimension *(c)*, and then these results may subsequently be summed over, for example, the block dimensions (*by* and *bx*).

VII.3 Evaluation

We have evaluated the performance of the DVQ video quality metric by comparing its predictions to judgments of impairment made by 25 human observers viewing five reference sequences as processed by 12 HRCs. The DVQ metric performs considerably better than models based on simple bit-rate or root mean square (rms) error. The quality of the predictions suggests the metric may be useful in practical applications. More recently we submitted our algorithm to the VQEG (Video Quality Experts Group) testing project. DVQ performed quite well over a wide range of HRC

subsets. It performed particularly well in the high quality regime, with a Rank Correlation of 0.72. Two of the tested conditions, multi-generation 1/2 inch professional record/play cycles and transmission errors, are outside the scope of our model. With the removal of these HRCs, the Spearman Rank Correlation was 0.82.

VII.4 References

[1] WATSON (A.B.): Toward a perceptual video quality metric in Human Vision, *Visual Processing, and Digital Display VIII*, San Jose, CA: SPIE, Bellingham, WA, 1998.

[2] WATSON (A.B.), *et al*.: Design and performance of a digital video quality metric in Human Vision, *Visual Processing, and Digital Display IX.* San Jose, CA: SPIE, Bellingham, WA, 1999.

[3] WATSON (A.B.), HU (J.), McGOWAN (J.F.), III: *DVQ:* A digital video quality metric based on human vision, *Journal of Electronic Imaging*, 2000. in press.

APPENDIX VIII

KPN/Swisscom CT

VIII.1 Introduction

In PVQM, the physical signals of input and output of the device under test (e.g. a codec, or a transmission chain) are mapped onto psychophysical representations (see Figure VIII.1) that match as close as possible the internal representations of the audio/video signals (representations inside our head). The quality of the device under test is judged on the basis of differences in the internal representation. In PVQM the internal representation, from which the quality is derived, is such that both, spatial and temporal distortions, are covered by the measurement method.

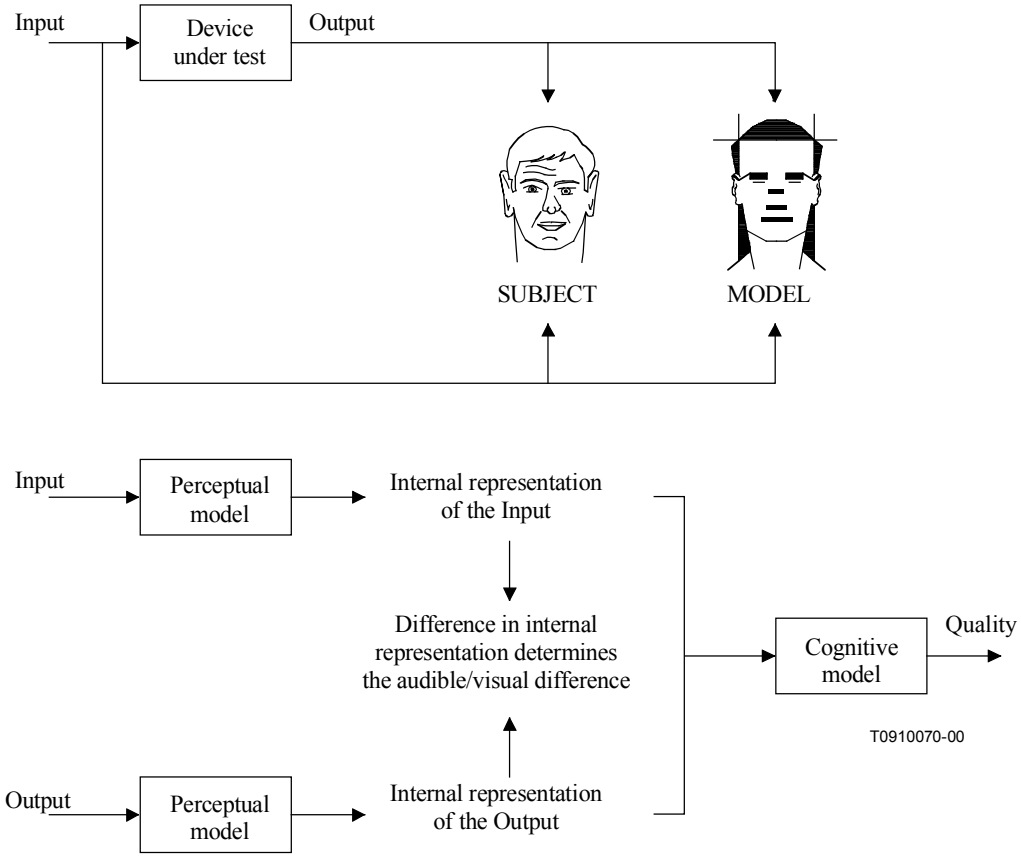


Figure VIII.1/J.144

Overview of the basic philosophy used in the development of PVQM. A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output of the device under test (e.g. a video codec) with the input, using any video signal.

In order to be able to use PVQM in practical situations a spatio-temporal-luminance alignment is included into the algorithm. It is well known that global changes in the brightness and contrast only have a limited impact on the subjectively perceived quality, especially when compared to the impact of distortions like blockiness. This effect is quantified in PVQM by using a special brightness/contrast adaptation of the distorted video sequence. Furthermore it is trivial that one can only calculate a relevant measure of distortion if one knows which parts of the input and output signal have to be compared. Therefore PVQM uses a kind of block matching spatio-temporal alignment procedure before the actual measurements are carried out.

The spatial luminance analysis part is based on edge detection of the Y signal, while the temporal part is based on difference frames analysis of the Y signal. It is well known that the Human Visual System (HVS) is much more sensitive to the sharpness of the luminance component than that of the chrominance components. Furthermore, the HVS has a contrast sensitivity function that decreases at high spatial frequencies. These basics of the HVS are reflected in the first pass of the PVQM algorithm that provides a first order approximation to the contrast sensitivity functions of the luminance and chrominance signals.

In the second step the edginess of the luminance Y is computed as a signal representation that contains the most important aspects of the picture. This edginess is computed by calculating the local gradient of the luminance signal in each frame. The relative error in the edginess between input and output video is aggregated over space and time using Lebesgue p-measures.

In the third step the chrominance error is computed as a weighted average of the colour error of both the Cb and Cr components normalized on the local saturation with a dominance of the Cr component.

In the last step the three different indicators are mapped onto a single quality indicator, using a simple multiple linear regression, which correlates well the subjectively perceived overall video quality of the sequence. The method has been validated at KPN Research using a wide variety of databases containing both codec (MPEG, ITU-T H.263, etc.) and artificially generated distortions. On all relevant databases the correlation between the objective PVQM values and subjective Mean Opinion Scores is above 0.9.

VIII.2 References

[1] BEERENDS (J.G.), HEKSTRA (A.P.): Objective measurement of video quality, *ITU-T Study Group 12, Document COM 12-7*, February 1997.

APPENDIX IX

NTIA

Introduction

This appendix provides full disclosure of the algorithm used to compute a Video Quality Metric (VQM) that accurately tracks subjective quality judgments of video scenes. This version of VQM contains several improvements over the model that was submitted to the Video Quality Experts Group (VQEG). These improvements were developed before the VQEG subjective data became available [1]. In addition to providing technology-independent perception-based estimates of subjective quality, the VQM has low computational complexity and can be used for continuous real-time in-service quality monitoring applications. Results are presented that compare the VQM with mean opinion scores from nine different double-stimulus subjective tests that span many different scenes, video systems, and coding technologies. Seven of these data sets contain mostly video scenes from contribution-quality and distribution-quality broadcast applications (> 1.5 Mbit/s) while two of these data sets contain mostly video scenes from multimedia applications (< 1.5 Mbit/s).

IX.1 Description of VQM algorithm

The VQM consists of a linear combination of four parameters that have been optimized for the standard viewing distance of six times picture height. Three parameters are extracted from spatial gradients of the luminance component (Y) of ITU-R BT.601 [2] input and output video streams while one parameter is extracted from the vector formed by the chrominance components (C_B, C_R).

The sampled input and output video streams are assumed to have been calibrated before the processes described herein are performed. This calibration includes compensation for system gain and level offset, as well as spatial and temporal registration of the images.

IX.2 Spatial gradient parameters

An overview of the algorithm used to extract the spatial gradient parameters is given in Figure IX.1. The Y components of the input and output video streams are processed using horizontal and vertical edge enhancement filters. Next, these processed video streams are divided into spatial-temporal (S-T) regions from which features, or summary statistics, are extracted that quantify the spatial activity as a function of angular orientation. Then, these features are clipped at the lower end to emulate perceptibility thresholds. Next, distortions in video quality due to gains and losses in the feature values are calculated for each S-T region by comparing their input and output values using functional relationships that emulate visual masking of impairments. These distortions are then

pooled across space (spatial collapsing) and time (temporal collapsing) to produce quality parameters for a video clip that is nominally 5 to 10 seconds in duration.

The edge enhancement filters, the S-T region size, and the perceptibility thresholds that are presented here were optimized based on correlation with perceptual distortions at six times picture height.

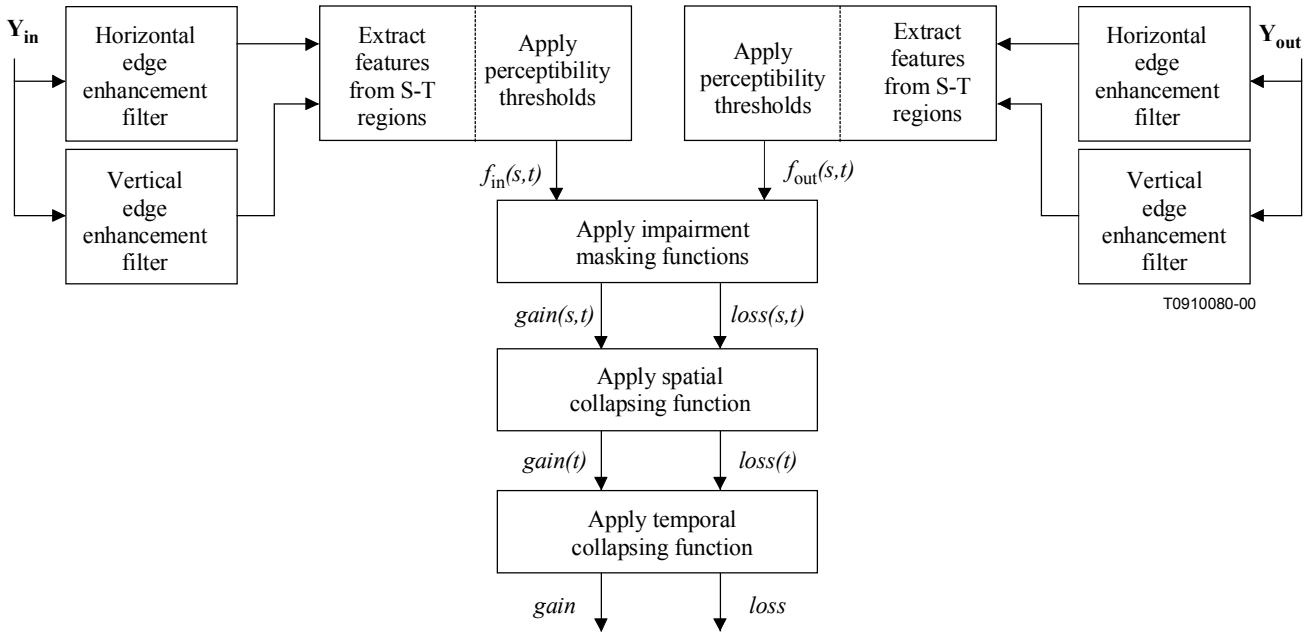


Figure IX.1/J.144 – Overview of algorithm used to extract spatial gradient parameters

IX.3 Edge enhancement filters

The input and output video *frames* are first processed with horizontal and vertical edge enhancement filters that enhance edges while reducing noise. The two filters shown in Figure IX.2 are applied separately, one to enhance horizontal pixel differences while smoothing vertically (left filter), and the other to enhance vertical pixel differences while smoothing horizontally (right filter).

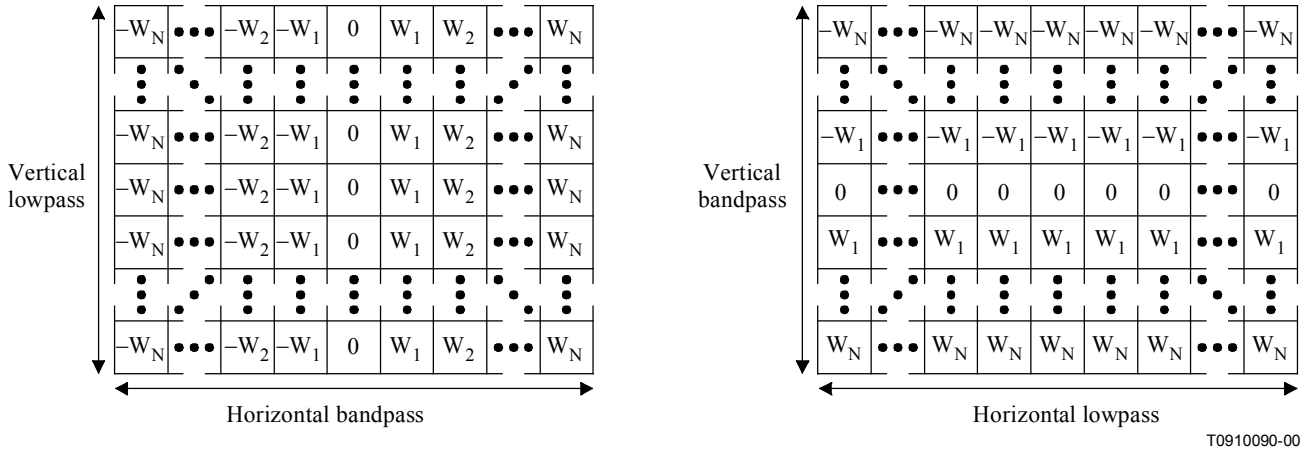


Figure IX.2/J.144 – Edge enhancement filters

The two filters are transposes of each other, have size 13 × 13, and have filter weights given by:

$$w_x = k * \left(\frac{x}{c}\right) * \exp\left\{-\frac{1}{2}\left(\frac{x}{c}\right)^2\right\}$$

where x is the pixel displacement from the centre of the filter (0, 1, 2, …, N), c is a constant that sets the width of the bandpass filter, and k is a normalization constant selected such that each filter would produce the same gain as a true Sobel filter. The optimal amount of horizontal bandpass filtering for a viewing distance of six times picture height was found to be given by the $c = 2$ filter, which has a peak response at about 4.5 cycles/degree. The bandpass filter weights that were used are given by:

[–0.0052625, –0.0173446, –0.0427401, –0.0768961, –0.0957739, –0.0696751, 0, 0.0696751, 0.0957739, 0.0768961, 0.0427401, 0.0173446, 0.0052625].

Notice that the filters in Figure IX.2 have a flat lowpass response. A flat lowpass response produced the best quality estimate and has the added advantage of being computationally efficient (e.g. for the left filter in Figure IX.2, one merely has to sum the pixels in a column and multiply once by the weight).

IX.4 S-T region size

The horizontal and vertical edge enhanced input and output video streams are each divided into localized S-T regions. Figure IX.3 gives the S-T region size (8 horizontal pixels × 8 vertical lines × 6 video frames) that achieved the maximum correlation with subjective ratings. It should be noted, however, that the correlation was found to worsen *slowly* as one moves away from the optimum point. Horizontal and vertical widths up to 32 pixels or lines and temporal widths up to 30 frames can be used with satisfactory results, giving the objective measurement system designer considerable flexibility in adapting the techniques presented here to different S-T region sizes.

Features are extracted from each S-T region by calculating summary statistics over the S-T region. A detailed description of the features that are extracted is given in IX.5.

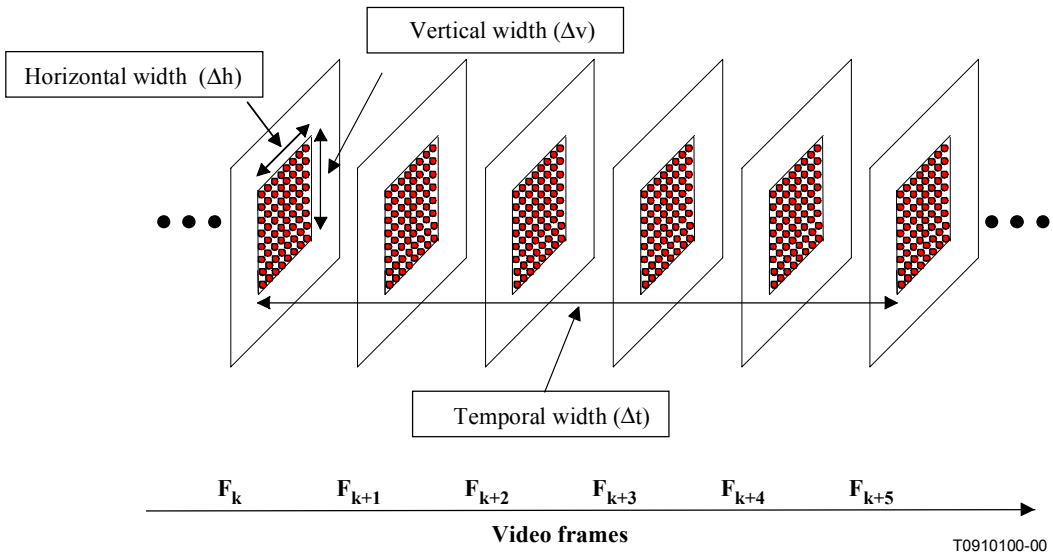


Figure IX.3/J.144 – Optimal spatial-temporal (S-T) region size for extracting features

IX.5 Description of features

This clause describes the extraction of two spatial activity features from S-T regions of the edge enhanced input and output video streams from IX.4. The filter shown in Figure IX.2 (left) enhances spatial gradients in the horizontal (H) direction while the transposes of these filters enhance spatial gradients in the vertical (V) direction. The response at each pixel from the H and V filters can be plotted on a two dimensional diagram such as the one shown in Figure IX.4 with the H filter response forming the abscissa value and the V filter response forming the ordinate value. For a given image pixel located at row i, column j, and time t, the H and V filter responses will be denoted as $H(i, j, t)$ and $V(i, j, t)$, respectively. These responses can be converted into polar coordinates (R, θ) using the relationships:

$$R(i,j,t) = \sqrt{H(i,j,t)^2 + V(i,j,t)^2}$$

and:

$$\theta(i,j,t) = \tan^{-1}\left[\frac{V(i,j,t)}{H(i,j,t)}\right]$$

The first feature, f_1, is computed simply as standard deviation (*stdev*) over the S-T region of the $R(i, j, t)$ samples, and then clipped at the perceptibility threshold of P (i.e. if the results of the *stdev* calculation falls below P, f_1 is set equal to P), namely:

$$f_1 = \left\{ stdev\left[R(i,j,t)\right]\right\}\Big|_P : \; i,j,t \in \{\text{S-T Region}\}$$

This feature is sensitive to changes in the overall amount of spatial activity within a given S-T region. For instance, localized blurring produces a reduction in the amount of spatial activity whereas noise produces an increase. The recommended threshold P for this feature is 12.

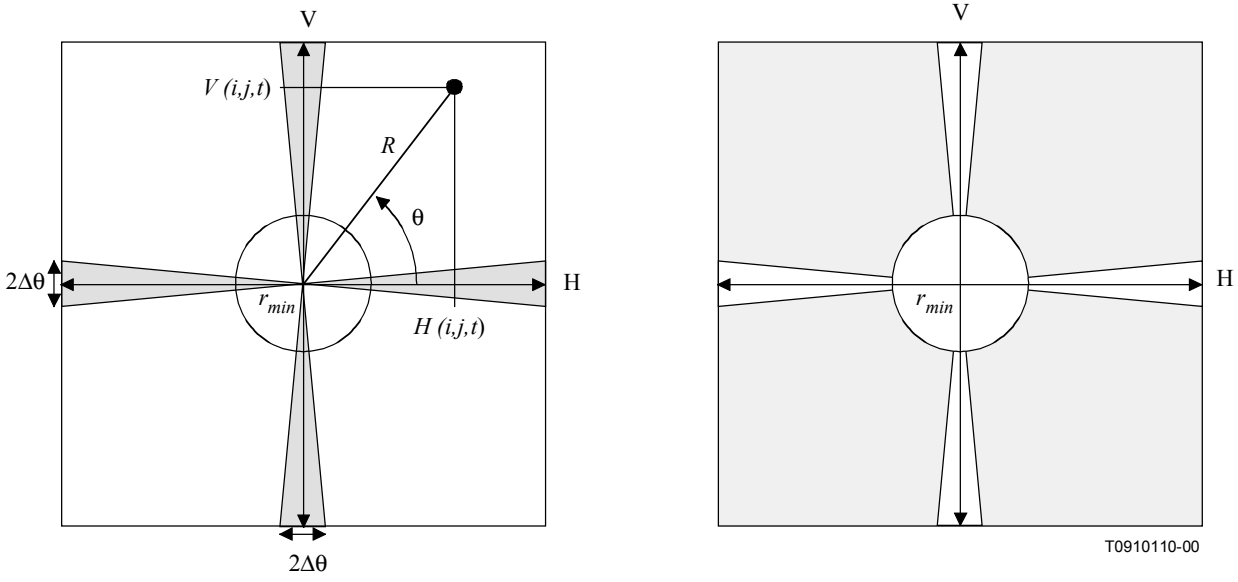


Figure IX.4/J.144 – Division of horizontal (H) and vertical (V) spatial activity into *HV* (left) and $\overline{HV}$ (right) distributions

The second feature, f_2, is sensitive to changes in the angular distribution, or orientation, of spatial activity. Complementary images are computed with the shaded spatial gradient distributions shown in Figure IX.4. The image with horizontal and vertical gradients, denoted as *HV*, contains the

$R(i, j, t)$ pixels that are horizontal or vertical edges (pixels that are diagonal edges are zeroed). The image with the diagonal gradients, denoted as $\overline{HV}$, contains the $R(i, j, t)$ pixels that are diagonal edges (pixels that are horizontal or vertical edges are zeroed). Gradient magnitudes $R(i, j, t)$ less than $r_{\min}$ are zeroed in both images to assure accurate θ computations. Pixels in HV and $\overline{HV}$ can be represented mathematically as:

$$HV(i,j,t) = \begin{cases} R(i,j,t) \quad if\ R(i,j,t) \geq r_{\min}\ and\ m\frac{\pi}{2} - \Delta\theta < \theta(i,j,t) < m\frac{\pi}{2} + \Delta\theta \quad (m = 0,1,2,3) \\ \\ 0 \quad otherwise \end{cases}$$

and:

$$\overline{HV}(i,j,t) = \begin{cases} R(i,j,t) \quad if\ R(i,j,t) \geq r_{\min}\ and\ m\frac{\pi}{2} + \Delta\theta \leq \theta(i,j,t) \leq (m+1)\frac{\pi}{2} - \Delta\theta \quad (m = 0,1,2,3) \\ \\ 0 \quad otherwise \end{cases}$$

where:

$$i, j, t \in \{\text{S-T Region}\}$$

For the computation of HV and $\overline{HV}$ above, the recommended value for $r_{\min}$ is 20 and the recommended value for $\Delta\theta$ is 0.05236 radians. Feature f_2 for one S-T region is then given by the ratio of the mean of HV to the mean of $\overline{HV}$, where these resultant means are clipped at their perceptibility thresholds P, namely:

$$f_2 = \{mean[HV(i,j,t)]\}\big|_P \Big/ \{mean[\overline{HV}(i,j,t)]\}\big|_P$$

The recommended perceptibility threshold P for the mean of HV and $\overline{HV}$ is 3. The f_2 feature is sensitive to changes in the angular distribution of spatial activity within a given S-T region. For example, if horizontal and vertical edges suffer more blurring than diagonal edges, f_2 of the output will be less than f_2 of the input. On the other hand, if erroneous horizontal or vertical edges are introduced, say in the form of blocking or tiling distortions, then f_2 of the output will be greater than f_2 of the input. The f_2 feature thus provides a simple means to include variations in the sensitivity of the human visual system with respect to angular orientation.

For the following discussion, an input feature stream will be denoted as $f_{in}(s, t)$ and the corresponding output feature stream will be denoted as $f_{out}(s, t)$, where s and t are indices that denote the spatial and temporal positions, respectively, of the S-T region within the calibrated input and output video streams.

IX.6 Impairment masking functions

Next, the perceptual impairment at each S-T region is calculated using a function that models visual masking of impairments. Gain and loss must be examined separately, since they produce fundamentally different effects on quality perception (e.g. loss of spatial activity due to blurring and gain of spatial activity due to noise or blocking). Of the many comparison functions that we have evaluated, two have consistently produced the best correlation to subjective ratings. These

comparison functions model the perceptibility of spatial or temporal impairments. For a given S-T region, gain and loss distortions are computed using:

$$gain(s,t) = pp\left\{\log_{10}\left[\frac{f_{out}(s,t)}{f_{in}(s,t)}\right]\right\}$$

and:

$$loss(s,t) = np\left\{\frac{f_{out}(s,t) - f_{in}(s,t)}{f_{in}(s,t)}\right\}$$

where *pp* is the positive part operator (i.e. negative values are replaced with zero), and *np* is the negative part operator (i.e. positive values are replaced with zero). These visual masking functions imply that impairment perception is inversely proportional to the amount of localized spatial or temporal activity in the input scene. In other words, spatial impairments become less visible as the spatial activity in the input scene is increased (i.e. spatial masking), and temporal impairments become less visible as the temporal activity in the input scene is increased (i.e. temporal masking). While the logarithmic and ratio comparison functions behave very similarly, the logarithmic function tends to be slightly more advantageous for gains while the ratio function tends to be slightly more advantageous for losses.

IX.7 Spatial collapsing function

Next, impairments from S-T regions with the same time index *t* are pooled using a spatial collapsing function. Extensive investigation has revealed that the optimal spatial collapsing functions normally involve some form of worst-case processing. This is because localized impairments tend to draw the focus of the viewer, making the worst part of the picture the predominant factor in the subjective quality decision. The spatial collapsing function is computed at each temporal index *t* as the average of the worst 5% of the measured distortions over the spatial index *s* (denoted as $worst\_5\%_{space}$). This amounts to rank sorting the gain distortions at each temporal index *t* and averaging those distortions that are above the 95% threshold. Similarly, the loss distortions are rank sorted at each temporal index *t*, but the average of those distortions that are below the 5% threshold is used (since losses are negative). Applying the $worst\_5\%_{space}$ function produces a time history of the gain and loss samples, namely *gain*(*t*) and *loss*(*t*), which must then be temporally collapsed.

IX.8 Temporal collapsing functions

Finally, the results from the spatial collapsing function are pooled using a temporal collapsing function to produce an objective parameter for the video clip, which is nominally 5 to 10 seconds in length. Viewers seem to use several temporal collapsing functions when subjectively rating video clips that are from 9 to 10 seconds in length. One temporal collapsing function is indicative of the average quality level of the clip while the other is indicative of the worst transient quality of the clip (e.g. digital transmission errors normally cause a 1 to 2 second disturbance in the output video).

The mean over time (denoted as $mean_{time}$) seems to be indicative of the average quality that is observed during the time period. For worst transient quality, the 10% level over time for loss parameters (denoted as $10\%_{time}$) and the 90% level over time for gain parameters (denoted as $90\%_{time}$) seems to capture most of the subjective impact (i.e. the time history samples of the loss parameter are rank sorted and the 10% level is used; the time history samples of the gain parameter are rank sorted and the 90% level is used). Further research needs to be performed to optimize these temporal collapsing functions.

IX.9 Three spatial gradient parameters

The three spatial gradient parameters that are used to compute VQM are given by:

f_1_loss (use temporal collapsing function $10\%_{time}$);

f_2_loss (use temporal collapsing function $mean_{time}$); and

f_2_gain (use temporal collapsing function $mean_{time}$).

Here, the f_1 and f_2 features are described in IX.5, the loss and gain functions are given in IX.6, the spatial collapsing function is given in IX.7, and the temporal collapsing functions are given in IX.8.

IX.10 Chrominance parameter

This clause presents a single chrominance distortion parameter that is included in the computation of VQM. For a given image pixel located at row i, column j, and time t, let $C_B(i, j, t)$ and $C_R(i, j, t)$ represent ITU-R BT.601 C_B and C_R values. The components of a two-dimensional chrominance feature vector, $\underline{f}_C$, are computed simply as the mean (*mean*) over the S-T region of the $C_B(i, j, t)$ and $C_R(i, j, t)$ samples, respectively, giving more perceptual weight to the C_R component:

$$\underline{f}_C(s,t) = \left(mean[C_B(i,j,t)], W_R * mean[C_R(i,j,t)]\right): i,j,t \in \{\text{S-T Region}\}, \text{ and } W_R = 1.5.$$

The recommended S-T region size is 8 horizontal pixels × 8 vertical lines × 1 video frames (actually 4 horizontal C_B and C_R pixels, since these signals are sub-sampled by two in ITU-R BT.601). Chrominance distortion for each S-T region, denoted as d_C (s, t), where s and t are indices that denote the spatial and temporal positions, respectively, of the S-T region within the calibrated input and output video streams, is computed as the Euclidean distance between the input and output chrominance feature vectors $\underline{f}_{Cin}$ and $\underline{f}_{Cout}$ shown by the dashed line in Figure IX.5, namely:

$$d_C(s,t) = \left\| \underline{f}_{Cout}(s,t) - \underline{f}_{Cin}(s,t) \right\|$$

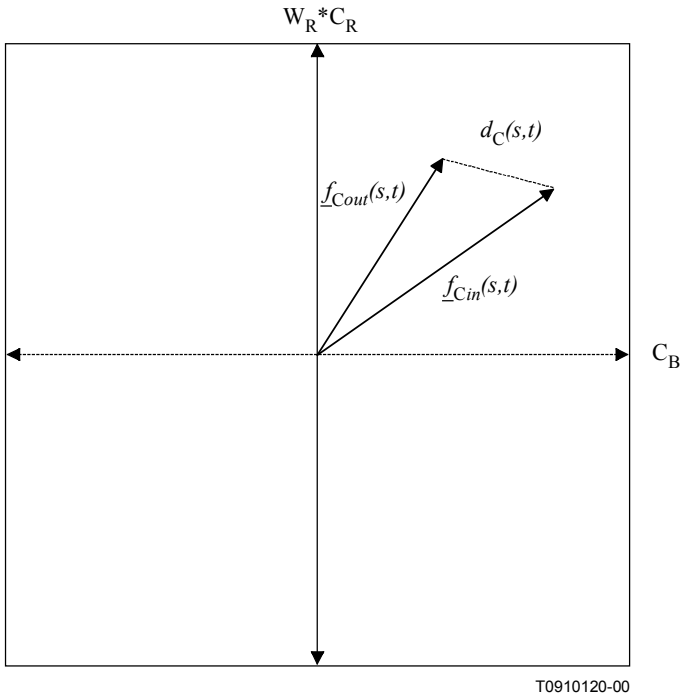


Figure IX.5/J.144 – Computation of chrominance distortion d_C (s, t) for a S-T region

The optimal spatial collapsing function for d_C (s, t) is the standard deviation over space (denoted as $stdev_{space}$), which is similar to the $worst\_5\%_{space}$ function given previously. The optimal temporal collapsing function is the 10% level over time (denoted as $10\%_{time}$), which represents the level of distortion that is nearly always present. The chrominance distortion value after spatial and temporal collapsing is clipped at a perceptibility threshold $P = 0.8$ and then this clipping value is subtracted to produce the d_C metric. In summary, the chrominance distortion parameter d_C is given by:

$$d_C = \left\{10\%_{\text{time}}\left[stdev_{\text{space}}\left(d_C(s,t)\right)\right]\right\}\Big|_P - P$$

IX.11 VQM computation

VQM is computed as:

$$\text{VQM} = -0.3609*f_1\_loss + 0.5031*(f_2\_loss)^2 + 0.1390*f_2\_gain + 0.0295*d_C$$

The square on the f_2_loss parameter is necessary to linearize this parameter response. The f_1_loss parameter requires a negative multiplier since this parameter is always less than or equal to zero (the f_2_loss parameter is also always less than or equal to zero but the square of this parameter is being used in the VQM calculation). The f_2_gain and d_C parameters are always greater than or equal to zero. VQM computed in this manner will have values greater than or equal to zero and have a nominal maximum value of one. VQM may occasionally exceed one for video scenes that are extremely distorted.

IX.12 Description of subjective data sets

The nine subjective experiments were collected from 1992 to 1999. All of the data sets were conducted in accordance with the most recent version of ITU-R BT.500-9 [3] that was available when the experiment was performed. All of the data sets used scenes from 9 to 10 seconds in duration and used double stimulus viewing (viewers saw both the original and impaired sequences). For brevity, only a summary of each subjective experiment is given here. The reader is directed to the accompanying references for more complete descriptions.

Data Set One [4, 5]

A panel of 48 viewers rated a total of 132 video clips that were generated by random and deterministic pairing of 36 test scenes with 27 video systems. The 36 test scenes contained widely varying amounts of spatial and temporal information. The 27 video systems included digital video compression systems operating at bit rates from 56 kbit/s to 45 Mbit/s with controlled error rates, NTSC encode/decode cycles, VHS and S-VHS record/play cycles, and VHF transmission. Viewers were shown the original version first, then the degraded version, and asked to rate the difference in perceived quality using the 5-point impairment scale (imperceptible, perceptible but not annoying, slightly annoying, annoying, very annoying).

Data Set Two [6, 7]

Viewer panels comprising a total of 30 viewers from three different laboratories rated 600 video clips that were generated by pairing 25 test scenes with 24 video systems. The 25 test scenes included scenes from five categories:

1) one person, mainly head and shoulders;
2) one person with graphics and/or more detail;
3) more than one person;
4) graphics with pointing; and
5) high object and/or camera motion.

The 24 video systems included proprietary and standardized video teleconferencing systems operating at bit rates from 56 kbit/s to 1.5 Mbit/s with controlled error rates, one 45 Mbit/s codec, and VHS record/play cycle. The subjective test procedure was the same as data set one.

Data Set Three [8]

A panel of 32 viewers rated the difference in quality between input scenes with controlled amounts of added noise and the resultant MPEG-2 compression-processed output. The data set contains a total of 105 video clips that were generated by pairing seven test scenes at three different noise levels with five MPEG-2 video systems. The seven test scenes were chosen to span a range of spatial detail, motion, brightness, and contrast. The five MPEG-2 video systems operated at bit rates from 1.8 Mbit/s to 13.9 Mbit/s. Viewers were shown the input and processed output in randomized A/B ordering and asked to rate the quality of B using A as a reference. The experiment utilized a seven-point comparison scale (B much worse than A, B worse than A, B slightly worse than A, B the same as A, B slightly better than A, B better than A, B much better than A).

Data Set Four [9]

A panel of 32 viewers rated a total of 112 video clips that were generated by pairing subgroups of eight scenes each (total number of scenes in the test was 16) with 14 different video systems. The 16 test scenes spanned a wide range of spatial detail, motion, brightness, and contrast and included scene material from movies, sports, nature, and classical ITU-R BT.601 test scenes. The 14 video systems included MPEG-2 systems operated at bit rates from 2 Mbit/s to 36 Mbit/s with controlled error rates, multi-generation MPEG-2, multi-generation 1/2 inch professional record/play cycles, VHS, and video teleconferencing systems operating at bit rates from 768 kbit/s to 1.5 Mbit/s. The subjective test procedure was the same as data set three.

Data Set Five [9]

A panel of 32 viewers rated a total of 42 video clips that were generated by pairing subgroups of six scenes each (total number of scenes in the test was 12) with seven different MPEG-2 systems. The 12 test scenes included sports material and classical ITU-R BT.601 test scenes. The nine MPEG-2 systems operated at bit rates from 2 Mbit/s to 8 Mbit/s. The subjective test procedure was the same as data set three.

Data Sets Six to Nine [10]

Four data sets (525-line high quality, 525-line low quality, 625-line high quality, 625-line low quality), each of 90 video clips were generated by pairing ten scenes with nine video systems. For each data set, a total of 60 to 80 viewers from four different laboratories (i.e. 15 to 20 viewers per laboratory) rated subjective quality using the double stimulus continuous quality scale (DSCQS). The twenty different test scenes (ten for 525-line, ten for 625-line) included sports material, classical ITU-R BT.601 test scenes, moving graphics, and stills. The video systems included MPEG-2 systems operating at bit rates from 2 Mbit/s to 50 Mbit/s, video teleconferencing systems operating at 768 kbit/s and 1.5 Mbit/s, some systems with digital transmission errors, multi-generation MPEG-2, multi-generation 1/2 inch professional record/play cycles, where composite and/or component signal formats were used.

IX.13 Results

The Pearson linear correlation coefficient between VQM and each of the individual subjective data sets is given in Table IX.1. VQM achieved an average Pearson correlation coefficient of 0.90.

Figure IX.6 shows the scatter plot of the subjective quality judgments from all nine subjective data sets versus VQM. In this scatter plot, the subjective mean opinion scores of the nine data sets have been mapped to fall between zero and one. The Pearson linear correlation coefficient between the subjective scores and VQM in the scatter plot is 0.94 (this correlation coefficient is higher than the average of the Table IX.1 values since the range of quality in the combined data set is larger than in

any of the individual data sets). The majority of the outliers in the scatter plot are from systems that have some form of time varying noise in the output (e.g. VHF transmission, multi-generation 1/2 inch professional record/play cycles, composite encode/decode cycles, digital transmission errors which produce transient error blocks). Future improvements to VQM are being developed that will include perception-based parameters to measure these time varying noise effects. One promising area of research is quality parameters derived from temporal gradient information (i.e. temporal activity).

Table IX.1/J.144 – Pearson linear correlation coefficient for VQM

Data Set	Pearson linear correlation coefficient
One	0.92
Two	0.90
Three	0.94
Four	0.88
Five	0.91
Six to Nine Combined	0.86

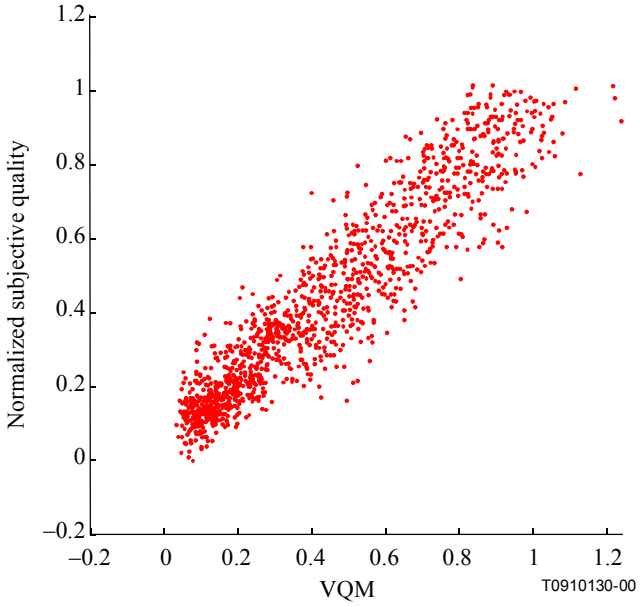


Figure IX.6/J.144 – Scatter plot of subjective quality vs VQM for nine data sets

IX.14 References

[1] WOLF (Stephen), PINSON (Margaret H.): Spatial-temporal distortion metrics for in-service quality monitoring on any digital video system, *SPIE International Symposium on Voice, Video, and Data Communications*, Boston, MA, 11-22 September 1999.

[2] ITU-R BT.601-5 (1995), *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*.

[3] ITU-R BT.500-9 (1998), *Methodology for subjective assessment of the quality of television pictures*.

[4] VORAN (Stephen), WOLF (Stephen): The Development and evaluation of an objective video quality assessment system that emulates human viewing panels, *International Broadcasting Convention (IBC)*, July 1992.

[5] WEBSTER (Arthur A.), JONES (Coleen T.), PINSON (Margaret H.), VORAN (Stephen D.), WOLF (Stephen): An objective video quality assessment system based on human perception, *Human Vision, Visual Processing, and Digital Display IV, Proceedings of the SPIE*, Vol. 1913, February 1993.

[6] ANSI Accredited Standards Working Group T1A1 contribution number T1A1.5/94-118R1, "*Subjective test plan (tenth and final draft)*", Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC, 3 October 1993.

[7] ANSI T1.801.01 (1995), *Digital Transport of Video Teleconferencing/Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment*.

[8] FENIMORE (Charles), *et al.*: Perceptual effects of noise in digital video compression, *SMPTE Journal*, Vol. 109, pp. 178-186, March 2000.

[9] WOLF (S.), PINSON (M.): In-service performance metrics for MPEG-2 video systems, *Made to Measure 98 – Measurement Techniques of the Digital Age Technical Seminar, technical conference jointly sponsored by the International Academy of Broadcasting (IAB), ITU, and the Technical University of Braunschweig (TUB)*, Montreux, Switzerland, 12-13 November 1998.

[10] Final report from the video quality experts group (VQEG) on the validation of objective models of video quality assessment, *VQEG meeting number 4*, Ottawa, Canada, March 2000.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
Series B	Means of expression: definitions, symbols, classification
Series C	General telecommunication statistics
Series D	General tariff principles
Series E	Overall network operation, telephone service, service operation and human factors
Series F	Non-telephone telecommunication services
Series G	Transmission systems and media, digital systems and networks
Series H	Audiovisual and multimedia systems
Series I	Integrated services digital network
Series J	**Cable networks and transmission of television, sound programme and other multimedia signals**
Series K	Protection against interference
Series L	Construction, installation and protection of cables and other elements of outside plant
Series M	TMN and network maintenance: international transmission systems, telephone circuits, telegraphy, facsimile and leased circuits
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Telephone transmission quality, telephone installations, local line networks
Series Q	Switching and signalling
Series R	Telegraph transmission
Series S	Telegraph services terminal equipment
Series T	Terminals for telematic services
Series U	Telegraph switching
Series V	Data communication over the telephone network
Series X	Data networks and open system communications
Series Y	Global information infrastructure and Internet protocol aspects
Series Z	Languages and general software aspects for telecommunication systems

Printed in Switzerland
Geneva, 2002