15

Video Codec Evaluation Scheme and Implementation Based on Characteristics of Human Visual Perception

KDD R&D laboratories
2-1-15 OHARA KAMIFUKUOKA, SAITAMA 356, JAPAN
ta-hamada@ kdd.co.jp

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 Classes I, II, III (Contribution, Primary and Secondary distribution), it is required to strive to achieve consistently high quality by constantly monitoring the quality of the transmitted pictures. In conventional analog FM transmission, there is little degradation in the picture due to the contents or to analog 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, KDD proposes a scheme to evaluate the picture quality of MPEG-2 based video codecs mainly used in Classes I, II, III. In these classes, following functions are considered to be necessary.

• Generic assessment for various types of video contents

• Analog/Digital•Composite/Component video formats are supported

• Real time assessment

•Precise temporal & spatial alignment between an original and a codec out signal

•Sensitive and accurate assessment to subtle and complex distortions

Considering above, we offer a new evaluation scheme and its implementation based on the characteristics of human visual perception, enabling very precise measurements of video quality.

2. Evaluation scheme based on characteristics of human visual perception

2.1 Three-layered bottom-up noise weighting model

Figure 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,

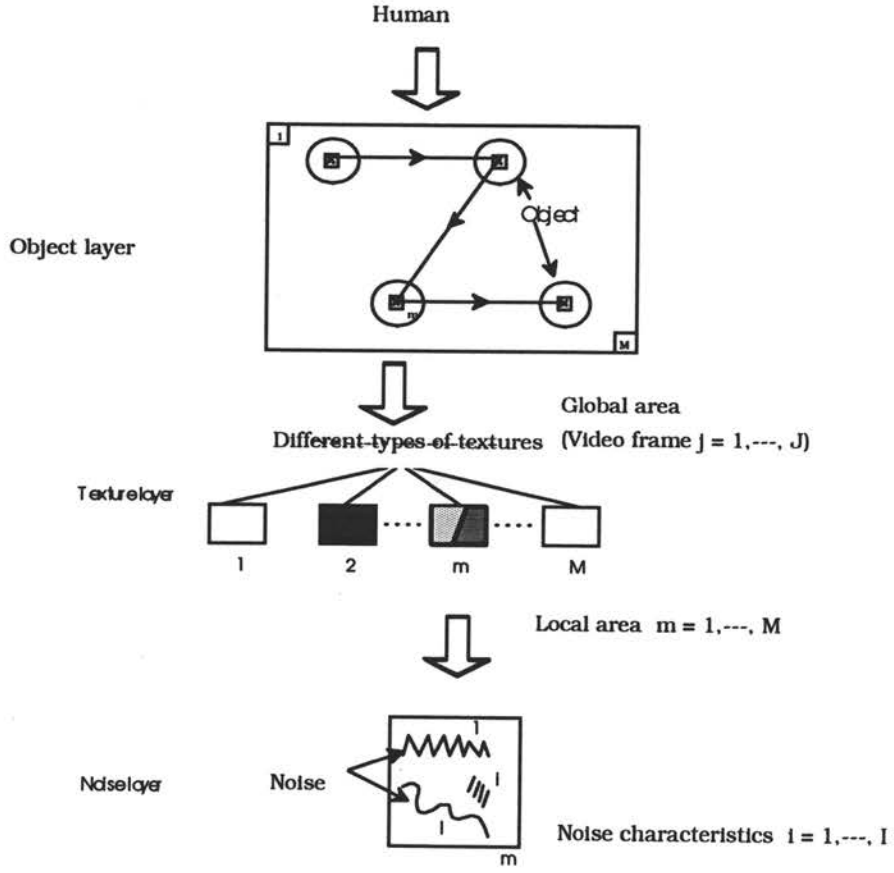


Figure 1 Three-layered model for video signal

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 2).

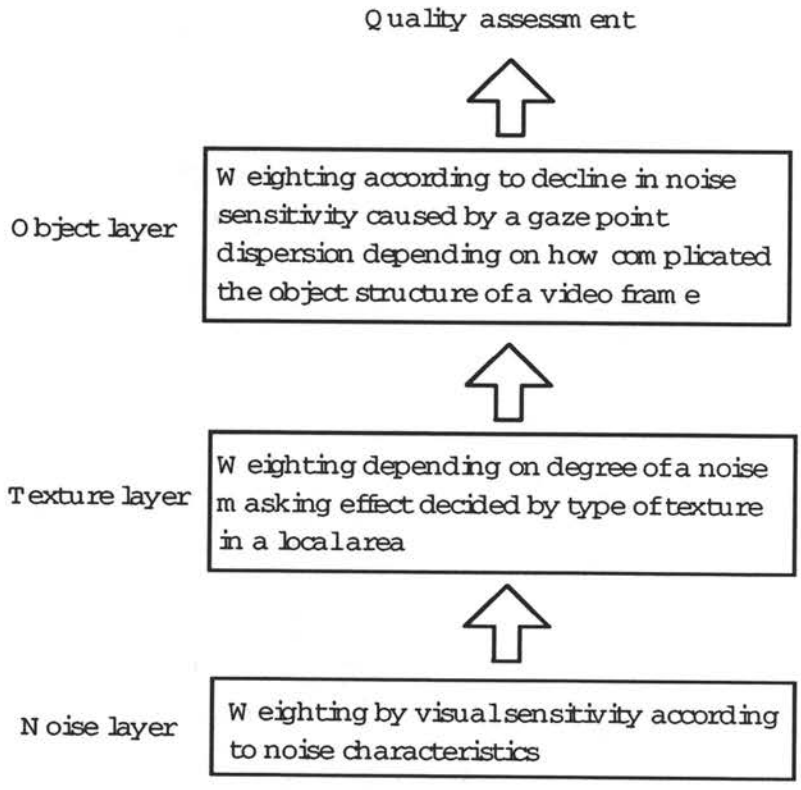


Figure 2 Three-layered bottom-up noise weighting

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 noise are strongly masked, and "flat texture" such as a human skin and a sky in which noise 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
hi : Weighting function for a noise i
C(j,m) : Texture in a local area (j,m)
tc : Noise weighting function in a texture C
G(j) : Parameter indicating how complicated the structure is of objects of a frame j
qG : 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 hi corresponding to noise characteristics in a local area (j,m), we calculate $WMSE_{NL}$ as,

$$WMSE_{NL}(j,m) = \frac{1}{I}\sum_{i=1}^{I} hi \cdot P(j,m,i) \qquad (1)$$

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

$$WMSE_{TL}(j) = \frac{1}{M}\sum_{m=1}^{M} t_c(j,m) \cdot WMSE_{NL}(j,m) \qquad (2)$$

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,

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

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

$$WSNR(dB) = 10\log_{10}\frac{255^2}{WMSE} \quad (4)$$

$$D(\%) = f(WSNR) \quad (5)$$

2.2 Parameters for quality assessment

We give detailed expressions of the formulas (1)-(3) to apply the proposed model to NTSC as a sample television signal. For component TV signals, it is feasible to take the same approach.

First, as a local spot area, we choose a small 8 pixel*8 line size square block. We apply a frame-based orthogonal transform such as DCT to convert this pixel domain block into a frequency domain block and classify noises according to their characteristics. For example, in case of evaluation of NTSC signals by WHT (Walsh Hadamard Transform) which is very suitable for expressing the NTSC noises in a transform domain, the common NTSC noises include DC and AC luminance noises, color sub-carrier noises, chrominance noises, flickers, jerkiness and so on as shown in Figure 3 and Table 1. In

0	1	2	3	4	4	5	5
1	2	3	5	11	6	6	7
2	3	5	10	9	11	7	7
3	5	10	8	8	9	7	7
4	6	6	6	7	7	7	7
4	6	6	7	7	7	7	7
12	13	7	7	7	7	7	7
12	12	13	13	7	7	7	7

Figure 3 Noise classifications in Hadamard transform domain (I)

these figures and tables, i=0 is luminance DC noise, i=1-7 are luminance AC noises, i=8 are color sub-carrier noises, i=9-11 are chrominance AC noises, i=12 are flickers and i=13 are jerkiness. The weighting function matrix obtained for NTSC signals in an (8x8) Hadamard Transform Domain is shown in Figure 4.

Table 1 Classifications of noise characteristics

Cluster	Noise type
0	Luminance DC noise
1	Luminance 1st AC noise
2	Luminance 2nd AC noise
3	Luminance 3rd AC noise
4	Luminance 4th AC noise
5	Luminance 5th AC noise
6	Luminance 6th AC noise
7	Luminance 7th AC noise
8	Color sub-carrier noise
9	Chrominance 1st AC noise
10	Chrominance 2nd AC noise
11	Chrominance 3rd AC noise
12	Flicker
13	Jerkiness

In order to determine a texture type C(j,m) in a local area (j,m) and it's a weighting function tc, the noise masking effect is introduced, in which noise is perceived differently according to how actively an original signal changes in a local area. Regarding this noise masking effect, approaches to measure it based on signal changes between pixels

1.00	1.00	0.75	0.60	0.50	0.50	0.40	0.40
1.00	0.75	0.60	0.40	0.30	0.30	0.30	0.15
0.75	0.60	0.40	0.60	0.75	0.30	0.15	0.15
0.60	0.40	0.60	1.00	1.00	0.75	0.15	0.15
0.50	0.30	0.30	0.30	0.15	0.15	0.15	0.15
0.50	0.30	0.30	0.15	0.15	0.15	0.15	0.15
0.75	0.50	0.15	0.15	0.15	0.15	0.15	0.15
0.75	0.75	0.50	0.50	0.15	0.15	0.15	0.15

Figure 4 Weighting function matrix hi reflecting visual sensitivity in (8×8) Hadamard Transform Domain

have been reported. But in the 8*8 block case, it is considered that a pixel-based

signal change is insufficient and block-based quantity should be used to measure a masking effect.

The C(j,m) is therefore defined as an average of AC coefficients power in a block, and the function tc is follows;

$$t_{c(j,m)} = \frac{1}{\sqrt{2..02 \log_{10} C(j,m)}} \tag{6}$$

where noises are more strongly masked for C(j,m) grows with a flat, a gradient and a detail texture block.

The parameter G(j) indicating a frame activity is defined as an average value for C(j,m) over a whole frame. The parameter qG is determined as follows. Namely, to several test pictures which have different values of G, we added noises so that noise power weighted by tc and hi in formulas (1)-(3) are equal among the test pictures. We measured the subjective assessment values (%) by subjective tests and normalized these in the 0 to 1 range (Figure 5). In Figure 5, a linear relation between G and qG is extracted, and qG is approximated by attaching the least mean square error as shown in eq.(7).

$$q_G = -0.0018G + 0.806 \tag{7}$$

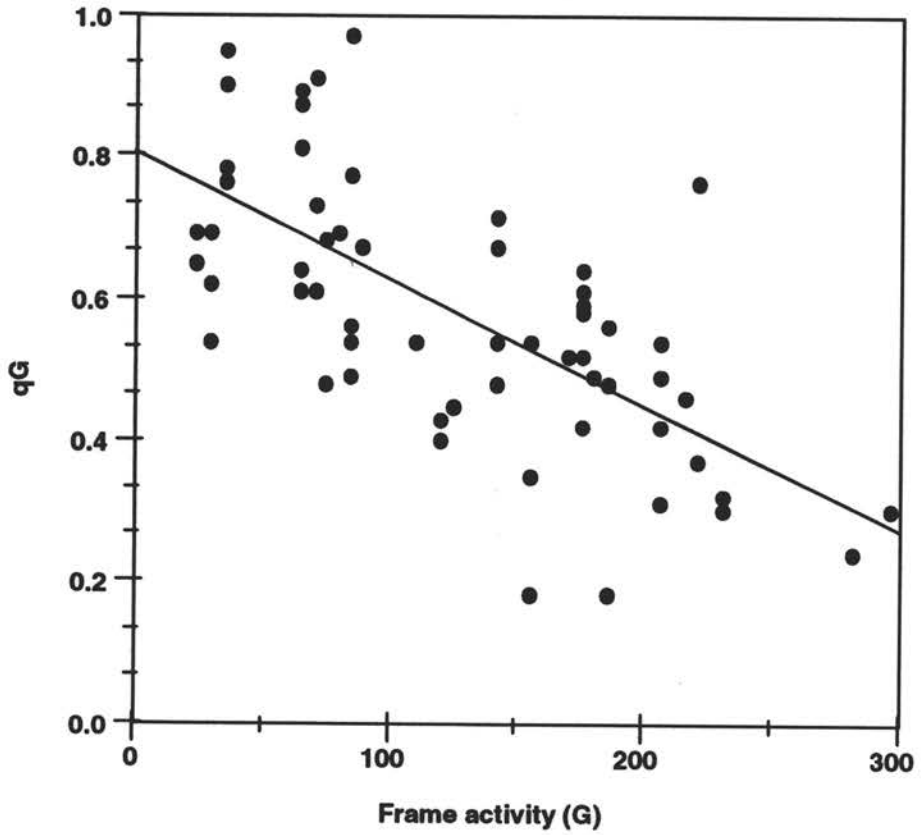


Figure 5 Frame activity G v.s. qG

To decide the function f in the formula (5), the approach is taken of giving a polynomial approximation to a non-linear relation between the WSNR derived from the formula (4) and the subjective assessment value (%) obtained by the ITU-R Rec. 500-7 experiment. The polynomial is determined by increasing the order of the polynomial and by deciding coefficients for each order to minimize the approximation errors. As a result, a saturation in approximation errors occurs at the fifth order and f is given as follows;

$$f(x) = -7.32\times10^{-6}x^5 + 1.22\times10^{-3}x^4 - 7.29\times10^{-2}x^3 + 1.8x^2 - 14.5x \qquad (8)$$

,where x=WSNR.

3. Implementation of the evaluation system

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

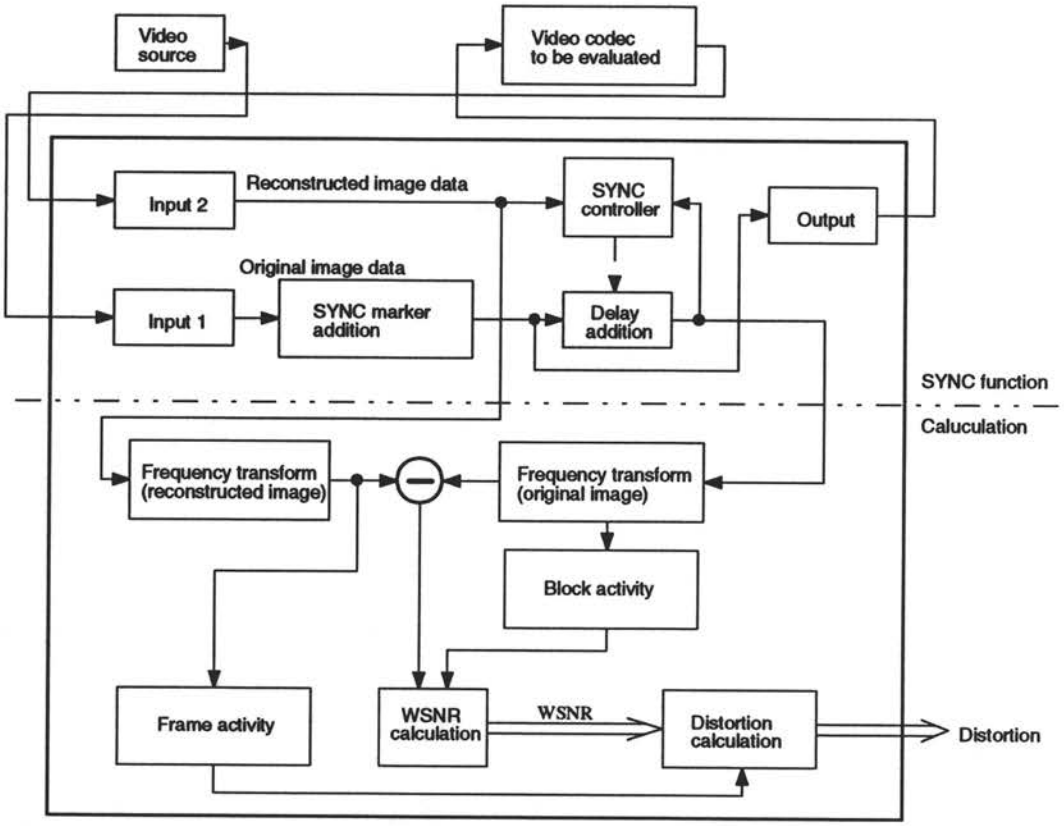


Figure 6 System Configuration

module, which determines video quality with reference to characteristics of human visual perception. Figure 6 shows the configuration of the system. And Table 2 describes principal parameters. As Table 2 shows, both composite (NTSC)/component signals with full samplings are supported.

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 frame are sent via the output module to the video codec that is to be evaluated. The video codec compresses the frame, 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 & 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.

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 2 Principal parameters

Applicable video signal format	NTSC composite signal 525/60 component signal D1 serial digital
Sampling frequency (Analog input)	14.318MHz (NTSC) 13.5MHz (Component Y) 6.75MHz (Component C)
Applicable codec	MPEG-1,2 based codec Composite codec etc.
Effective evaluation area	768pixels•480lines (NTSC) 720pixels•480lines (Component Y) 360pixels•480lines (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

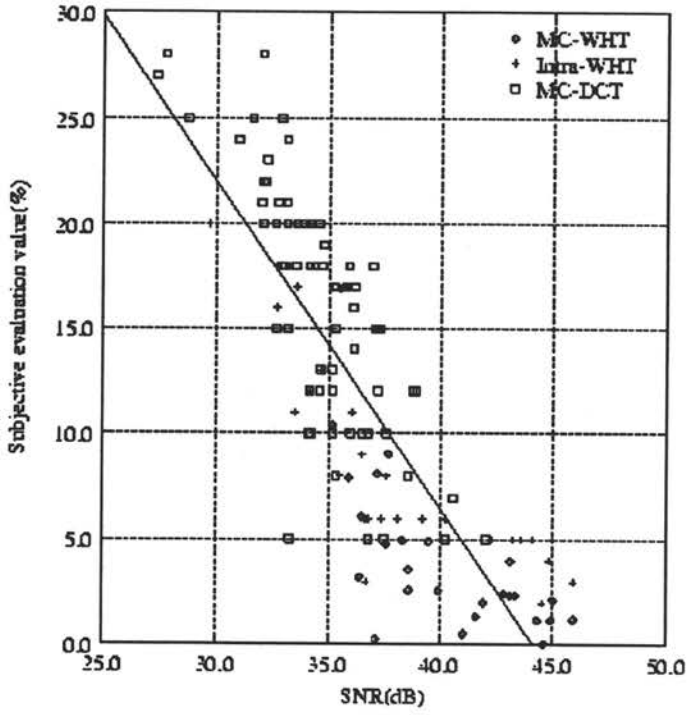
4. Comparisons with subjective assessment tests

4.1 Composite TV signal (NTSC) test

We prepared 16 NTSC test materials for evaluation of the described scheme and three different types of video codecs were tested, which are the MC DCT MPEG-2 codec (including color encoding/decoding functions, 12Mbps and 9Mbps), the Intra-frame WHT composite codec (22Mbps and 15Mbps), and the MC WHT composite codec (22Mbps and 15Mbps).

As assessment references, subjective tests were also conducted based on ITU-R Rec. 500-7 with the same test materials and 20 expert viewers.

The SNR and the assessment values by this system versus subjective quality (ITU-R) are shown in Figure 7 and Figure 8 respectively. In these figures, picture



quality is

Figure 7 SNR(dB) vs. Subjective assessment value(%) by ITU-R BT.500-7

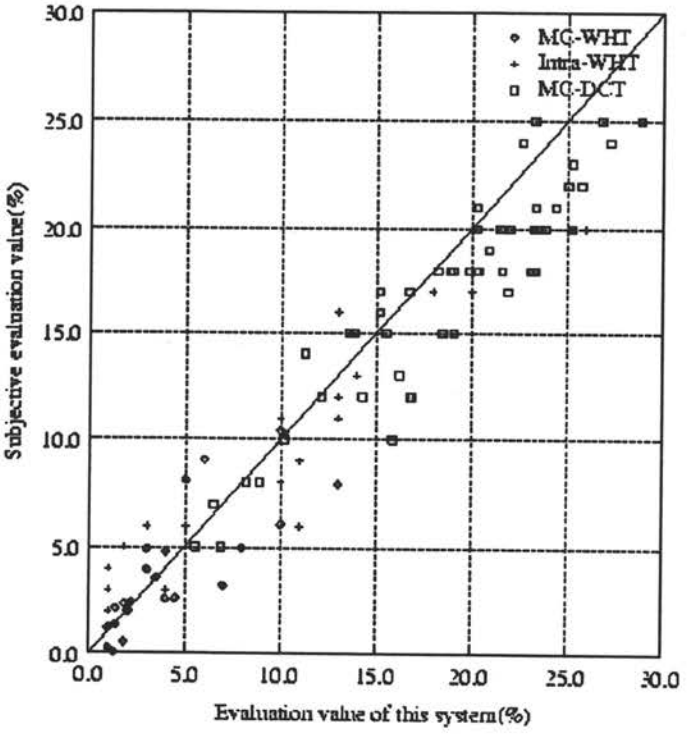


Figure 8 Output(%) by the proposed scheme(Hardware) vs. Subjective assessment value(%) by ITU-R BT.500-7

scaled from 0-100%, in which 0% indicates there is no difference between the codec output signal and original signal. Also in Table 3, the average and maximum values of

Table 3 Assessment error from ITU-R Rec.500-7

	Error (%)			
	SNR		Proposed scheme (Hardware)	
Range (%)	$\sqrt{MSE}$	Worst	$\sqrt{MSE}$	Worst
0~5	4.76	10.9	1.65	3.88
5~10	4.56	12.1	2.79	6.35
10~15	3.27	5.78	2.09	4.25
15~20	3.19	6.68	2.68	6.59
20~25	3.63	6.71	1.73	4.26
25~30	5.61	9.10	4.47	7.33
Total	4.25	12.1	2.45	7.33

root mean square error from linear approximations in these figures are shown for each 5% subjective quality range (%). Generally, plots are more widely scattered in Figure 7 than in Figure 8. More precisely, the system gives smaller values both in average and maximum values than the SNR in every range. Particularly, much improvement in accuracy can be obtained at 0 to 10% range where high quality video codecs are assessed for use in applications such as contribution and primary distribution.

A correlation for each layer is calculated as shown in Table 4, where 1.00 indicates a perfect correlation and 0 indicater there is no correlation. The SNR gives a low

Table 4 Assessment accuracy at each layer (correlation)

	Assessment accuracy (correlation)			
Compression scheme	MSE(SNR)	WMSE$_{NL}$	WMSE$_{TL}$	WMSE$_{OL(D)}$
MC-WHT	0.680	0.697	0.716	0.804
Intra-WHT	0.743	0.768	0.808	0.912
MC-DCT	0.772	0.790	0.847	0.891

correlation between 0.68-0.77 because human visual perception is not considered. But by taking human visual perception into account at each layer, correlation grows higher and finally reaches 0.8-0.9, regardless of the type of video codecs.

Six modes (MC-WHT-22M, MC-WHT-15M, Intra-WHT-22M, Intra-WHT-15M, MC-DCT-12M, MC-DCT-9M) are also graded by the SNR, by the proposed scheme and by the ITU-R Rec.500-7 assessments. The "grading error" is defined here as the total absolute difference value from the Rec.500-7 grading. For example, if the SNR grades six modes as 341265 and the Rec.500-7 grades then as 143625, the grading error is 2+0+2+4+4+0=10. An averaged grading error over 16 materials is shown in Table 5. A result of 3.63 by SNR grading means it often happens that a lower grade codec is graded as a high grade or vice versa. This is a very dangerous measurement for a codec discrimination. 1.50 by the proposed scheme, can be regarded as minor errors among codecs having nearly equal performance.

Table 5 Grading error

Scheme	Grading error
SNR	3.63
Proposed scheme	1.50

From the evaluation results mentioned above, it is concluded that, by the proposed assessment scheme, efficient and practical solutions can be obtained for objective assessments of subjective picture quality of a variety of video codecs using different kinds of test materials.

4.2 Component TV signal test

We compared the evaluation results by proposed scheme with subjective assessment test results which are already graded following ITU-R Rec.500-7. Assessment targets are MPEG-2 SP@ML with 5Mbps, 7Mbps and 10Mbps applied for ITU-R Rec.601, 4:2:2 component TV test signals. These are 21 data including Mobile, Flower garden, Cheer leaders etc. Therefore, we have totally 21 data x 3 bit-rate = 63 samples.

For these samples, we conducted subjective assessment test on two different days (March 23 & 24, 1995) with same conditions and viewers. The "triangle" of the proposed scheme and two days assessment results are shown in Figure 9. This figure proves that assessment accuracy expressed by rmse, rwse and correlation

of three assessment results are nearly equal from the triangle center, which is the true assessment value.

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

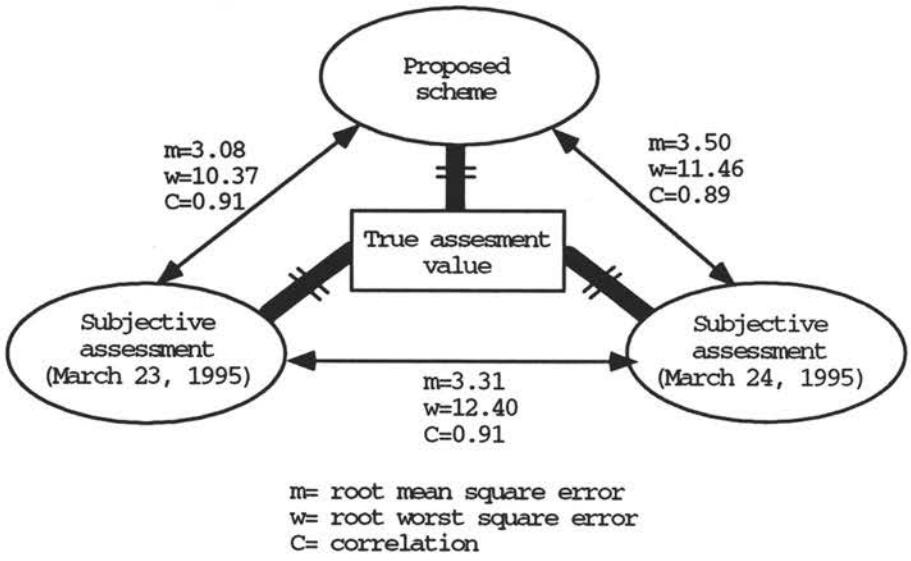


Figure 9 Comparisons with Subjective Assessment Tests

5. Using the system

Following area a few examples of how the system could be used. First, the configuration shown in Figure 10 could be used to evaluate the quality of individual video codecs. This would enable users to determine whether a particular codec has the performance to meet their quality requirements. At the same time, it would clarify the

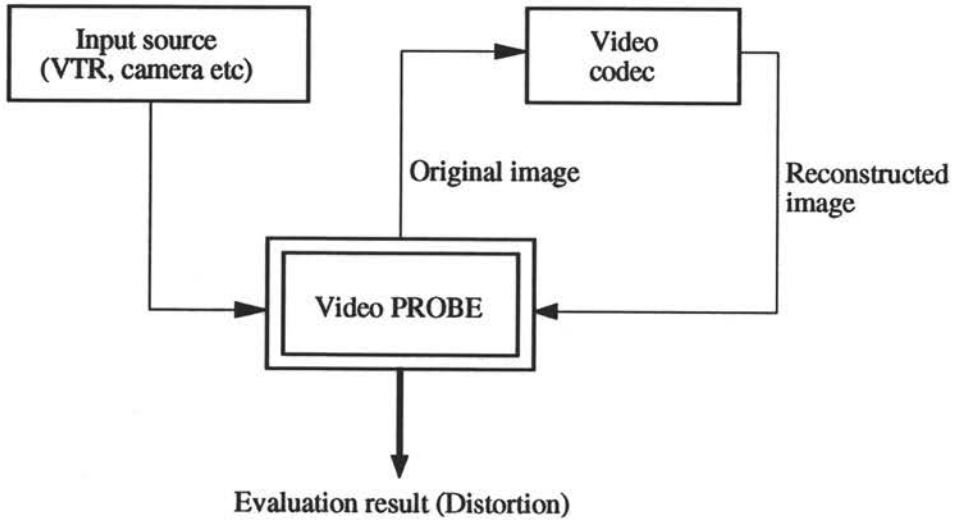


Figure 10 Evaluation of Video Codec Quality

types of content for which the codec is best and least suited.

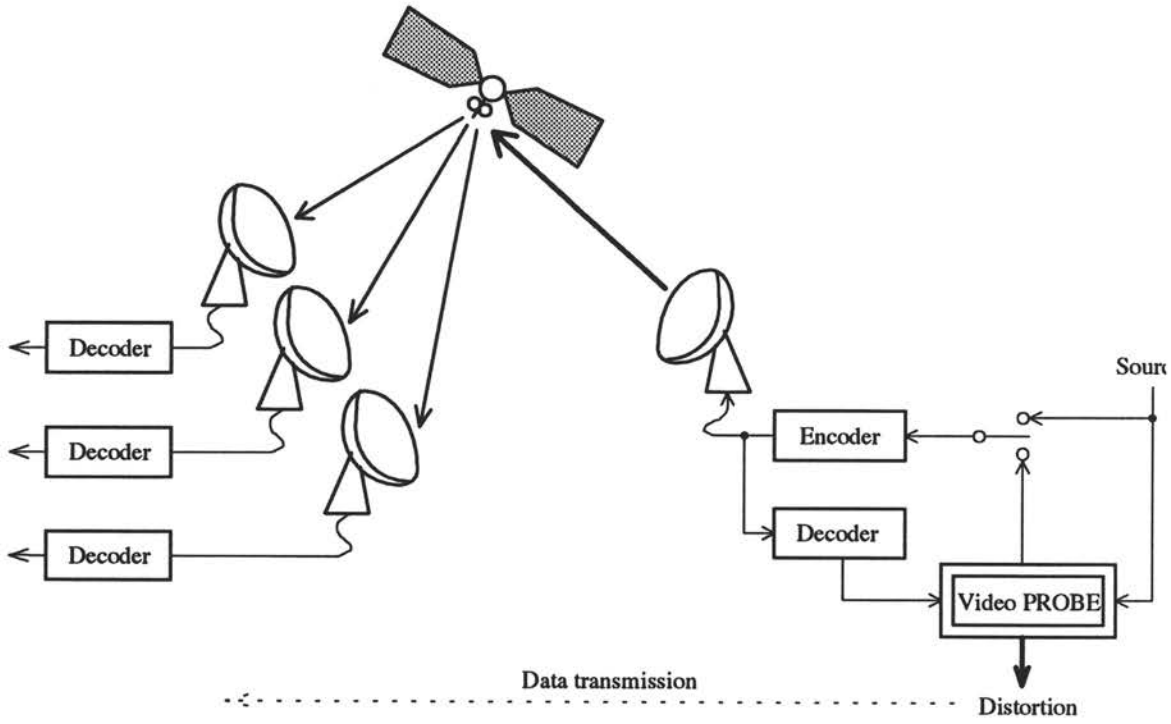


Figure 11 Quality Management in Digital Broadcasting

The system could also be used to monitor the quality of video transmission. When a codec is used in digital broadcasting or transmission of digital television, the quality of the received video depends on compression operations performed by the encoder on the sending side. This means that it is important to check on the transmitting side. As shown in Figure 11, a monitor decoder could be installed on the sending side so that this system could be used to measure the amount of degradation in the compressed and reconstructed video. This would make it possible to systematically manage the quality of transmission video. The results of evaluation with this system could also be transmitted as data so that the receiving side could monitor transmission quality at the same time. In addition to regular monitoring of compression distortion on the sending side, the receiving side could use the evaluation results of this system to switch over to other channels in the case of transmission faults.

6• Display of evaluation results

The evaluation results output by this system are displayed as percentage values representing the amount of degradation from the original picture, where 0% indicates no degradation. Figure 12 shows an example. Displayed on the same screen are a

graph of results over time, comments, and an indication of how difficult the video is to compress.

The quality evaluation results of this system have been shown to be very close to

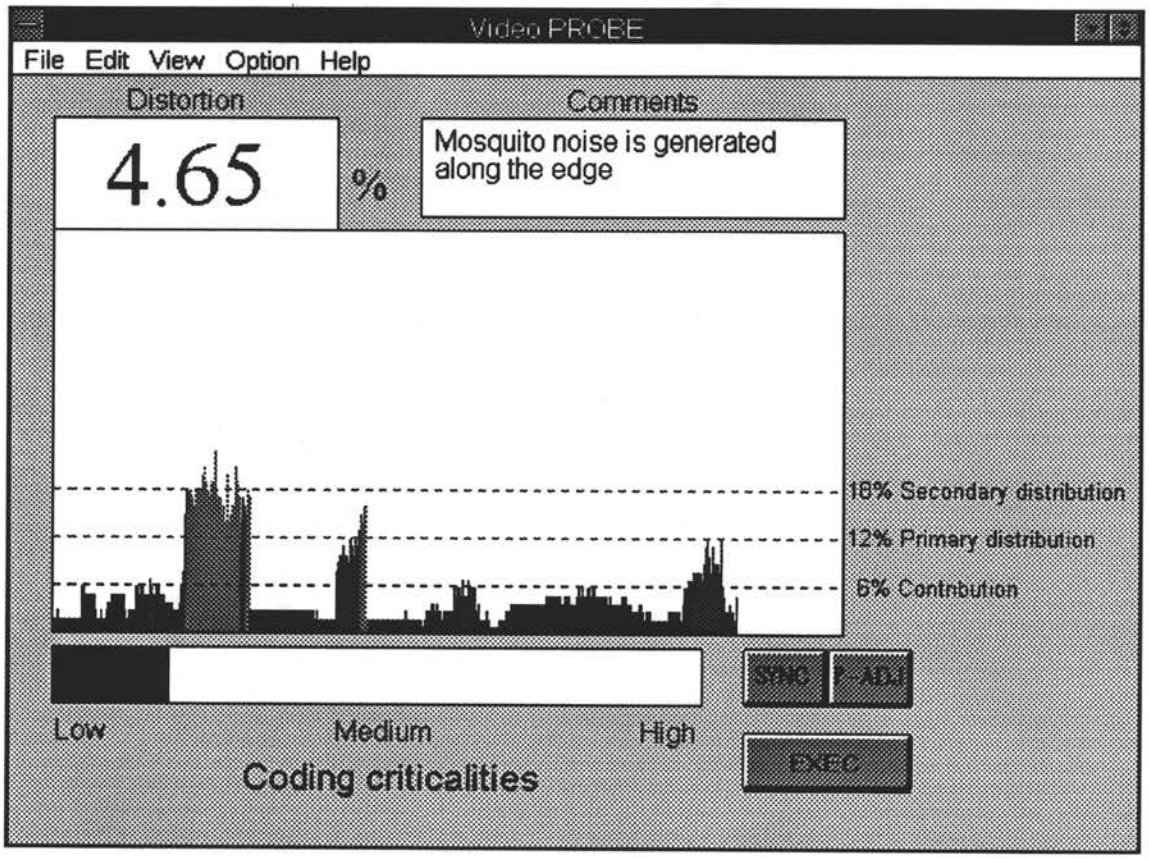


Figure 12 Display of Evaluation Results

evaluations performed by humans who actually view the picture. This indicates that its accuracy has been greatly enhanced by the incorporation of the three major characteristics of human visual perception.

This system makes it possible to perform precise, real-time evaluation of the complex and constantly changing quality of compressed digital video. It is expected to become a valuable tool for providing stable, high quality transmission of television signals in the fields of digital broadcasting and digital television transmission.