

International Telecommunication Union

ITU-T

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

J.246

(08/2008)

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

Measurement of the quality of service

**Perceptual visual quality measurement
techniques for multimedia services over digital
cable television networks in the presence of a
reduced bandwidth reference**

Recommendation ITU-T J.246



Recommendation ITU-T J.246

Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference¹

Summary

The term multimedia as defined in Recommendation ITU-T J.148 is the combination of multiple forms of media such as: video, audio, text, graphics, fax, and telephony in the communication of information. A three stage approach has been adopted to recommending objective assessment methods for multimedia. The first two stages will identify perceptual quality tools appropriate for measuring video and audio individually. The third stage will identify objective assessment methods for the combined audiovisual media. This Recommendation contains the first stage video only used in multimedia applications.

Recommendation ITU-T J.246 provides guidelines on the selection of appropriate objective perceptual video quality measurement methods when a reduced reference signal is available. The following are example applications that can use this Recommendation:

- 1) Internet multimedia streaming
- 2) Video telephony and conferencing over cable and other networks
- 3) Progressive video television streams viewed on LCD monitors over cable networks including those transmitted over the Internet using Internet Protocol. (VGA was the maximum resolution in the validation test.)
- 4) Mobile video streaming over telecommunications networks
- 5) Some forms of IPTV video payloads (VGA was the maximum resolution in this validation test.)
- 6) Video quality monitoring at the receiver when side-channels are available.

Source

Recommendation ITU-T J.246 was approved on 13 August 2008 by ITU-T Study Group 9 (2005-2008) under Recommendation ITU-T A.8 procedures.

¹ The name was changed to reflect the fact that this Recommendation covers video only.

FOREWORD

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

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

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

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

NOTE

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

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

INTELLECTUAL PROPERTY RIGHTS

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

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

© ITU 2010

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

CONTENTS

	Page
1 Scope	1
1.1 Application	2
1.2 Limitations.....	2
2 References.....	3
2.1 Normative references.....	3
2.2 Informative references.....	3
3 Definitions	4
3.1 Terms defined elsewhere.....	4
3.2 Terms defined in this Recommendation.....	4
4 Abbreviations and acronyms	5
5 Conventions.....	6
6 Description of the reduced reference measurement method.....	6
7 Findings of the Video Quality Experts Group (VQEG)	7
Annex A – Yonsei University Reduced Reference Method	10
A.1 Introduction	10
A.2 The EPSNR Reduced Reference Models	10
A.3 Conclusions	22
Appendix I – Optimal side-channel bandwidths.....	23
Appendix II – Excerpts from the Synopsis from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment, phase I.....	26
II.1 Introduction	26
II.2 Model performance evaluation techniques.....	27
II.3 RR model performance.....	27
II.4 Data analysis executed by ILG.....	29
Appendix III – Equations for Model Evaluation Metrics	30
III.1 Evaluation Metrics.....	30
Bibliography.....	33

Recommendation ITU-T J.246

Perceptual visual quality measurement techniques for multimedia services over digital cable television networks in the presence of a reduced bandwidth reference

1 Scope

This Recommendation provides guidelines and recommendations on the selection of appropriate perceptual video quality measurement equipment for use in multimedia applications when the reduced reference measurement method can be used.

The reduced reference measurement method can be used when features extracted from the unimpaired reference video signal are 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. The estimation methods are based on processing video in VGA, CIF, and QCIF resolution.

The validation test material contained both multiple coding degradations and various transmission error conditions (e.g., bit errors, dropped packets). In the case where coding distortions are considered in the video signals, the encoder can utilize various compression methods (e.g., MPEG-2, H.264, etc.). The models proposed in this Recommendation may be used to monitor the quality of deployed networks to ensure their operational readiness. The visual effects of the degradations may include spatial as well as temporal degradations (e.g., frame repeats, frame skips, frame rate reduction). The models in this Recommendation can also be used for lab testing of video systems. When used to compare different video systems, it is advisable to use a quantitative method (such as that in [ITU-T J.149]) to determine the models' accuracy for that particular context.

This Recommendation is deemed appropriate for telecommunications services delivered at 4 Mbit/s or less presented on mobile/PDA and computer desktop monitors. The following conditions were allowed in the validation test for each resolution:

- PDA/Mobile (QCIF): 16 kbit/s to 320 kbit/s
- CIF: 64 kbit/s-2 Mbit/s (C01 has several 2 Mbit/s)
- VGA: 128 kbit/s-4 Mbit/s (V13 has one HRC with 6 Mbit/s)

Table 1 – Factors for which J.246 has been evaluated

Test factors
Transmission errors with packet loss
Video resolution QCIF, CIF and VGA
Video bit rates <ul style="list-style-type: none">• QCIF: 16 kbit/s to 320 kbit/s• CIF: 64 kbit/s-2 Mbit/s• VGA: 128 kbit/s-4 Mbit/s
Temporal errors (pausing with skipping) of maximum 2 seconds
Video frame rates from 5 fps to 30 fps
Coding technologies
H.264/AVC (MPEG-4 Part 10), VC-1, Windows Media 9, Real Video (RV 10), MPEG-4 Part 2. See Note 1.

Table 1 – Factors for which J.246 has been evaluated

Applications
Real-time, in-service quality monitoring at the source
Remote destination quality monitoring when side-channels are available for features extracted from source video sequences
Quality measurement for monitoring of a storage or transmission system that utilizes video compression and decompression techniques, either a single pass or a concatenation of such techniques
Lab testing of video systems
NOTE 1 – The validation testing of models included video sequences encoded using 15 different video codecs. The five codecs listed in this table were most commonly applied to encode test sequences and any recommended models may be considered appropriate for evaluating these codecs. In addition to these five codecs a smaller proportion of test sequences were created using the following codecs: Cinepak, DivX, H.261, H.263, H.263+ (Note 2), JPEG-2000, MPEG-1, MPEG-2, Sorenson, H.264 SVC, Theora. It can be noted that some of these codecs were used only for CIF and QCIF resolutions because they are expected to be used in the field mostly for these resolutions. Before applying a model to sequences encoded using one of these codecs the user should carefully examine its predictive performance to determine whether the model reaches acceptable predictive performance.
NOTE 2 – H.263+ is a particular configuration of H.263 (1998).

1.1 Application

This Recommendation provides video quality estimations for video classes TV3 to MM5B, as defined in Annex B of [ITU-T P.911]. Note that the maximum resolution was VGA and the maximum bit rate covered well in the test was 4 Mbit/s. The applications for the estimation models described in this Recommendation include but are not limited to:

- 1) potentially real-time, in-service quality monitoring at the source;
- 2) remote destination quality monitoring when side-channels are available for features extracted from source video sequences;
- 3) quality measurement for monitoring of a storage or transmission system that utilizes video compression and decompression techniques, either a single pass or a concatenation of such techniques;
- 4) lab testing of video systems.

1.2 Limitations

The estimation models described in this Recommendation cannot be used to fully replace subjective testing. Correlation values between two carefully designed and executed subjective tests (i.e., in two different laboratories) normally fall within the range 0.95 to 0.98. If this Recommendation is utilized to make video system comparisons (e.g., comparing two codecs), it is advisable to use a quantitative method (such as that in [ITU-T J.149]) to determine the models' accuracy for that particular context.

The models in this Recommendation were validated by measuring video that exhibits frame freezes up to 2 seconds.

The models in this Recommendation were not validated for measuring video that has a steadily increasing delay (e.g., video which does not discard missing frames after a frame freeze).

It should be noted that in case of new coding and transmission technologies producing artifacts which were not included in this evaluation, the objective models may produce erroneous results. Here a subjective evaluation is required.

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

2 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. The reference to a document within this Recommendation does not give it, as a stand-alone document, the status of a Recommendation.

- [ITU-T J.143] Recommendation ITU-T J.143 (2000), *User requirements for objective perceptual video quality measurements in digital cable television.*
- [ITU-T P.910] Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications.*
- [ITU-T P.911] Recommendation ITU-T P.911 (1998), *Subjective audiovisual quality assessment methods for multimedia applications.*

2.2 Informative references

- [ITU-T H.261] Recommendation ITU-T H.261 (1993), *Video codec for audiovisual services at p x 64 kbits.*
- [ITU-T H.263] Recommendation ITU-T H.263 (1996), *Video coding for low bit rate communication.*
- [ITU-T H.263+] Recommendation ITU-T H.263 (1998), *Video coding for low bit rate communication (H.263+).*
- [ITU-T H.264] Recommendation ITU-T H.264 (2003), *Advanced video coding for generic audiovisual services.*
- [ITU-T J.144] Recommendation ITU-T J.144 (2001), *Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference.*
- [ITU-T J.148] Recommendation ITU-T J.148 (2003), *Requirements for an objective perceptual multimedia quality model.*
- [ITU-T J.149] Recommendation ITU-T J.149 (2004), *Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM).*
- [ITU-T J.244] Recommendation ITU-T J.244 (2008), *Calibration methods for constant misalignment of spatial and temporal domains with constant gain and offset.*
- [ITU-T P.931] Recommendation ITU-T P.931 (1998), *Multimedia communications delay, synchronization and frame rate measurement.*
- [ITU-R BT.500-11] Recommendation ITU-R BT.500-11 (in force), *Methodology for the subjective assessment of the quality of television pictures.*

[VQEG] Final report from the video quality experts group on the validation of objective models of multimedia quality-Phase I, 2008.
ftp://vqeg.its.bldrdoc.gov/Documents/Projects/multimedia/MM_Final_Report/VQEG_MM_Report_Final_v2.6.pdf

3 Definitions

3.1 Terms defined elsewhere

This Recommendation uses the following terms defined elsewhere:

3.1.1 objective perceptual measurement (picture): [ITU-T J.144]

3.1.2 Proponent: [ITU-T J.144]

3.1.3 subjective assessment (picture): [ITU-T J.144]

3.2 Terms defined in this Recommendation

This Recommendation defines the following terms:

3.2.1 anomalous frame repetition: An event where the HRC outputs a single frame repeatedly in response to an unusual or out of the ordinary event. Anomalous frame repetition includes but is not limited to the following types of events: an error in the transmission channel, a change in the delay through the transmission channel, limited computer resources impacting the decoder's performance, and limited computer resources impacting the display of the video signal.

3.2.2 constant frame skipping: An event where the HRC outputs frames with updated content at an effective frame rate that is fixed and less than the source frame rate.

3.2.3 effective frame rate: Number of unique frames (i.e., total frames – repeated frames) per second.

3.2.4 frame rate: Number of (progressive) frames displayed per second (fps).

3.2.5 intended frame rate: Number of video frames per second physically stored for some representation of a video sequence. The intended frame rate may be constant or may change with time. Two examples of constant intended frame rates are a BetacamSP tape containing 25 fps and a VQEG FR-TV Phase I compliant 625-line YUV file containing 25 fps; these both have an intended frame rate of 25 fps. One example of a variable intended frame rate is a computer file containing only new frames; in this case the intended frame rate exactly matches the effective frame rate. The content of video frames is not considered when determining intended frame rate.

3.2.6 live network conditions: Errors imposed upon the digital video bit stream as a result of live network conditions. Examples of error sources include packet loss due to heavy network traffic, increased delay due to transmission route changes, multi-path on a broadcast signal, and fingerprints on a DVD. Live network conditions tend to be unpredictable and unrepeatable.

3.2.7 pausing with skipping: Events where the video pauses for some period of time and then restarts with some loss of video information. In pausing with skipping, the temporal delay through the system will vary about an average system delay, sometimes increasing and sometimes decreasing. One example of pausing with skipping is a pair of IP Videophones, where heavy network traffic causes the IP Videophone display to freeze briefly; when the IP Videophone display continues, some content has been lost. Another example is a videoconferencing system that performs constant frame skipping or variable frame skipping. Constant frame skipping and variable frame skipping are subsets of pausing with skipping. A processed video sequence containing pausing with skipping will be approximately the same duration as the associated original video sequence.

3.2.8 pausing without skipping: Any event where the video pauses for some period of time and then restarts without losing any video information. Hence, the temporal delay through the system must increase. One example of pausing without skipping is a computer simultaneously downloading and playing an AVI file, where heavy network traffic causes the player to pause briefly and then continue playing. A processed video sequence containing pausing without skipping events will always be longer in duration than the associated original video sequence.

3.2.9 refresh rate: The rate at which the computer monitor is updated.

3.2.10 simulated transmission errors: Errors imposed upon the digital video bit stream in a highly controlled environment. Examples include simulated packet loss rates and simulated bit errors. Parameters used to control simulated transmission errors are well defined.

3.2.11 source frame rate (SFR): The intended frame rate of the original source video sequences. The source frame rate is constant. For the VQEG MM Phase I test the SFR was either 25 fps or 30 fps.

3.2.12 transmission errors: Any error imposed on the video transmission. Example types of errors include simulated transmission errors and live network conditions.

3.2.13 variable frame skipping: An event where the HRC outputs frames with updated content at an effective frame rate that changes with time. The temporal delay through the system will increase and decrease with time, varying about an average system delay. A processed video sequence containing variable frame skipping will be approximately the same duration as the associated original video sequence.

4 Abbreviations and acronyms

This Recommendation uses the following abbreviations and acronyms:

ACR	Absolute Category Rating
ACR-HR	Absolute Category Rating with Hidden Reference
AVI	Audio Video Interleave
CIF	Common Intermediate Format (352 × 288 pixels)
DMOS	Difference Mean Opinion Score
FR	Full Reference
FRTV	Full Reference TeleVision
HRC	Hypothetical Reference Circuit
ILG	VQEG's Independent Laboratory Group
LCD	Liquid Crystal Display
MM	Multimedia
MOS	Mean Opinion Score
MOSp	Mean Opinion Score, predicted
NR	No (or Zero) Reference
PDA	Personal Digital Assistant
PSNR	Peak Signal to Noise Ratio
PVS	Processed Video Sequence
QCIF	Quarter Common Intermediate Format (176 × 144 pixels)

RMSE	Root Mean Square Error
RR	Reduced Reference
SFR	Source Frame Rate
SRC	Source Reference Channel or Circuit
VGA	Video Graphics Array (640 × 480 pixels)
VQEG	Video Quality Experts Group
VQM	Video Quality Metric
YUV	Color Space and file format

5 Conventions

None.

6 Description of the reduced reference measurement method

The double-ended measurement method with reduced reference, for objective measurement of perceptual video quality, evaluates the performance of systems by making a comparison between features extracted from 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 reduced reference method to test a codec in the laboratory.

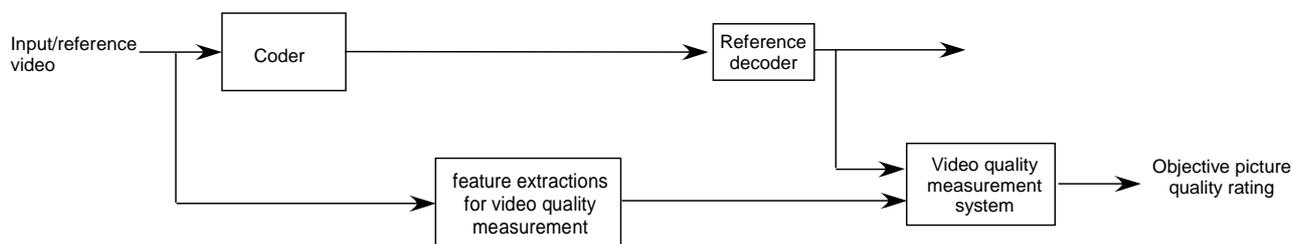


Figure 1 – Application of the reduced reference perceptual quality measurement method to test a codec in the laboratory

The comparison between input and output signals may require a temporal alignment or a spatial alignment process, the latter 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.

Alignment and gain adjustment is known as registration. This process is required because most reduced reference methods compare the features extracted from reference pictures and processed pictures on what is effectively a pixel-by-pixel basis. The video quality metrics described in Annex A include registration methods.

As the video quality metrics are typically based on approximations to human visual responses, rather than on the measurement of specific coding artefacts, they are in principle equally valid for analogue systems and for digital systems. They are 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 reduced reference method to test a transmission chain.

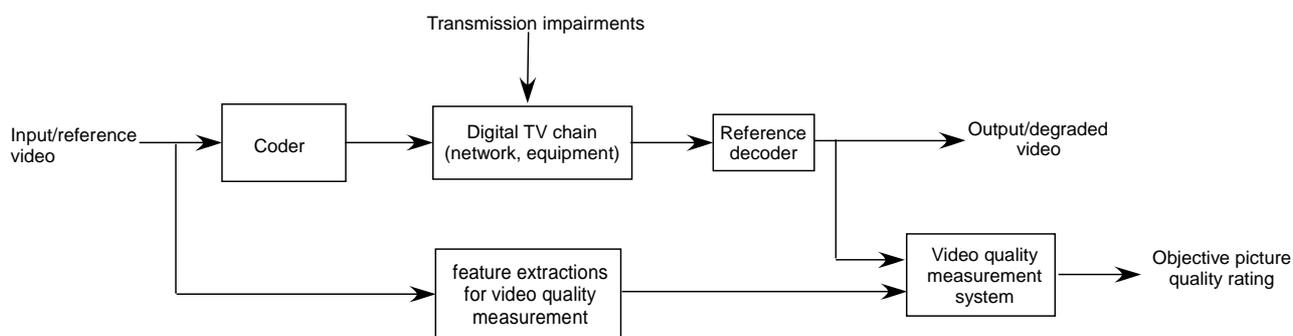


Figure 2 – Application of the reduced reference perceptual quality measurement method to test a transmission chain

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 ACR-HR methods specified in [ITU-T P.910].

7 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 9 and 12 and ITU-R Study Group 6. The recently completed Multimedia Phase I test of VQEG assessed the performance of proposed reduced reference perceptual video quality measurement algorithms for QCIF, CIF, and VGA formats.

Based on present evidence, the following method can be recommended by ITU-T at this time:

Annex A – VQEG Proponent: Yonsei University, Korea

The technical descriptions of this model can be found in Annex A.

Table 2 below provides informative details on the models' performances in the VQEG Multimedia Phase I test.

Table 2 – VGA resolution: Informative description on the models' performances in the VQEG Multimedia Phase I test: Averages over 13 subjective tests

Statistic	Yonsei RR10k	Yonsei RR64k	Yonsei RR128k	PSNR
Correlation	0.803	0.803	0.803	0.713
RMSE	0.599	0.599	0.598	0.714
Outlier Ratio	0.556	0.553	0.552	0.615

Table 3 – CIF resolution: Informative description on the models' performances in the VQEG Multimedia Phase I test: Averages over 14 subjective tests

Statistic	Yonsei RR10k	Yonsei RR64k	PSNR
Correlation	0.780	0.782	0.656
RMSE	0.593	0.590	0.720
Outlier Ratio	0.519	0.511	0.632

Table 4 – QCIF resolution: Informative description on the models' performances in the VQEG Multimedia Phase I test: Averages over 14 subjective tests

Statistic	Yonsei RR1k	Yonsei RR10k	PSNR
Correlation	0.771	0.791	0.662
RMSE	0.604	0.578	0.721
Outlier Ratio	0.505	0.486	0.596

The average correlations of the primary analysis for the RR VGA models were all 0.80, and PSNR was 0.71. Individual model correlations for some experiments were as high as 0.93. The average RMSE for the RR VGA models were all 0.60, and PSNR was 0.71. The average outlier ratio for the RR VGA models ranged from 0.55 to 0.56, and PSNR was 0.62. All proposed models performed statistically better than PSNR for 7 of the 13 experiments. Based on each metric, each RR VGA model was in the group of top performing models the following number of times:

Statistic	Yonsei RR10k	Yonsei RR64k	Yonsei RR128k	PSNR
Correlation	13	13	13	7
RMSE	13	13	13	6
Outlier Ratio	13	13	13	10

The average correlations of the primary analysis for the RR CIF models were 0.78, and PSNR was 0.66. Individual model correlations for some experiments were as high as 0.90. The average RMSE for the RR CIF models were all 0.59, and PSNR was 0.72. The average outlier ratio for the RR CIF models were 0.51 and 0.52, and PSNR was 0.63. All proposed models performed statistically better than PSNR for 10 of the 14 experiments. Based on each metric, each RR CIF model was in the group of top performing models the following number of times:

Statistic	Yonsei RR 10k	Yonsei RR64k	PSNR
Correlation	14	14	5
RMSE	14	14	4
Outlier Ratio	14	14	5

The average correlations of the primary analysis for the RR QCIF models were 0.77 and 0.79, and PSNR was 0.66. Individual model correlations for some experiments were as high as 0.89. The average RMSE for the RR QCIF models were 0.58 and 0.60, and PSNR was 0.72. The average outlier ratio for the RR QCIF models were 0.49 and 0.51, and PSNR was 0.60. All proposed models performed statistically better than PSNR for at least 9 of the 14 experiments. Based on each metric, each RR QCIF model was in the group of top performing models the following number of times:

Statistic	Yonsei RR1k	Yonsei RR10k	PSNR
Correlation	14	14	5
RMSE	14	14	4
Outlier Ratio	12	13	4

Annex A

Yonsei University Reduced Reference Method

(This annex forms an integral part of this Recommendation)

A.1 Introduction

Although PSNR has been widely used as an objective video quality measure, it is also reported that it does not well represent perceptual video quality. By analysing how humans perceive video quality, it is observed that the human visual system is sensitive to degradation around the edges. In other words, when the edge pixels of a video are blurred, evaluators tend to give low scores to the video even though the PSNR is high. Based on this observation, the reduced reference models which mainly measure edge degradations have been developed.

Figure A.1 illustrates how a reduced-reference model works. Features which will be used to measure video quality at a monitoring point are extracted from the source video sequence and transmitted. Table A.1 shows the side-channel bandwidths for the features, which have been tested in the VQEG MM test.

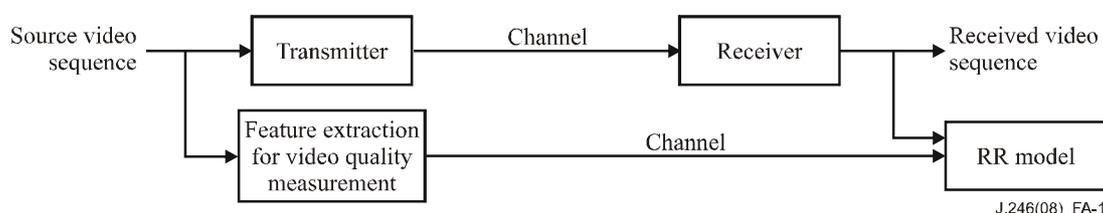


Figure A.1 – Block diagram of reduced reference model

Table A.1 – Side-channel bandwidths

Video Format	Tested Bandwidths
QCIF	1 kbit/s, 10 kbit/s
CIF	10 kbit/s, 64 kbit/s
VGA	10 kbit/s, 64 kbit/s, 128 kbit/s

A.2 The EPSNR Reduced Reference Models

A.2.1 Edge PSNR (EPSNR)

The RR models mainly measure on edge degradations. In the models, an edge detection algorithm is first applied to the source video sequence to locate the edge pixels. Then, the degradation of those edge pixels is measured by computing the mean squared error. From this mean squared error, the edge PSNR is computed.

One can use any edge detection algorithm, though there may be minor differences in the results. For example, one can use any gradient operator to locate edge pixels. A number of gradient operators have been proposed. In many edge detection algorithms, the horizontal gradient image $g_{horizontal}(m,n)$ and the vertical gradient image $g_{vertical}(m,n)$ are first computed using gradient operators. Then, the magnitude gradient image $g(m,n)$ may be computed as follows:

$$g(m,n) = |g_{horizontal}(m,n)| + |g_{vertical}(m,n)|$$

Finally, a thresholding operation is applied to the magnitude gradient image $g(m,n)$ to find edge pixels. In other words, pixels whose magnitude gradients exceed a threshold value are considered as edge pixels.

Figures A.2-6 illustrate the procedure. Figure A.2 shows a source image. Figure A.3 shows a horizontal gradient image $g_{horizontal}(m,n)$, which is obtained by applying a horizontal gradient operator to the source image of Figure A.2. Figure A.4 shows a vertical gradient image $g_{vertical}(m,n)$, which is obtained by applying a vertical gradient operator to the source image of Figure A.2. Figure A.5 shows the magnitude gradient image (edge image) and Figure A.6 shows the binary edge image (mask image) obtained by applying thresholding to the magnitude gradient image of Figure A.5.



Figure A.2 – A source image (original image)

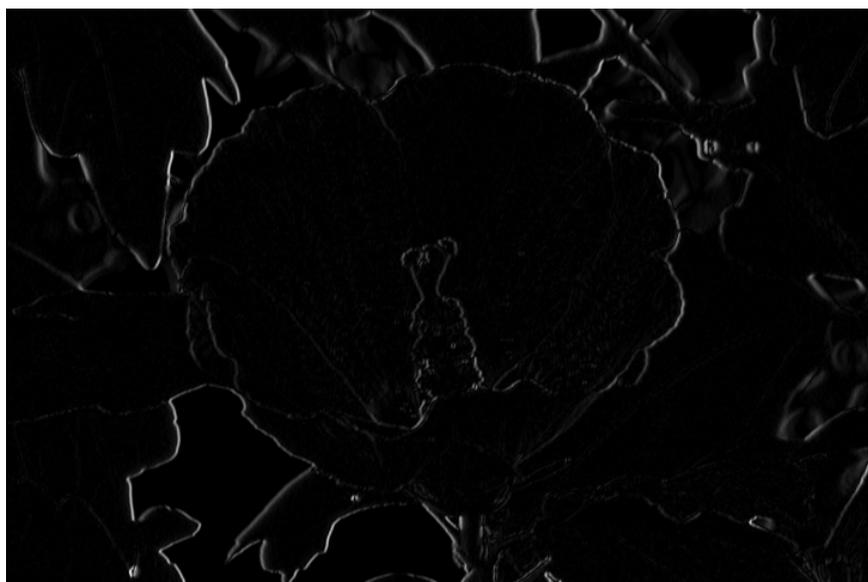


Figure A.3 – A horizontal gradient image, which is obtained by applying a horizontal gradient operator to the source image of Figure A.2

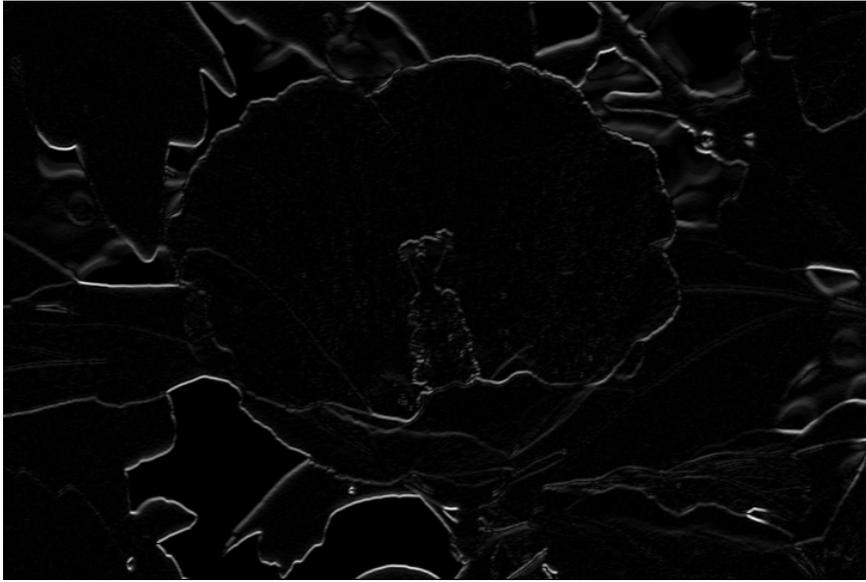


Figure A.4 – A vertical gradient image, which is obtained by applying a vertical gradient operator to the source image of Figure A.2

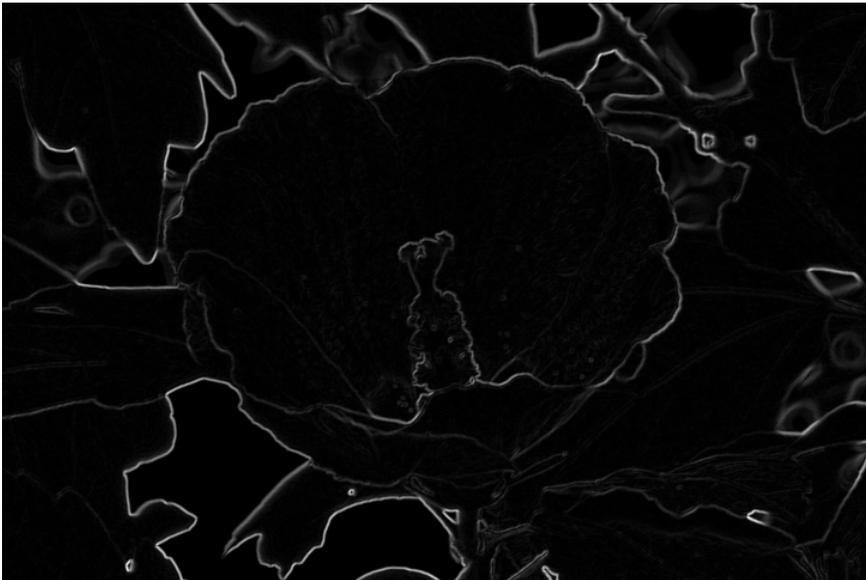


Figure A.5 – A magnitude gradient image



Figure A.6 – A binary edge image (mask image) obtained by applying thresholding to the magnitude gradient image of Figure A.5

Alternatively, one may use a modified procedure to find edge pixels. For instance, one may first apply a vertical gradient operator to the source image, producing a vertical gradient image. Then, a horizontal gradient operator is applied to the vertical gradient image, producing a modified successive gradient image (horizontal and vertical gradient image). Finally, a thresholding operation may be applied to the modified successive gradient image to find edge pixels. In other words, pixels of the modified successive gradient image, which exceed a threshold value, are considered as edge pixels. Figures A.7-9 illustrate the modified procedure. Figure A.7 shows a vertical gradient image $g_{vertical}(m,n)$, which is obtained by applying a vertical gradient operator to the source image of Figure A.2. Figure A.8 shows a modified successive gradient image (horizontal and vertical gradient image), which is obtained by applying a horizontal gradient operator to the vertical gradient image of Figure A.7. Figure A.9 shows the binary edge image (mask image) obtained by applying thresholding to the modified successive gradient image of Figure A.8.

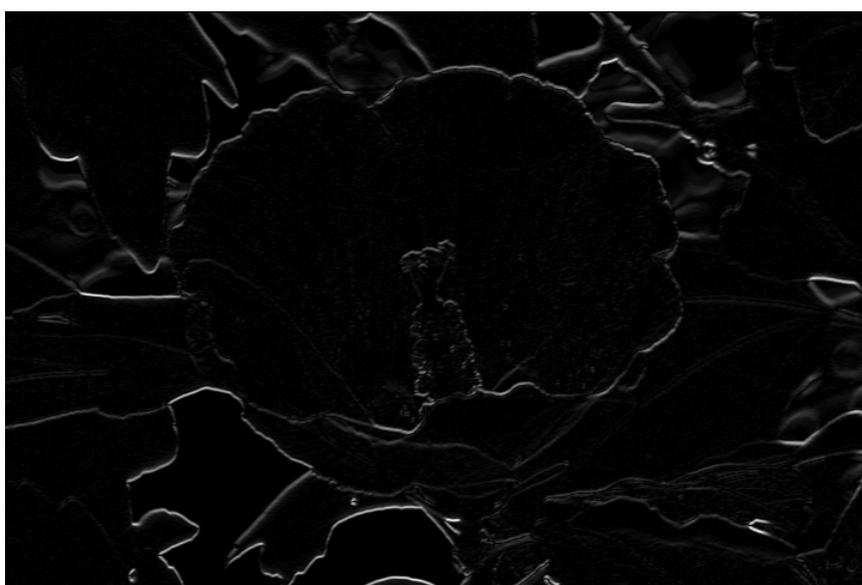


Figure A.7 – A vertical gradient image, which is obtained by applying a vertical gradient operator to the source image of Figure A.2

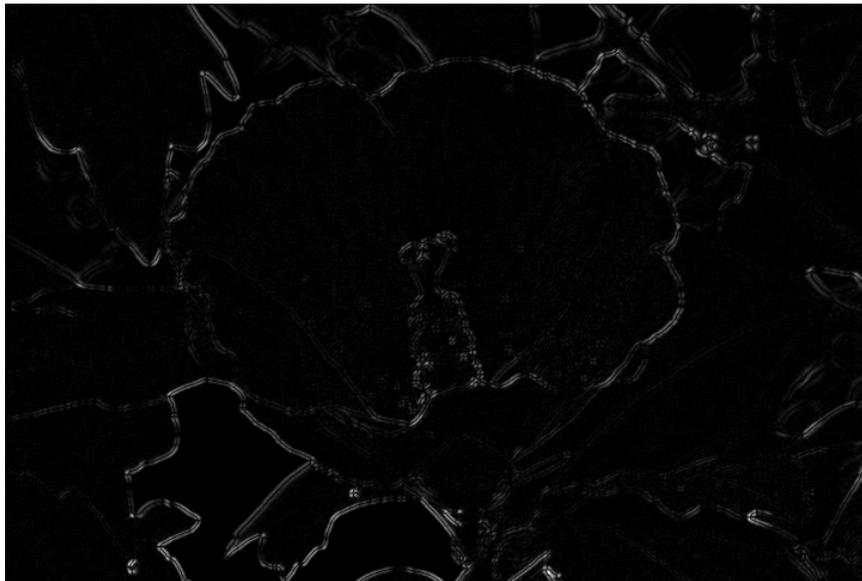


Figure A.8 – A modified successive gradient image (horizontal and vertical gradient image), which is obtained by applying a horizontal gradient operator to the vertical gradient image of Figure A.7

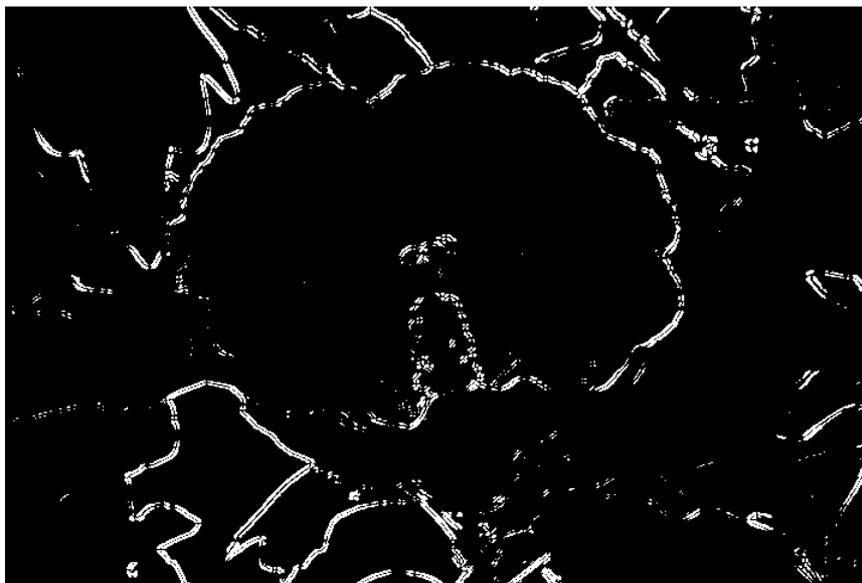


Figure A.9 – A binary edge image (mask image) obtained by applying thresholding to the modified successive gradient image of Figure A.8

It is noted that both methods can be understood as an edge detection algorithm. One may choose any edge detection algorithm depending on the nature of videos and compression algorithms. However, some methods may outperform other methods.

Thus, in the model, an edge detection operator is first applied, producing edge images (Figures A.5 and A.8). Then, a mask image (binary edge image) is produced by applying thresholding to the edge image (Figures A.6 and A.9). In other words, pixels of the edge image whose value is smaller than threshold t_e are set to zero and pixels whose value is equal to or larger than the threshold are set to a non-zero value. Figures A.6 and A.9 show some mask images. Since a video can be viewed as a sequence of frames or fields, the above-stated procedure can be applied to each frame or field of

videos. Since the model can be used for field-based videos or frame-based videos, the terminology "image" will be used to indicate a field or frame.

A.2.2 Selecting features from source video sequences

Since the model is a reduced-reference (RR) model, a set of features need to be extracted from each image of a source video sequence. In the EPSNR RR model, a certain number of edge pixels are selected from each image. Then, the locations and pixel values are encoded and transmitted. However, for some video sequences, the number of edge pixels can be very small when a fixed threshold value is used. In the worst scenario, it can be zero (blank images or very low frequency images). In order to address this problem, if the number of edge pixels of an image is smaller than a given value, the user may reduce the threshold value until the number of edge pixels is larger than a given value. Alternatively, one can select edge pixels which correspond to the largest values of the horizontal and vertical gradient image. When there are no edge pixels (e.g., blank images) in a frame, one can randomly select the required number of pixels or skip the frame. For instance, if 10 edge pixels are to be selected from each frame, one can sort the pixels of the horizontal and vertical gradient image according to their values and select the largest 10 values. However, this procedure may produce multiple edge pixels at the identical locations. To address this problem, one can first select several times of the desired number of pixels of the horizontal and vertical gradient image and then randomly choose the desired number of edge pixels among the selected pixels of the horizontal and vertical gradient image. In the models tested in the VQEG multimedia test, the desired number of edge pixels is randomly selected among a large pool of edge pixels. The pool of edge pixels is obtained by applying a thresholding operation to the gradient image.

In the EPSNR RR models, the locations and edge pixel values are encoded. It is noted that during encoding process, cropping may be applied. In order to avoid selecting edge pixels in the cropped areas, the model selects edge pixels in the middle area (Figure A.10). Table A.2 shows the sizes after cropping. Table A.2 also shows the number of bits required to encode the location and pixel value of an edge pixel.

Table A.2 – Bits requirement per edge pixel

Video Format	Size	Size after cropping	Bits for location	Bits for pixel value	Total bits per pixel
QCIF	176 × 144	168 × 136	15	8	23
CIF	352 × 288	338 × 274	17	8	25
VGA	640 × 480	614 × 454	19	8	27

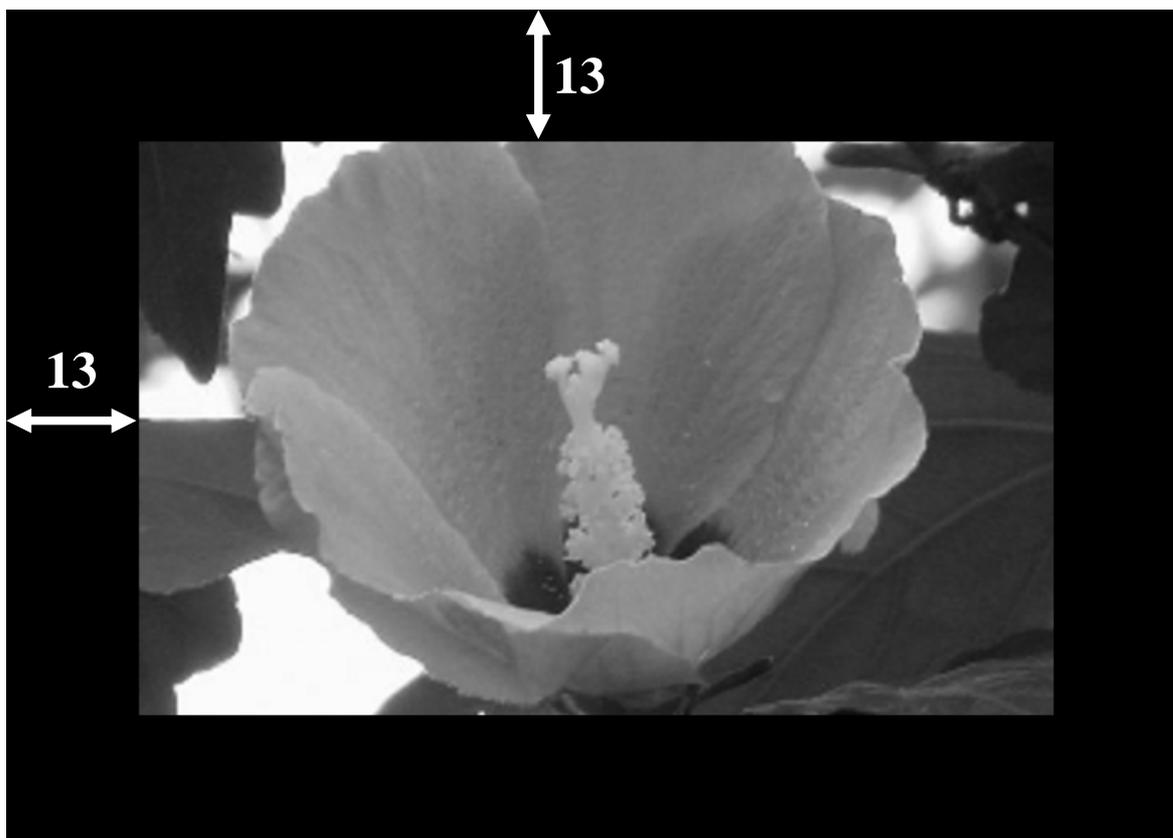


Figure A.10 – An example of cropping (VGA) and the middle area

The model selects edge pixels from each frame in accordance with the allowed bandwidth (Table A.1). Tables A.3-4 show the number of edge pixels per frame which can be transmitted for the tested bandwidths.

Table A.3 – Number of edge pixels per frame (30 frames per second)

Video Format	1 kbit/s	10 kbit/s	64 kbit/s	128 kbit/s
QCIF	1	14		
CIF		13	85	
VGA		12	79	158

Table A.4 – Number of edge pixels per frame (25 frames per second)

Video Format	1 kbit/s	10 kbit/s	64 kbit/s	128 kbit/s
QCIF	1	17		
CIF		16	102	
VGA		14	94	189

A.2.3 Spatial/temporal registration and gain/offset adjustment

Before computing the difference between the edge pixels of the source video sequence and those of the processed video sequence which is the received video sequence at the receiver, the model first applies a spatial/temporal registration and gain/offset adjustment. First, a full search algorithm is applied to find global spatial and temporal shifts along with gain and offset values (Figure A.11).

Then, for every possible spatial shifts ($\Delta x, \Delta y$), a temporal registration is performed and the EPSNR is computed. Finally the smallest EPSNR is chosen as a video quality metric (VQM).

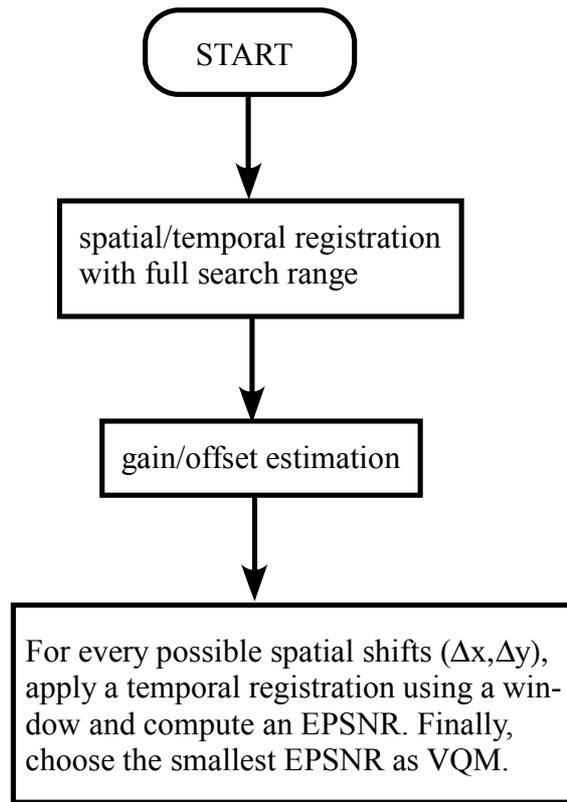


Figure A.11 – Flowchart of the model

At the monitoring point, the processed video sequence should be aligned with the edge pixels extracted from the source video sequence. However, if the side-channel bandwidth is small, only a few edge pixels of the source video sequence are available (Figure A.12). Consequently, the temporal registration can be inaccurate if the temporal registration is performed using a single frame (Figure A.13). To address this problem, the model uses a window for temporal registration. Instead of using a single frame of the processed video sequence, the model builds a window which consists of a number of adjacent frames to find the optimal temporal shift. Figure A.14 illustrates the procedure. The mean squared error within the window is computed as follows:

$$MSE_{window} = \frac{1}{N_{win}} \sum (E_{SRC}(i) - E_{PVS}(i))^2$$

where MSE_{window} is the window mean squared error, $E_{SRC}(i)$ is an edge pixel within the window which has a corresponding pixel in the processed video sequence, $E_{PVS}(i)$ is a pixel of the processed video sequence corresponding to the edge pixel, and N_{win} is the total number of edge pixels used to compute MSE_{window} . This window mean squared error is used as the difference between a frame of the processed video sequence and the corresponding frame of the source video sequence.

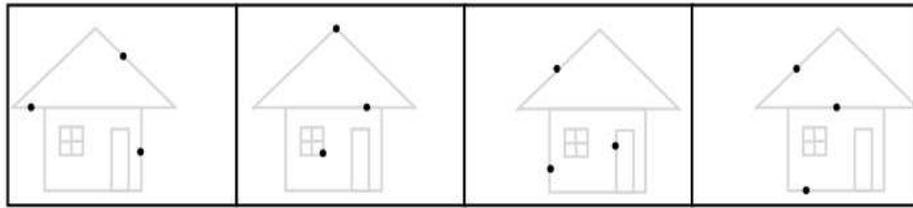


Figure A.12 – Edge pixel selection of the source video sequence

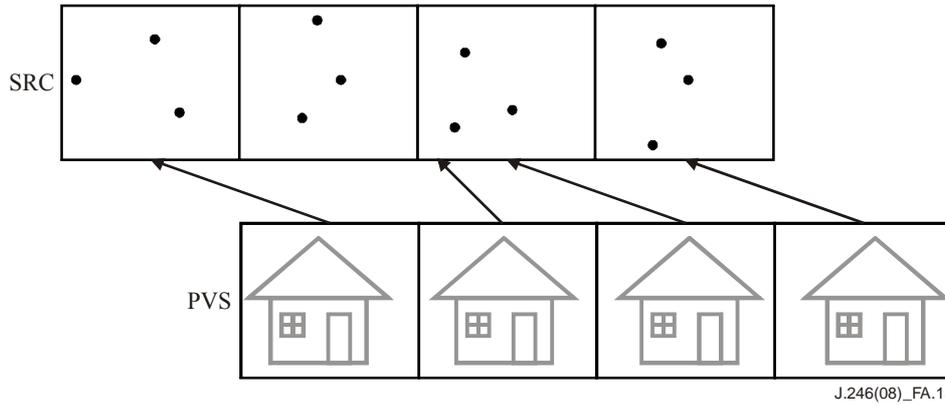


Figure A.13 – Aligning the processed video sequence to the edge pixels of the source video sequence

The window size can be determined by considering the nature of the processed video sequence. For a typical application, a window corresponding to two seconds is recommended. Alternatively, various sizes of windows can be applied and the best one which provides the smallest mean squared error can be used.

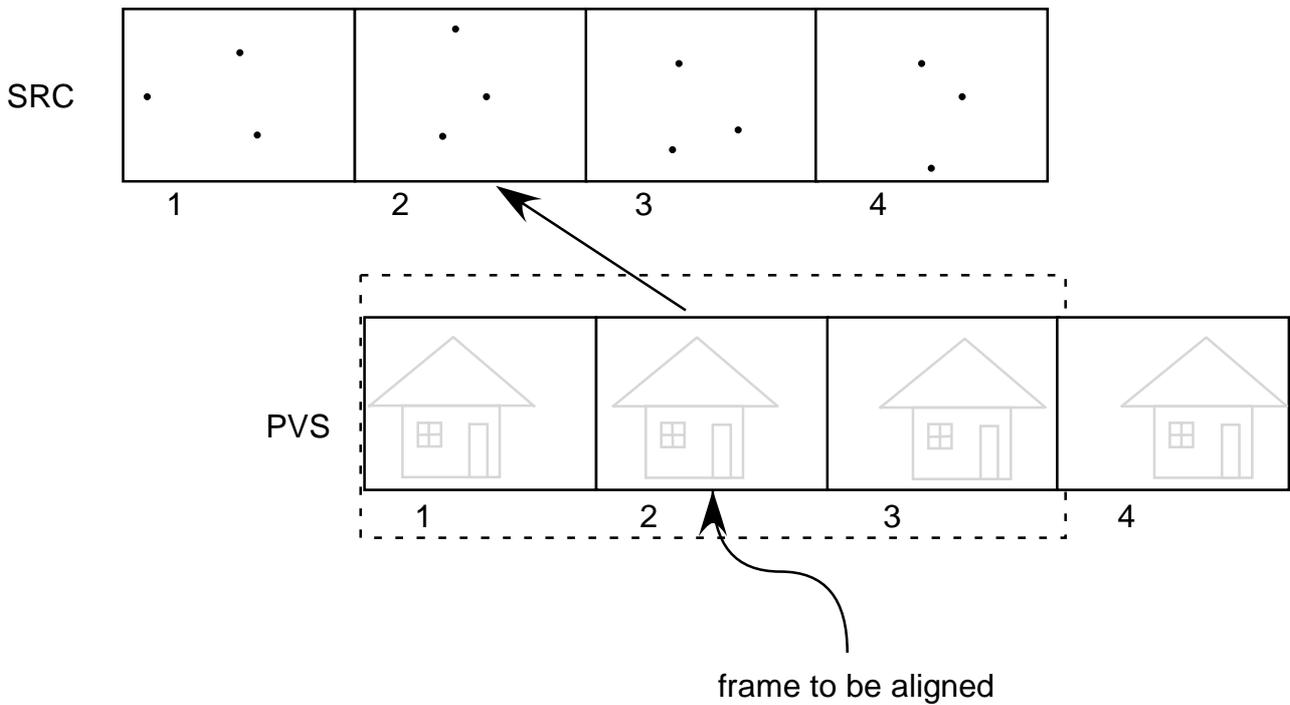


Figure A.14 – Aligning the processed video sequence to the edge pixels using a window

When the source video sequence is encoded at high compression ratios, the encoder may reduce the number of frames per second and the processed video sequence has repeated frames (Figure A.15). In Figure A.15, the processed video sequence does not have frames corresponding to some frames of the source video sequence (2, 4, 6, 8th frames). In this case, the model does not use repeated frames in computing the mean squared error. In other words, the model performs temporal registration using the first frame (valid frame) of each repeated block. Thus, in Figure A.16, only three frames (3, 5, 7th frames) within the window are used for temporal registration.

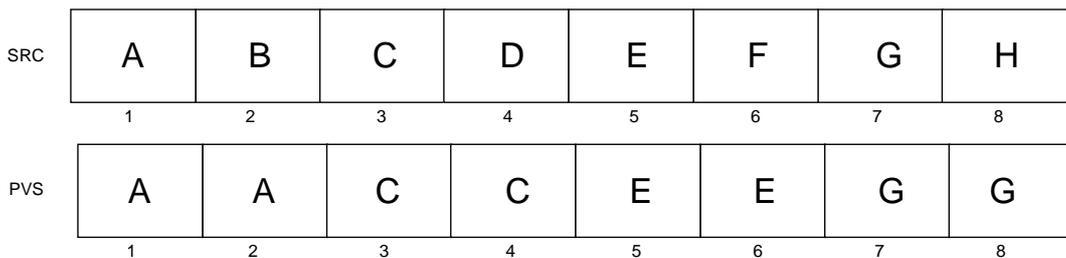


Figure A.15 – Example of repeated frames

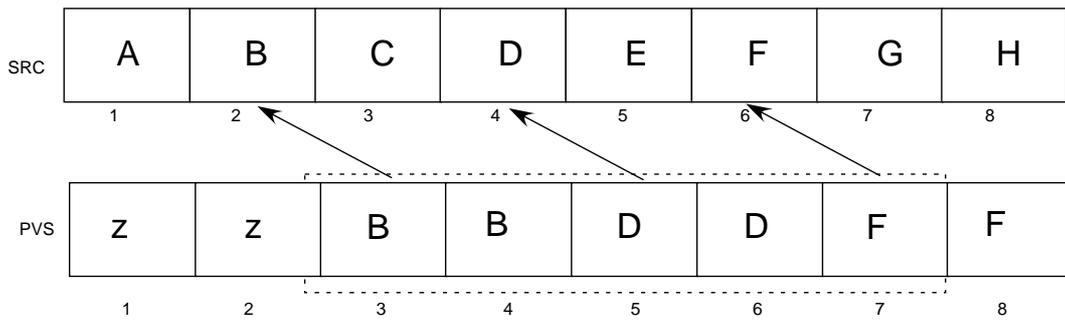


Figure A.16 – Handling repeated frames

It is possible to have a processed video sequence with irregular frame repetition, which may cause the temporal registration method using a window to produce inaccurate results. To address this problem, it is possible to locally adjust each frame of the window within a given value (e.g., ± 1) as shown in Figure A.18 after the temporal registration using a window. Then, the local adjustment which provides the minimum MSE is used to compute the EPSNR.

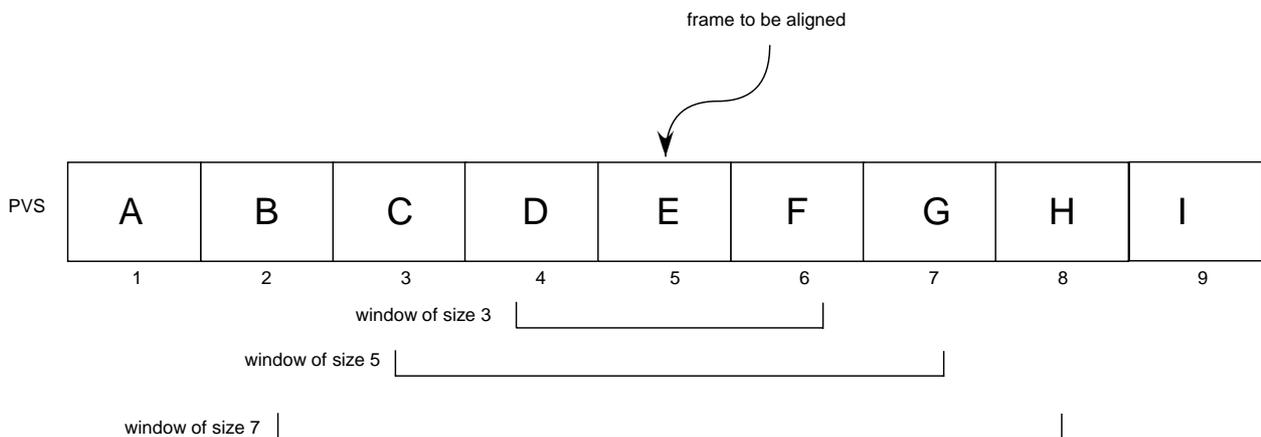
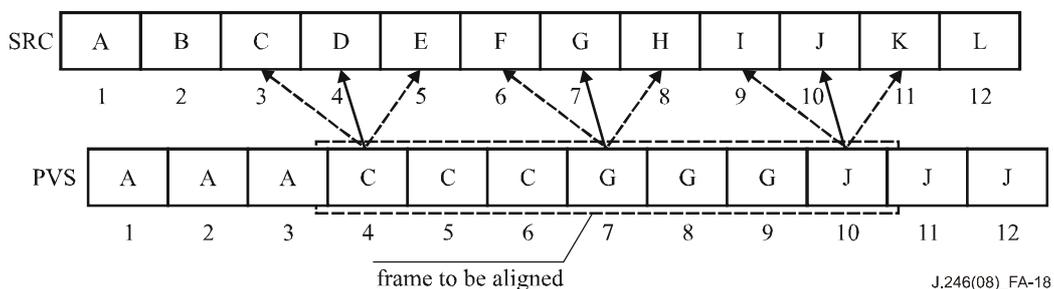


Figure A.17 – Windows of various sizes



J.246(08)_FA-18

Figure A.18 – Local adjustment for temporal registration using a window

A.2.4 Computing EPSNR and post-processing

After temporal registration is performed, the average of the differences between the edge pixels of the source video sequence and the corresponding pixels of the processed video sequence is computed, which can be understood as the edge mean squared error of the processed video sequence (MSE_{edge}). Finally, the EPSNR (edge PSNR) is computed as follows:

$$EPSNR = 10 \log_{10} \left(\frac{P^2}{MSE_{edge}} \right)$$

where p is the peak value of the image.

In multimedia video encoding, there can be frame repeating due to reduced frame rates and frame freezing due to transmission error, which will degrade the perceptual video quality. In order to address this effect, the model applies the following adjustment before computing the EPSNR:

$$MSE_{frozen_frame_considered} = MSE_{edge} \times \frac{K \times N_{total_frame}}{N_{total_frame} - N_{total_frozen_frame}}$$

where $MSE_{frozen_frame_considered}$ is the mean squared error which takes into account repeated and frozen frames, N_{total_frame} is the total number of frames, $N_{total_frozen_frame}$ is the total number of frozen frames, K is a constant. In the model tested in the VQEG multimedia test, K was set to 1.

When the EPSNR exceeds a certain value, the perceptual quality becomes saturated. In this case, it is possible to set the upper bound of the EPSNR. Furthermore, when a linear relationship between the EPSNR and DMOS (difference mean opinion score) is desirable, one can apply a piecewise linear function as illustrated in Figure A.19. In the model tested in the VQEG multimedia test, only the upper bound is set to 50 since polynomial curve fitting was used.

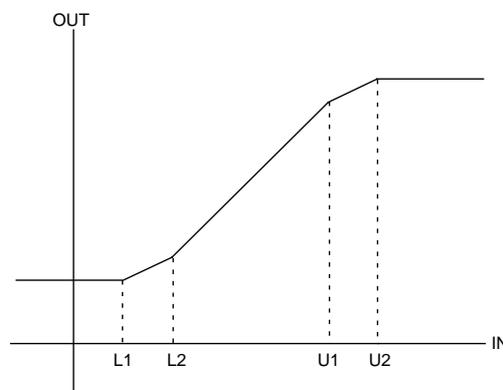


Figure A.19 – Piecewise linear function for linear relationship between the EPSNR and DMOS

A.2.5 Optimal bandwidth of side channel

Appendix I shows the performance comparison as the bandwidth of the side-channel increases. For the QCIF format, it is observed that the correlation coefficients are almost saturated at about 10 kbit/s. After that, increasing the bandwidth produces about 1% improvement. For the CIF format, it is observed that the correlation coefficients are almost saturated at about 15 kbit/s. After that, increasing the bandwidth produces about 0.5% improvement. For the VGA format, it is observed that the correlation coefficients are almost saturated at about 30 kbit/s. After that, increasing the bandwidth produces about 0.5% improvement.

A.3 Conclusions

The EPSNR reduced reference models for the objective measurement of the video quality are proposed based on edge degradation. The models can be implemented in real time with moderate use of computing power. The models are well suited to applications which require real-time video quality monitoring where side-channels are available.

Appendix I

Optimal side-channel bandwidths

(This appendix does not form an integral part of this Recommendation)

Figure I.1 shows the correlation coefficient for different side-channel bandwidths for the QCIF video sets. It can be seen that the correlation coefficients are almost saturated at about 10 kbit/s. After that, increasing the bandwidth produces about 1% improvement.

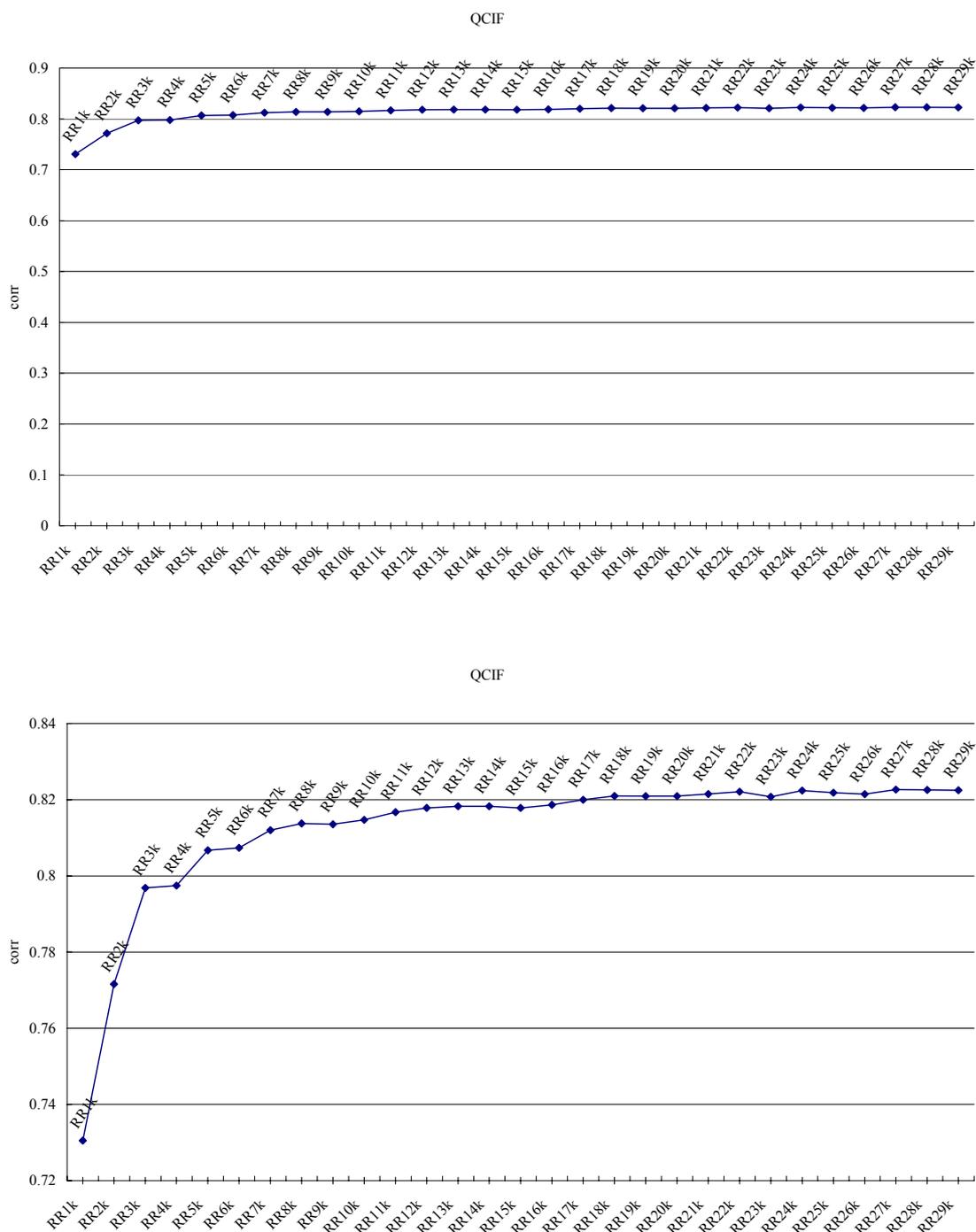


Figure I.1 – Performance improvement as the side-channel bandwidth increases (QCIF)

Figure I.2 shows the correlation coefficient for different side-channel bandwidths for the CIF video sets. It can be seen that the correlation coefficients are almost saturated at about 15 kbit/s. After that, increasing the bandwidth produces about 0.5% improvement.

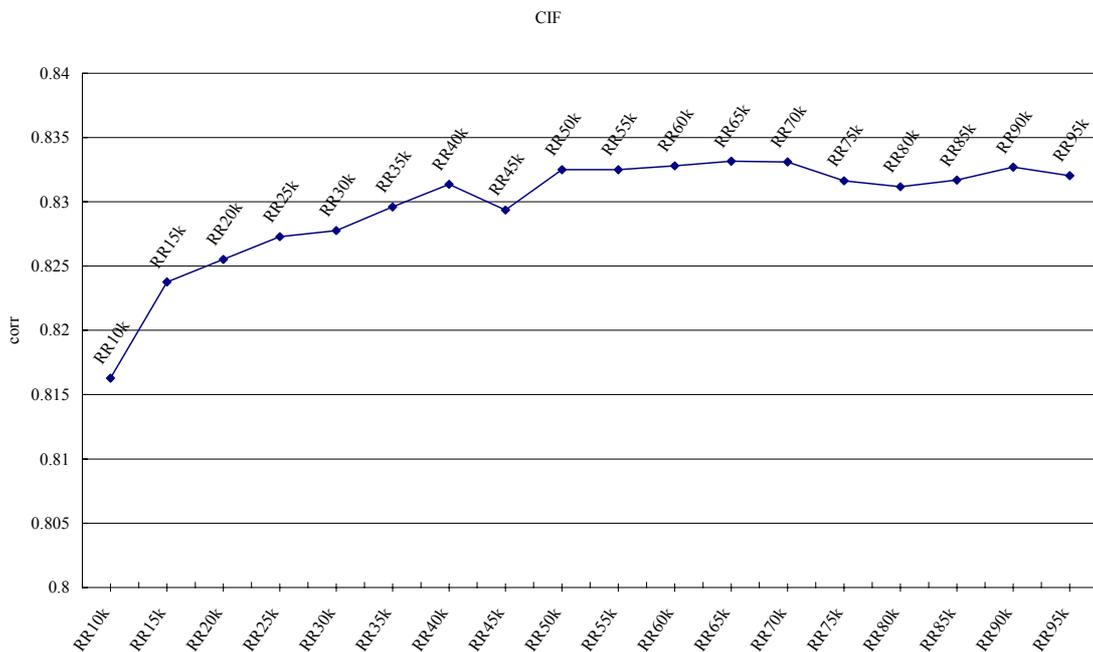
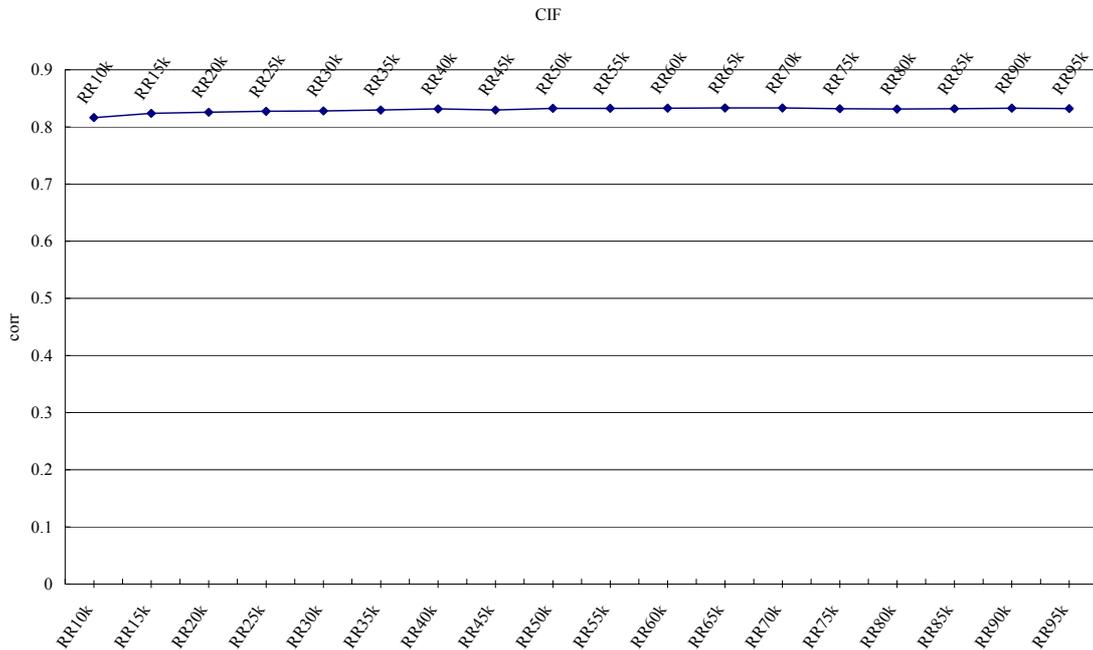


Figure I.2 – Performance improvement as the side-channel bandwidth increases (CIF)

Figure I.3 shows the correlation coefficient for different side-channel bandwidths for the VGA video sets. It can be seen that the correlation coefficients are almost saturated at about 30 kbit/s. After that, increasing the bandwidth produces about 0.5% improvement.

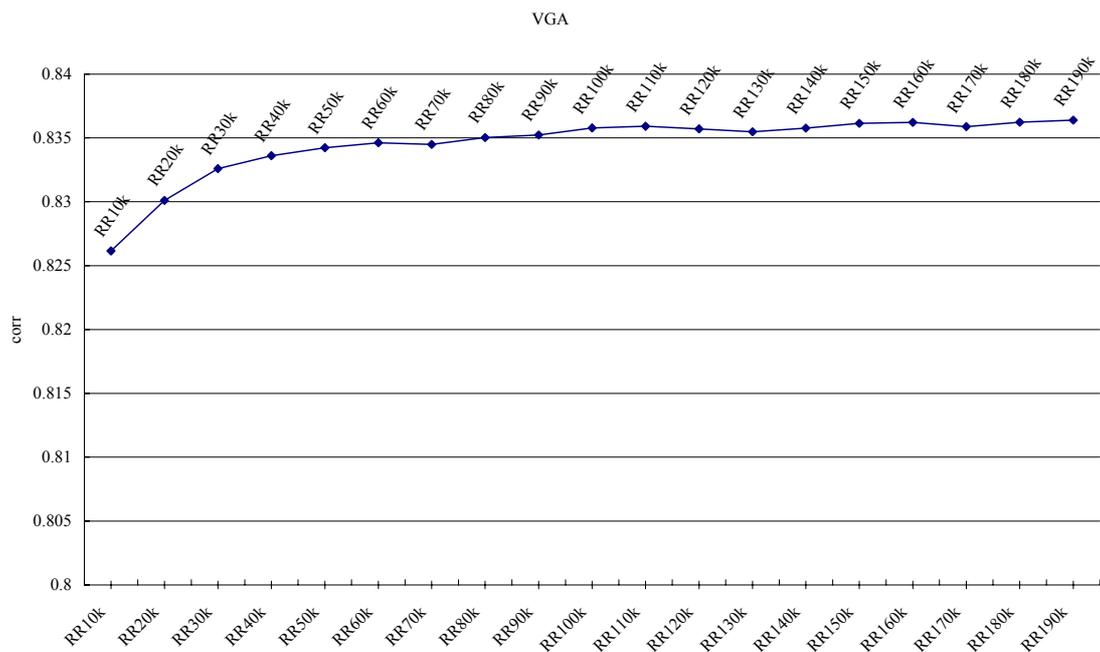
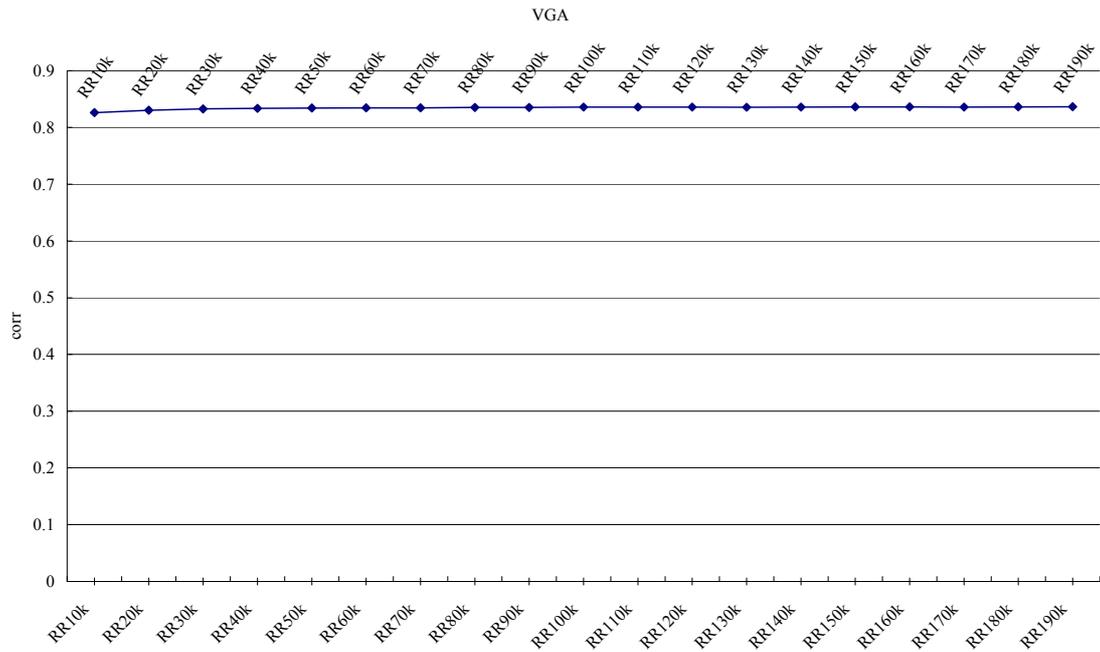


Figure I.3 – Performance improvement as the side-channel bandwidth increases (VGA)

Appendix II

Excerpts from the Synopsis from the Video Quality Experts Group on the validation of objective models of multimedia quality assessment, phase I

(This appendix does not form an integral part of this Recommendation)

II.1 Introduction

This appendix presents results from the Video Quality Experts Group (VQEG) Multimedia validation testing of objective video quality models for mobile/PDA and broadband internet communications services.²

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

The MM test discussed addresses three video resolutions (VGA, CIF, and QCIF) and three types of models: full reference (FR), reduced reference (RR), and no reference (NR). FR models have full access to the source video; RR models have limited bandwidth access to the source video; and NR models do not have access to the source video. RR models can be used in certain applications that cannot be addressed by FR models, such as in-service monitoring in networks. NR models can be used in certain applications that cannot be addressed by FR or RR approaches. Typically, no-reference models are applied in situations where the user does not have access to the source. Proponents were given the option of submitting different models for each video resolution and model type.

Forty one subjective experiments provided data against which model validation was performed. The experiments were divided among the three video resolutions and two frame rates (25 fps and 30 fps). A common set of carefully chosen video sequences were inserted identically into each experiment at a given resolution, to anchor the video experiments to one another and assist in comparisons between the subjective experiments. The subjective experiments included processed video sequences with a wide range of quality, and both compression and transmission errors were present in the test conditions. These forty one subjective experiments included 346 source video sequences and 5320 processed video sequences. These video clips were evaluated by 984 viewers.

A total of thirteen organizations performed subjective MM testing. Of these organizations, five were model proponents (NTT, OPTICOM, Psytechnics, SwissQual, and Yonsei University) and the remainder were either independent testing laboratories (Acreo, CRC, IRCCyN, France Telecom, FUB, Nortel, NTIA, and Verizon) or laboratories that helped by running processed video sequences (PVS) and subjective experiments (KDDI and Symmetricom). Objective models were submitted prior to scene selection, PVS generation, and subjective testing, to ensure none of the models could be trained on the test material. Of the 31 models submitted, six were withdrawn; therefore, 25 are presented in this appendix. A model is considered in this context to be a model type (i.e., FR or RR or NR) for a specified resolution (i.e., VGA or CIF or QCIF). Results for models submitted by the Yonsei University (Korea) are included in this appendix.

² This appendix has been adapted from [VQEG] with permission. This synopsis is a shortened version of the VQEG MM synopsis, as this appendix only addresses RR models.

VQEG cautions that the MM data should not be used as evidence to standardize any other objective video quality model that was not tested within this phase. Such a comparison would not be valid, because another model could have been trained on the MM data.

II.2 Model performance evaluation techniques

The models were evaluated using three statistics that provide insights into model performance: Pearson Correlation, Root-Mean Squared Error (RMSE) and Outlier Ratios. These statistics compare the objective model's predictions with the subjective quality as judged by a panel of human observers. Each model was fitted to each subjective experiment, by optimizing Pearson Correlation with subjective data first, and minimizing RMSE second.

Each of these statistics (Pearson Correlation, RMSE, and Outlier Ratios) can be used to determine whether a model is in the group of top performing models for one video format/resolution (i.e., a group of models that include the top performing model and models that are statistically equivalent to the top performing model). Note that a model that is not in the top performing group and is statistically worse than the top performing model but may be statistically equivalent to one or more of the models that are in the top performing group. Statistical significances are computed for each metric separately, and therefore the models' ranking per video resolution is accomplished per each statistical metric.

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

Primary analysis considers each video sequence separately. Secondary analysis averages over all video sequences associated with each video system (or condition), and thus reflects how well the model tracks the average hypothetical reference circuit (HRC) performance. The common set of video sequences are included in primary analysis but eliminated from secondary analysis. The following sections report on model performance across model type and resolution. The reader should be aware that performance is reported according to primary evaluation metrics and secondary evaluation metrics. Secondary analysis is presented to supplement the primary analysis. The primary analysis is the most important determinant of a model's performance.

PSNR was computed as a reference measure, and compared to all models. PSNR was computed using an exhaustive search for calibration and one constant delay for each video sequence. Models were required to perform their own calibration, where needed. While PSNR serves as a reference measure, it is not necessarily the most useful benchmark for recommendation of models.

II.3 RR model performance

RR models were submitted by Yonsei for the following resolutions and bit rates: VGA at 128 kbit/s, 64 kbit/s and 10 kbit/s; CIF at 64 kbit/s and 10 kbit/s; and QCIF at 10 kbit/s and 1 kbit/s. When comparing these RR models to PSNR, it must be noted that PSNR is an FR model (i.e., PSNR needs full access to the source video).

II.3.1 Primary Analysis of RR Models

The average correlations of the primary analysis for the RR VGA models were all 0.80, and PSNR was 0.71. Individual model correlations for some experiments were as high as 0.93. The average RMSE for the RR VGA models were all 0.60, and PSNR was 0.71. The average outlier ratio for the RR VGA models ranged from 0.55 to 0.56, and PSNR was 0.62. All proposed models performed statistically better than PSNR for 7 of the 13 experiments. Based on each metric, each RR VGA model was in the group of top performing models the following number of times:

Statistic	Yon_RR10k	YonRR64k	YonRR128k	PSNR
Correlation	13	13	13	7
RMSE	13	13	13	6
Outlier Ratio	13	13	13	10

The average correlations of the primary analysis for the RR CIF models were 0.78, and PSNR was 0.66. Individual model correlations for some experiments were as high as 0.90. The average RMSE for the RR CIF models were all 0.59, and PSNR was 0.72. The average outlier ratio for the RR CIF models were 0.51 and 0.52, and PSNR was 0.63. All proposed models performed statistically better than PSNR for 10 of the 14 experiments. Based on each metric, each RR CIF model was in the group of top performing models the following number of times:

Statistic	Yon_RR10k	YonRR64k	PSNR
Correlation	14	14	5
RMSE	14	14	4
Outlier Ratio	14	14	5

The average correlations of the primary analysis for the RR QCIF models were 0.77 and 0.79, and PSNR was 0.66. Individual model correlations for some experiments were as high as 0.89. The average RMSE for the RR QCIF models were 0.58 and 0.60, and PSNR was 0.72. The average outlier ratio for the RR QCIF models were 0.49 and 0.51, and PSNR was 0.60. All proposed models performed statistically better than PSNR for at least 9 of the 14 experiments. Based on each metric, each RR QCIF model was in the group of top performing models the following number of times:

Statistic	Yon_RR1k	YonRR10k	PSNR
Correlation	14	14	5
RMSE	14	14	4
Outlier Ratio	12	13	4

II.3.2 Secondary analysis of RR models

The secondary analysis shows, in principle, a similar picture. The VGA RR models all tend to perform similarly. The CIF RR models all tend to perform similarly. For QCIF, Yonsei's 10k RR model slightly outperforms Yonsei's 1k RR model. The average correlation coefficients increase to 0.87 for VGA, 0.85 for CIF, and 0.91 for Yonsei's 10k model.

II.3.3 RR model conclusions of VQEG

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

- The secondary analysis requires averaging over a well-defined set of sequences while the tested system including all processing steps for the video sequences must remain exactly the same for all clips. Averaging over arbitrary sequences will lead to much worse results.

It should be noted that in case of new coding and transmission technologies, which were not included in this evaluation, the objective models can produce erroneous results. Here, a subjective evaluation is required.

II.4 Data analysis executed by ILG

Subjective data included virtually in this appendix is being made available by the Video Quality Experts Group (VQEG) to assist the research community. Statistics from the VQEG synopsis can be used in papers by anyone provided that identification that the VQEG synopsis was the source of the data is made explicitly in such papers.

VQEG validation subjective experiment data is placed in the public domain; however, the video sequences themselves are only available for further experiments from the content provider and with restrictions required by the relevant copyright holder for the particular video sequence. VQEG objective validation test data may only be used with the proponent's approval. Interested parties should contact the VQEG for additional information. Nevertheless, any summary data contained in this appendix is available for users of this Recommendation.

The ILG has analysed the data collected by the VQEG and provided it to the ITU for distribution with this Recommendation. The official ILG data analysis is provided in the associated file indicated below.

The associated zip file for Recommendation ITU-T J.246 contains the following files in the Software folder:

- General instructions: readme.txt
- Analysis of data: performance_analysis.xls
- Copyright Notice: copyright_notice.txt

That zip file is available for free download here:

<http://www.itu.int/rec/T-REC-J.246>.

Appendix III

Equations for Model Evaluation Metrics

(This appendix does not form an integral part of this Recommendation)

III.1 Evaluation Metrics

III.1.1 Pearson Correlation Coefficient

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

$$R = \frac{\sum_{i=1}^N (X_i - \bar{X}) * (Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} * \sqrt{\sum (Y_i - \bar{Y})^2}} \quad (\text{III.1})$$

X_i denotes the subjective score (DMOS(i) for FR models) and Y_i the objective score (DMOSp(i) for FR). N in equation III.1 represents the total number of video clips considered in the analysis.

Therefore, in the context of this test, the value of N in equation III.1 is:

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

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

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

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

$$\sigma_z = \sqrt{\frac{1}{N-3}} \quad (\text{III.2})$$

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

$$CI = \pm K1 * \sigma_z \quad (\text{III.3})$$

NOTE 1 – For a Gaussian distribution, $K1 = 1.96$ for the 95% confidence interval. If $N < 30$ samples are used then the Gaussian distribution must be replaced by the appropriate Student's t distribution, depending on the specific number of samples used.

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

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

$$\text{LowerBound} = z - K1 * \sigma_z$$

$$\text{UpperBound} = z + K1 * \sigma_z$$

NOTE 2 – The values of Fisher's z of lower and upper bounds are then converted back to Pearson's R to get the CI of correlation R .

III.1.2 Root Mean Square Error

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

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

$$Perror(i) = DMOS(i) - DMOSp(i) \quad (III.4)$$

where the index *i* denotes the video sample.

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

$$rmse = \sqrt{\left(\frac{1}{N-d} \sum_N Perror[i]^2 \right)} \quad (III.5)$$

where *N* denotes the total number of video clips considered in the analysis and *d* the number of degrees of freedom of the mapping function.

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

In the context of this test plan, the value of *N* in equation III.5 is:

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

The root mean square error is approximately characterized by a $\chi^2(n)$ [B-1], where *n* represents the degrees of freedom and it is defined by equation III.8:

$$n = N - d \quad (III.6)$$

where *N* represents the total number of samples.

Using the $\chi^2(n)$ distribution, the 95% confidence interval for the rmse is given by equation III.7 [B-1]:

$$\frac{rmse * \sqrt{N-d}}{\sqrt{\chi_{0.025}^2(N-d)}} < rmse < \frac{rmse * \sqrt{N-d}}{\sqrt{\chi_{0.975}^2(N-d)}} \quad (III.7)$$

III.1.3 Outlier ratio (using standard error of the mean)

The consistency attribute of the objective metric is evaluated by the outlier ratio (OR) which represents the number of "outlier-points" to total points *N*:

$$OR = \frac{TotalNoOutliers}{N} \quad (III.8)$$

where an outlier is a point for which

$$|Perror(i)| > K2 * \frac{\sigma(DMOS(i))}{\sqrt{Nsubjs}} \quad (III.9)$$

where $\sigma(DMOS(i))$ represents the standard deviation of the individual scores associated with the video clip *i*, and *Nsubjs* is the number of viewers per video clip *i*. In this test plan, a number of 24 viewers (*Nsubjs*=24) per video clip was used.

NOTE 1 – DMOS(i) is used for FR models.

NOTE 2 – For a Gaussian distribution, $K2 = 1.96$ for the 95% confidence interval. If the mean (DMOS) is based on less than thirty samples (i.e., $N_{\text{subjs}} < 30$), then the Gaussian distribution must be replaced by the appropriate Student's t distribution, depending on the specific number of samples in the mean [B-1]. In the case of 24 viewers per video (i.e., the number of samples in the mean is 24), the number of degrees of freedom is $df=23$ and therefore the associated $K2 = 2.069$ is used for the 95% confidence interval.

Therefore, in the context of this test plan, $K2 = 2.069$.

The outlier ratio represents the proportion of outliers in N number of samples. Thus, the binomial distribution could be used to characterize the outlier ratio. The outlier ratio is represented by a distribution of proportions [B-1] characterized by the mean p (equation III.10) and standard deviation σ_p (equation III.11).

$$p = \frac{\text{TotalNoOutliers}}{N} \quad (\text{III.10})$$

$$\sigma_p = \sqrt{\frac{p*(1-p)}{N}} \quad (\text{III.11})$$

where N is the total number of video clips considered in the analysis.

For $N > 30$, the binomial distribution, which characterizes the proportion p , can be approximated with the Gaussian distribution. Therefore, the 95% confidence interval (CI) of the outlier ratio is given by equation III.12.

$$\text{CI} = \pm 1.96*\sigma_p \quad (\text{III.12})$$

NOTE 3 – If the mean is based on less than thirty samples (i.e., $N < 30$), then the Gaussian distribution must be replaced by the appropriate Student's t distribution, depending on the specific number of samples in the mean [B-1].

Bibliography

- [B-1] Spiegel M., "*Theory and problems of statistics*", McGraw Hill, 1998.

SERIES OF ITU-T RECOMMENDATIONS

Series A	Organization of the work of ITU-T
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	Telecommunication management, including TMN and network maintenance
Series N	Maintenance: international sound programme and television transmission circuits
Series O	Specifications of measuring equipment
Series P	Terminals and subjective and objective assessment methods
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, open system communications and security
Series Y	Global information infrastructure, Internet protocol aspects and next-generation networks
Series Z	Languages and general software aspects for telecommunication systems