ELSEVIER

Signal Processing 70 (1998) 279–294

SIGNAL PROCESSING

An objective measurement tool for MPEG video quality

K.T. Tan*, M. Ghanbari, D.E. Pearson

Department of Electronic Systems Engineering, University of Essex, Wivenhoe Park,Colchester CO4 3SQ, UK

Received 30 July 1998

Abstract

A two-stage objective measurement model for MPEG-coded video is proposed. The first stage weights the coded video distortion according to the human visual system's response. It computes the frame-by-frame perceptual impairment in the decoded picture with respect to a reference picture; this includes low-pass spatial filtering, a Sobel operation to derive masking coefficients, and spatial masking on the raw error between reference and compressed pictures. The second stage, a cognitive emulator, provides a simulation of human high-level processing of visual information. This includes the very low temporal response of human viewers to image quality changes, and asymmetric behaviour in respect of picture quality changes from bad to good, and vice versa. With this model, we have been able to mimic quite accurately the temporally varying subjective picture quality of video sequences as recorded by the ITU-R SSCQE method.

Zusammenfassung

In dieser Arbeit wird ein zweistufiges objektives Meßmodell für MPEG-codierte Videobilder vorgeschlagen. Die erste Stufe gewichtet die Verzerrung des codierten Videobildes entsprechend den Eigenschaften des menschlichen visuellen Systems. Dabei wird Frame für Frame die perzeptuelle Verschlechterung in decodierten Bild bezüglich eines Referenzbildes berechnet; diese Berechnung beinhaltet eine räumliche Tiefpaßfilterung, eine Sobel-Operation zur Ableitung von maskierungskoeffizienten sowie eine räumliche Maskierung des ursprünglichen Fehlers zwischen Referenzbild und komprimiertem Bild. Die zweite Stufe, ein Kognitiver Emulator, simuliert die menschliche höherstufige Verarbeitung von visueller information. Diese Stufe berücksichtigt die sehr geringe zeitliche Reaktion des menschlichen Betrachters auf Änderungen der Bildqualität sowie das asymmetrische Verhalten bezüglich Veränderungen von schlechter zu guter Qualität und umgekehrt. Mit diesem modell konnten wir die zeitlich variierende subjektive Bildqualität von Videosequenzen, wie sie mit der ITU-R SSCQE-Methode aufgezeichnet wurde, recht genau nachahmen.

Résumé

Un modèle de mesure objective en deux étapes pour les vidéos codées MPEG est proposé dans cet article. Dans la première étape on pondère la distortion de la vidéo codée en fonction de la réponse du système visuel humain. La déterioration perceptuelle trame par trame dans l'image décodée vis-à-vis d'une image de référence est calculée; ce calcul inclut un filtrage spatial passe-bas, une opération de Sobel pour dériver les coefficients de masquage, et un masquage

* Corresponding author.

0165-1684/98/$ – see front matter
PII: S0165-1684(98)00129-7

spatial sur l'erreur brute entre images de référence et comprimée. Dans la seconde étape, l'émulation cognitive, prend place une simulation du traitement haut-niveau humain de l'information visuelle. Ceci inclut la réponse temporelle très faible des observateurs humains aux changements de qualité d'image, et un comportement asymetrique vis-à-vis des changements de qualité d'image de mauvais à bon et vice versa. Avec ce modèle, nous avons été capables d'imiter très précisément la qualité subjective variant dans le temps de séquences vidéo enregistrées avec la méthode ITU-R SSCQE.

1. Introduction

Digital image compression technology makes digital image communications possible and efficient. Image compression could be lossless or lossy, depending on the application. Whilst lossless compression allows perfect recovery of the digital image, lossy approaches offer much higher compression ratio which is very attractive to image archiving where storage capacity is a main concern, or digital video transmission where channel bandwidth is precious. However, high compression ratio is very often associated with poorer picture quality. In most applications, picture quality is a key factor, and hence knowing the perceived quality of the digital pictures is advantageous. Over the years, many methodologies have been proposed to measure the quality of digital pictures. This could involve human observers, which is a very straightforward approach as human beings are the ultimate end users of digital imaging systems; or it could involve computers running very sophisticated algorithms to estimate the picture quality that would be perceived by human observers. The former class, known as subjective assessments, has been used in practice for years. Under strictly controlled conditions, subjective assessment methodologies could yield very accurate and reliable result. Its shortcoming is the cost and time required to prepare and conduct the subjective assessment. The latter class is known as objective measurements, using mathematical model to simulate the human visual system. This approach is relatively faster and cheaper than subjective assessments, and also offers the possibility of being incorporated into digital video system. We will first review some subjective assessment methodologies and their associated problems, and then present the objective model we propose.

2. Subjective assessment methodologies

Since human beings are the ultimate end users of many digital video systems, intuitively human observers should be used to judge the quality of digital images. There are three classes of subjective assessment methodologies: single stimulus methods, comparison methods and double stimulus methods.

In single stimulus assessment methods, subjects are presented with a series of independent video sequences, and at the end of each presentation, the subjects are asked to give an overall rating of the quality of the preceding presentation. This is best described by Fig. 1.

The rating scale used is typically the descriptive 5-grade category scale [4]: excellent, good, fair, poor and bad. The problem of this method is a well-known phenomenon called adaptation [5], where the grating process of picture quality by the subjects are very much affected by quality of the preceding pictures, hence making the order of the presentations very critical. There are suggestions that inserting anchor pictures (presentation with best- and worst-quality levels) may reduce the effect of adaptation, but this raises more questions on

Fig. 1. Single stimulus method.

insertion of anchors, for example, how frequent should the anchor stimuli be presented?

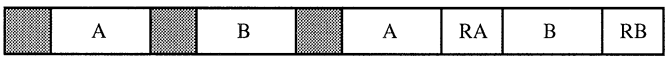
Another member of single stimulus method is single stimulus continuous quality evaluation (SSCQE) [3,9] developed under RACE project MOSAIC, which has now been accepted as a standard by ITU-R [18]. In this methodology, the human evaluators are asked to adjust a slider mechanism according to the variation in the picture quality during the test. The position of the slider is sampled typically at 2.5 Hz, and the reading is normalised to a range of 0.0–1.0. Aldridge et al. [3] have demonstrated the repeatability and stability of SSCQE method in recording the temporal variations in subjective quality. However, one problem associated with SSCQE is the lack of reference, hence making comparison of different continuous quality measurements difficult. Aldridge [1] addressed the need to anchor continuous quality measurement to a common reference, and proposed some calibration to be carried out. Despite this problem, SSCQE does offer some very attractive advantages, which will be discussed after the presentation of double stimulus methods.

Comparison methods present pairs of pictures contaminated by different levels of distortion to human evaluators, and the subjects make relational judgements between the two stimuli using a 7-grade categorical scale: much better, better, slightly better, the same, slightly worse worse and much worse. The popularity of comparison methods has been declining, mainly due to its failure in providing meaningful distance information between two stimuli.

Double stimulus methods are strictly speaking, another form of paired comparison method, but with constant reference. There are two forms of double stimulus methods: double stimulus impairment scale (DSIS) and double stimulus continuous quality scale (DSCQS). In double stimulus methods, reference pictures are presented together with impaired pictures. This provides the evaluator a constant quality level functioning as an anchor. Regular presentation of anchors throughout a double stimulus test session helps to alleviate the adaptation problem suffered by all the single stimulus methodologies. This offers a consolidated frame of reference within which picture quality judgements are made, thus allowing double stimulus methods to yield stable and reliable results [15]. In this paper, we will only discuss DSCQS, and the reader is referred to [18] for details on DSIS.

DSCQS [18] has become the most popular approach used in evaluating subjective picture quality. Fig. 2 illustrates the general format of DSCQS, and the 5-grade category scale used for rating is given in Fig. 3. The stimuli are 10 s in duration, and each pair of reference and test pictures (A and B in Fig. 2) are presented twice to the human evaluators. At the end of each second presentation, the subjects are asked to give a retrospective rating of the preceding presentation (RA and RB, in Fig. 2). The order of the reference and test pictures within a pair is randomised, and the raters are not informed of the order.

The DSCQS method has been used in so many experiments of picture quality evaluation that it has become the bench mark on which objective measurements models' performance are always compared to. However there are arguments about the suitability of DSCQS in evaluation of video sequences quality. Lodge [13] argues that the 10 s



3-5 sec containing test number

A, B = 10 sec sequences

RA, RB = 5-11 sec during which subjects rate for sequences A and B respectively

Fig. 2. Presentation structure for DSCQS method.

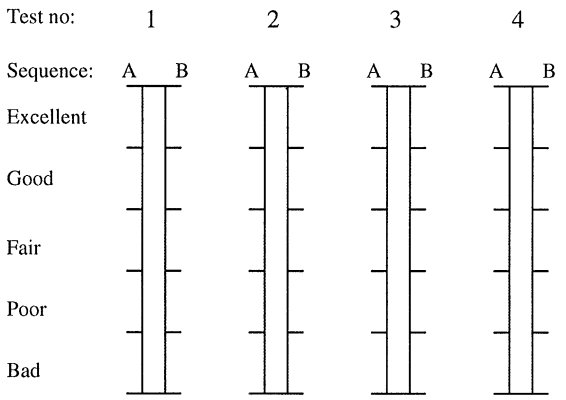


Fig. 3. Five-grade continuous category rating scale used in DSCQS method.

duration sequences used in DSCQS method are too short to be statistically-representative selection of scene content to be assessed. As a result, the process of selection of test sequences could steer the outcome of comparison between different coding schemes into favour of a particular coder. Using longer sequences may alleviate this problem, but there is a worry about recency effect in evaluating long-sequence quality using DSCQS [2,10]. Repetition of the test sequence pairs also raises the worry that subjects may identify the artefacts during the first presentation, and focus on the impaired section of the picture during the second presentation, which can result in underestimated judgement of the picture quality. Finally, there are also concern about the use of reference material, which is a test environment very different from home viewing situation.

Compared to DSCQS, the SSCQE has the advantage of better simulation of home viewing environment, where the programme is watched only once without the reference material. The SSCQE also offers more detailed information of the picture quality variation, which is recorded continuously, compared to the single rating given by DSCQS, averaged over the test sequence duration. Use of longer test sequences in SSCQE also does not suffer from recency effect [10], therefore allowing more statistical-representative set of picture quality variation to be evaluated.

3. Video distortion meter

Subjective assessment methodologies are generally time consuming and expensive. The advantages of having a computational image quality metric are then obvious: faster and cheaper evaluation of picture quality for selection of codecs or bit-rates; the possibility of incorporating quality control into coding process for optimisation; and for the world of image coding, a tool more reliable than PSNR for the researchers to assess their progress. However, a common tool covering this wide range of applications will be too complex, if not unrealistic. Therefore, over the years, this vast desire in evaluating picture quality objectively has spurred so many objective measurement models being proposed. These models differ very much in terms of approaches, application areas and complexity. Some examples are Lubin [14], Boch [6], Horita [11] and van den Branden Lambrecht [7].

Many models available today are mainly designed to return a single rating representing the quality of 10 s sequence. The disadvantage of having only a single score is the loss of information about the variation of picture quality within the sequence. As a result, an encoder that causes bursty impairment but otherwise fairly good quality may be wrongly interpreted as better than another encoder that generates slight but evenly distributed distortion over the sequence. Hence, it is important in some applications to have full details of the picture quality variation. There have been some attempts [17] to develop video distortion meters that produce an output trace showing the temporal variation of picture quality of the video under test, but so far there is little effort in simulating the mental processes involved in judgement of picture quality.

The accuracy of objective models is usually judged by the degree of correlation between the objective model output and the subjective data. For continuous video quality evaluation, the subjective assessment tool available is SSCQE. We therefore propose a video distortion meter [8,19] which emphasises on simulating the process of image quality evaluation using SSCQE methodology. The outline of the model is illustrated in Fig. 4. The human visual system properties are included in the

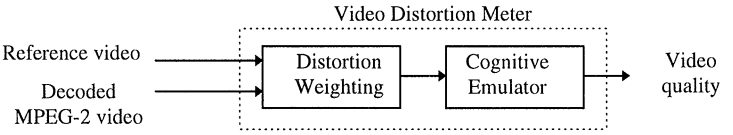


Fig. 4. Outline of a video distortion meter.

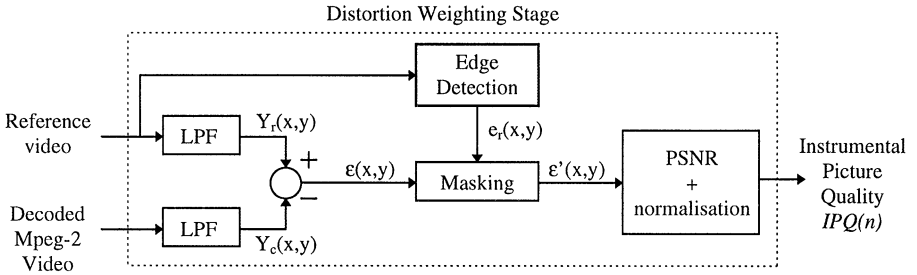


Fig. 5. Distortion weighting stage.

distortion weighting stage. This stage measures the frame-by-frame perceptible distortion in the decoded MPFG-2 picture with respect to the reference picture. The output of the distortion weighting stage, denoted as the instrumental picture quality (IPQ), is a quality metric ranging from 0.0 to 1.0, with 0.0 represents the worst quality, and 1.0 the best quality. If the input video sequences are of frame rate of 25 frames/s, it follows that the IPQ is a 25 samples/s discrete-time signal. This signal acts as an input to the cognitive emulator for further processing. The main task of the cognitive emulator is to simulate the process of judgement and decision making of the human evaluators. More details about these stages will be discussed in the subsequent sections.

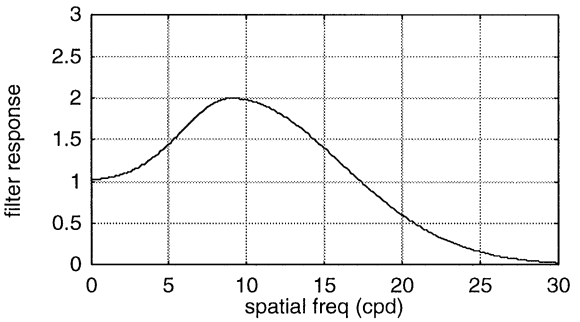


Fig. 6. Low-pass filter response.

3.1. Distortion weighting

The distortion weighting stage estimates the perceptibility of the error in the decoded frame using the low-level human visual system model. Fig. 5 shows the interior of this stage.

Both the reference and decoded pictures are first low-pass filtered. The low-pass filter has the response given in Fig. 6. After the filtering, the absolute error between the reference and decoded pictures is computed:

$$\varepsilon(x, y) = |Y_r(x, y) - Y_c(x, y)| \tag{3.1}$$

where $Y_r(x, y)$ and $Y_c(x, y)$ are the low-pass filtered luminance level of the reference and decoded pictures, respectively. Due to spatial masking effect, error occurring at sharp luminance transitions are less perceptible. This could be countered for by masking the error signal $\varepsilon(x, y)$ using the context information from the reference picture. A mask is constructed by first detecting the horizontal and vertical sharp luminance transitions in the reference picture using 3×3 Sobel filters all over the

reference picture:

$$e_h(x, y) = \text{sobel}_h(x, y),$$

$$e_v(x, y) = \text{sobel}_v(x, y), \tag{3.2}$$

where $e_h(x, y)$ and $e_v(x, y)$ are the horizontal and vertical luminance gradients at point (x, y) in the reference picture, respectively. These orthogonal gradients at each pixel are then combined to give a single gradient image, $e_r(x, y)$:

$$e_r(x, y) = \sqrt{e_h(x, y)^2 + e_v(x, y)^2}. \tag{3.3}$$

Note that $e_r(x, y)$ should be clipped to 255. After acquiring the luminance gradient information of each pixel in the picture, the mask could then be constructed as follows:

$$m(x,y,\delta) = \begin{cases} \left[255 - e_r(x, y) \left| \dfrac{5 - \delta}{5} \right| \right] \\ \quad \text{for } e_r(x, y) \geqslant 100, \\ 255 \quad \text{else}, \end{cases} \tag{3.4}$$

where $m(x, y,\delta)$ is the local spatial-masking function at point (x, y), and δ is the distance (in pixels) from the luminance transition. There is no masking effect if the luminance gradient is below 100. This masking function is a simple simulation of the spatial-masking effect. The reader is referred to [16] for more accurate masking function. Fig. 7. shows the 'cross section' of the masking function. The masking effect is maximum (smallest $m(x, y,\delta)$) at the point (x, y), i.e. $\delta = 0$, where the sharp luminance transition occurs, and gradually decreases as it gets farther from the sharp edge. According to [16], the masking typically spreads across 5 arc of degree, which is equivalent to 5 pixels when viewing the picture at a distance of six times the height of the broadcast picture. Therefore. the effective masking region is made ± 5 pixels from the centre of the luminance transition (hence $\delta = -5, -4, \ldots, 4, 5$).

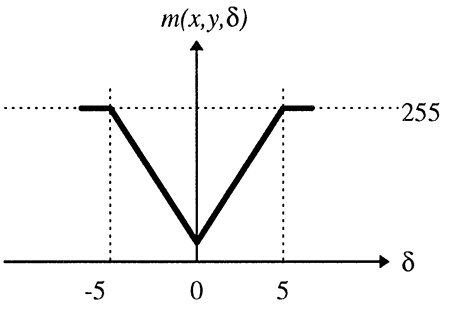


Fig. 7. Local spatial masking function, $m(x,y,\delta)$.

For each pixel in the gradient image, the corresponding local masking function is computed. After the masking function of each pixel is obtained, a global mask $m'(x, y\}$ which is the combination of the local masking functions $m(x, y, \delta)$ for all (x, y), is constructed. Where there is an overlapping of local masking functions (e.g. adjacent pixels in gradient image having large luminance gradients), the resultant masking equals to the stronger masker. This is best explained by Eq. (3.5), where $\text{MIN}\{\cdot\}$ returns the lowest value of the elements enclosed by the brackets.

$$m'(x, y) = \text{MIN}\{m(x + k, y - k), \\ m(x, y, 0), m(x, y + k, -k)\},$$

$$\text{where } k = -5, -4, \ldots, -1,1,2, \ldots 5. \tag{3.5}$$

Superimposing the global masking function onto the error signal $\varepsilon(x,y)$, we yield the masked error signal:

$$\varepsilon'(x, y) = \frac{m'(x, y) \times \varepsilon(x, y)}{255}, \tag{3.6}$$

in which we divide $m'(x, y)$ by 255 to normalise the masking function to 1.0 before applying it onto the error signal. Finally, the peak-to-peak signal-to-noise ratio is computed over the whole picture area, of width w and height h,

$$\text{PSNR} = 10 \times \log \frac{255^2}{(\sum_w \sum_h \varepsilon'(x, y)^2/(w \times h))}. \tag{3.7}$$

Since the final aim is to map the PSNR to visual rating, where the former is in dB and the latter is a mean opinion score, the PSNR had to be normalised for this purpose. The output of the distortion weighting stage is converted into another metric, the instrumental picture quality (IPQ), derived from the PSNR by its proper normalisation as below:

$$\text{IPQ} = \frac{\text{PSNR} - 20}{50 - 20}. \tag{3.8}$$

The normalisation boundaries were set to 20 and 50 dB, as we found these values gave the best fit to

the subjective results. IPQ is also clipped within a range of 0.0 to 1.0, to be mapped to the mean opinion score of SSCQE.

3.2. Cognitive emulator

The ultimate goal of the video distortion meter is to predict the results of subjective quality evaluation on video sequences that would be obtained from the SSCQE tests. The decision making and judgement process involved in SSCQE tests are very much different from those involved in quality evaluation of still images. Decision making tasks have four essential components: receiving of information (stimuli) from the external environment by the decision-maker; assimilation of the information in relation to some working hypothesis; action activated; and finally, making a response. In other words, decision-making is a very cognitive-demanding task, and very often necessitates selective processing of the input stimuli. Consequently, the decision, and hence response could be biased. In evaluating video quality, the variation of picture quality could be very frequent, demanding very rapid decision making from the human evaluator. As a result, biased judgement could be expected.

We try to identify and understand these biases, and if possible, model them with mathematical equations. We group these models into a block which we call the cognitive emulator. Fig. 8 shows the cognitive emulator we are proposing. It contains four elements: smoothing, perceptual saturation, asymmetric tracking and delay. These functions are explained in the following sections.

3.2.1. Smoothing effect

At a frame rate of 25 frames/s, the variation of picture quality in a video sequence is too fast for the human observer to segregate the individual video frame distortions. Depending on the bit-rate and picture activity (texture for spatial and motion for temporal), each frame will have different degree of impairments. For each variation in picture quality, a stimulus is sent to the human observer, and an associated response is generated (see the first case in Fig. 9(a)). Due to short-term human memory, the influence of a strong stimulus persists for a short

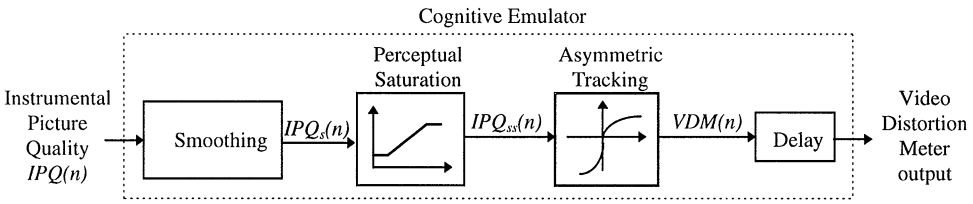


Fig. 8. Block diagram of the cognitive emulator.

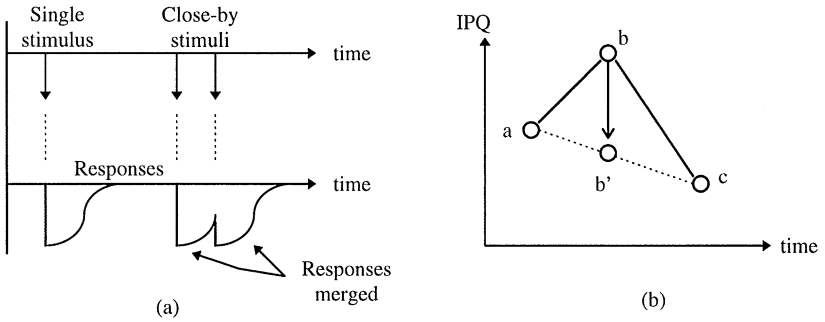


Fig. 9. Merging of responses results in the smoothing effect.

while, and fades out gradually. When two stimuli occur within an interval shorter than the memory duration, responses to these two stimuli may merge, as depicted in the second case in Fig. 9(a).

In terms of picture quality evaluation, when short durations of unimpaired frames interleave with distorted ones, all frames appear distorted. This implies some smoothing effect. It is believed, however, that the merging effect is limited to masking of undistorted frames by impaired ones, but not vice versa. Further validation on this phenomenon is necessary. The impact of this phenomenon on the picture quality judgement process is the quality metric (IPQ) from the preceding distortion weighting stage goes through some sort of smoothing. The implementation of this smoothing process is demonstrated in Fig. 9(b). Consider three consecutive IPQ data a, b and c, corresponding to three frames A, B and C, respectively. Since frames A and C are more impaired than frame B, the relatively better quality frame B is masked. Effectively the IPQ for frame B is modified from b to b′. Mathematically, this means

$$\mathrm{IPQ_s}(n) = \begin{cases} \frac{1}{2}(\mathrm{IPQ}(n-1) + \mathrm{IPQ}(n+1)), \\ \quad \text{if } [\mathrm{IPQ}(n) > \mathrm{IPQ}(n-1) \\ \qquad \text{and } \mathrm{IPQ}(n) > \mathrm{IPQ}(n+1)], \\ \mathrm{IPQ}(n), \quad \text{else.} \end{cases} \tag{3.9}$$

Taking the average of two adjacent points is the simplest approach to simulate the smoothing process. Iterative averaging may be employed to reflect a longer masking duration, which at this stage is still unclear.

3.2.2. Perceptual saturation

In many distortion meters, the dynamic range of the picture distortion (and fidelity) that humans can observe is not taken into consideration. A common assumption is that the picture quality perceived by the viewers is directly proportional to that estimated by the distortion meter (i.e., proportional to the IPQ) . However, our SSCQE test results suggest that this relation does not hold true if the picture quality goes to the extreme limits, either severely distorted, or with very high fidelity. In other words there are limitations in viewers' ability to observe any further changes in the picture quality after it exceeds certain thresholds, either towards better or worse quality. It is therefore anticipated that the perceived distortion can be related to the IPQ with a straight-line function in the middle region, but with non-linear characteristics at the boundary regions, as depicted in Fig. 10(a).

Fig. 10(b) shows the straight-line approximation we use in our model. This transformation could be integrated with the normalisation process performed at the distortion weighing stage, thus reducing the number of parameters. However, to make the cognitive emulator easier to be cascaded to other models besides the distortion weighting, the normalisation and transformation are done separately. The signal after the perceptual saturation process is therefore

$$\mathrm{IPQ_{ss}}(n) = \mathrm{PS}(\mathrm{IPQ_s}(n)), \tag{3.10}$$

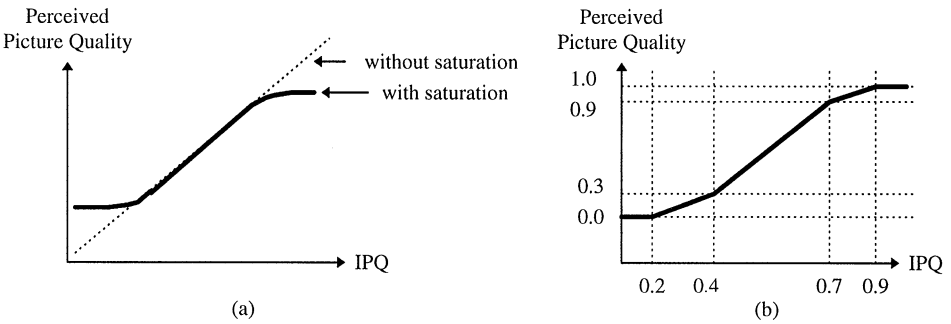


Fig. 10. Illustration of perceptual saturation.

where PS( ) is the transfer function emulating the perceptual saturation. We use straight-line approximation in our model for simplicity. All the parameters were chosen empirically, hence validation of this transfer function using DSCQS is necessary in the future.

3.2.3. Asymmetric tracking

In general, humans are better able to remember unpleasant experiences than pleasant moments, and also experience greater intensity of feelings from disliked situations compared to favourable situations. This could be described by a value function proposed by Kahneman and Tversky [12], graphically presented in Fig. 11(a).

This function illustrates the way in which people experience the displeasure of a loss more intensely than the pleasure of an objectively equivalent gain. Interpreting this behaviour in terms of subjective image quality assessment means that viewers are more sensitive to degradation than to improvement in picture quality. This phenomenon results in an asymmetric tracking ability of observers in trailing the variation of picture quality during the assessment process of video quality. The viewers respond decisively (and hence quickly) to degradation in picture quality, but hesitate (and thus slowly) in the cases of picture improvement.

In our model, we equate the gain to improvement in picture quality and the loss to drop in picture quality. Pleasure, in our case, is the subjective picture quality gain, and displeasure is translated into losses in subjective picture quality. We relate the instrumental gains/losses to the subjective gains/losses by Eq. (3.11):

$$g_s = \begin{cases} \alpha[1-(1-g_i)^{1.5}] & \text{for } g_i \geqslant 0, \\ -\beta[1-(1+g_i)^{1.5}] & \text{for } g_i < 0, \end{cases} \tag{3.11}$$

where g_s is the subjective gain, g_i is the instrumental gain, and α and β are the parameters controlling the degree of asymmetry between gain and loss. Note that $\alpha < \beta$, gives the value function shown in Fig. 12(b). Using this transfer function, we model how human evaluators respond to variation in picture quality. Let us assume that the current output of the distortion meter (which is also the present position of the slider) is VDM$(n-1)$, as illustrated in Fig. 12. During the SSCQE test, subjects try to track the temporally varying picture quality. At $t = n\tau$, where τ is the interval between $\mathrm{IPQ_{ss}}$ data samples (0.04 s for 26 frame/s sequences), the picture quality perceived by the subjects is $\mathrm{IPQ_{ss}}(n)$. Trying to track the picture quality, the subjects estimate the error between the current slider position and the previous picture quality to be

$$g_i = \mathrm{IPQ_{ss}}(n) - \mathrm{VDM}(n-1). \tag{3.12}$$

Due to asymmetric tracking capability, the picture quality variation perceived by the subject is modified by the value function, and consequently the error g_i becomes g_s, according to Eq. (3.11).

The subject then attempts to reposition the slider to compensate for this quality change. To simulate the process of adjusting the slider mechanism, we introduce another transfer function given in

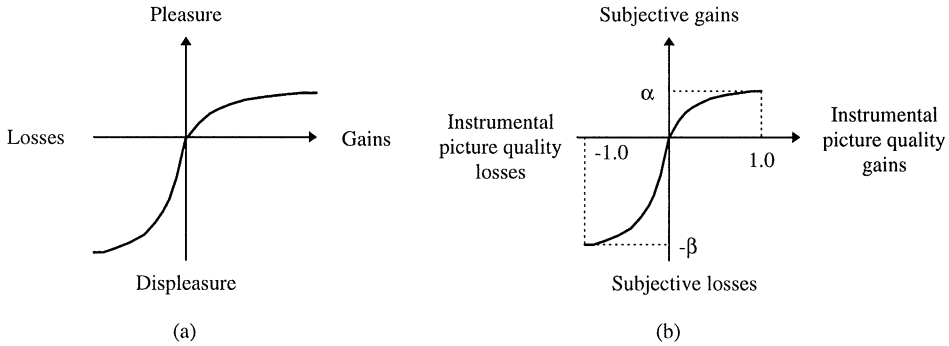


Fig. 11. The asymmetric nature of Kaheman and Tversky's value function.

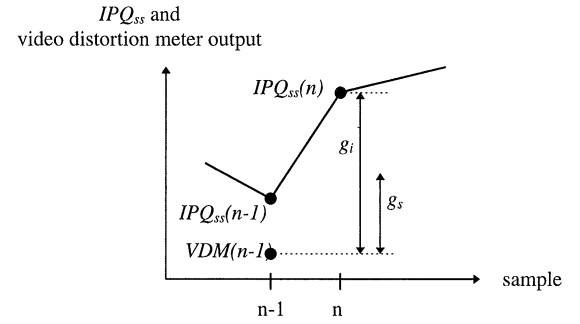


Fig. 12. Effect of value function on judgement of quality change.

Eq. (3.13). $S(g_s)$ represents the amount of slider movement due to subjective gain g_s. Since the slider's moving speed is limited by the friction, the change in the slider position is scaled down by the factor λ, where λ is normally $\ll 1.0$. To avoid having too many adjustable parameters in this video distortion meter, the parameter μ, which controls the sensitivity of the slider, is set to 1.0, hence $S(g_s)$ is simplified to Eq. (3.14).

$$S(g_s) = \begin{cases} \lambda(1-(1-g_s)^{\mu}) & \text{for } g_s \geqslant 0, \\ -\lambda(1-(1+g_i)^{\mu}) & \text{for } g_s < 0, \end{cases} \tag{3.13}$$

$$S(g_s) = \lambda \times g_s, \tag{3.14}$$

$S(g_s)$ is indeed the displacement of the slider. Therefore, the slider position is updated according to this distance, and the new position of the slider (video distortion meter output) is given by

$$\mathrm{VDM}(n) = \mathrm{VDM}(n-1) + S(g_s). \tag{3.15}$$

The process of computing the weighted noise, smoothing and non-linear transformations (perceptual saturation and asymmetric tracking) is repeated for each pair of reference and coded frames, and at the output of the asymmetric tracking stage, we obtain a 25 samples/s discrete-time signal reporting the variation of picture quality that would be perceived by the human observers.

3.2.4. Response time

The output of the asymmetric tracking stage is synchronised to the input frames, but it is not the case in SSCQE. The human observers make decisions responding to every slight variation of picture quality and displace the slider to reflect their opinion. As mentioned earlier, decision-making is a very demanding cognitive process, taking finite time to yield the response. The consequence is a delay between the moment the stimulus is captured by the subject and the moment the slider is brought to its right position. We need to delay the objective measurement result by the same amount in order to have the objective and subjective results synchronised. Unfortunately, this delay is not constant, depending on many factors to be identified. However, de Ridder and Hamberg [9] estimate the human response time to variation in picture quality to be about 1 s. This includes both the time needed to respond to stimuli as well as the delay due to the finite movement speed of the slider mechanism used in SSCQE. Therefore, the asymmetric tracking stage output is delayed by the same amount to provide a very crude temporal alignment between video distortion meter output and the SSCQE data.

4. Experimental set-up

This two-stage video distortion meter has been evaluated using three broadcast standard (720×576 pixels) MPEG-2 MP@ML coded video sequences, containing short scenes (Playground, Wind Machine and Photocopier) from a feature programme "Exam Conditions". The quality of the digitally coded video was controlled by varying the encoding bit-rate, according to the setting shown in Fig. 13. The first sequence, "Playground", was coded at 7.5 Mbits/s for first 35 s, followed by another 44 s coded at 4 Mbits/s. The bit rate is further reduced to 2 Mbits/s during the 79–128 s, and finally very low bit-rate (1 Mbits/s) is used for the last 52 s. The second sequence, "Wind Machine", is divided almost equally into three sections. The beginning and ending sections were coded at 4 Mbits/s, while the middle section used 2 Mbits/s. The last sequence, "Photocopier", is also divided into three sections, using 7.5, 1 and 4 Mbits/s for each section, respectively. These three sequences had earlier been used in a subjective test involving a panel of 15 subjects using the SSCQE method. An average result was obtained from these subjects to produce the SSCQE curves, and these were

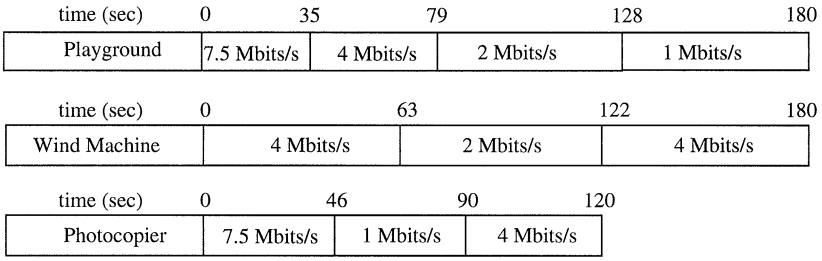


Fig. 13. Coding bit-rates of test sequences.

subsequently used to assess the performance of the video distortion meter.

For the cognitive emulation stage, parameters have been set empirically. Assuming that the displeasure due to losses is double the intensity of pleasure due to gains, we set α to 0.5 and β to 1.0. The slider sensitivity, λ is set to 0.03, which implies moving speed of 75% of the whole scale per sec.

We have also investigated the performance of our model without weighting the coding distortions with the low-level vision of the first stage. We fed the normalised PSNR computed directly from MSE data to the cognitive emulator and compared its output with the result obtained from using IPQ as the input to the second stage. This comparison demonstrates the role and significance of the cognitive emulator in the distortion meter. The normalisation of the PSNR in this case uses different boundaries: 12 and 42 dB. This is, in fact, derived from the average shift of 8 dB between the PSNRs computed from unweighted (MSE only) and weighted (distortion weighting stage) noise, respectively.

5. Results and discussions

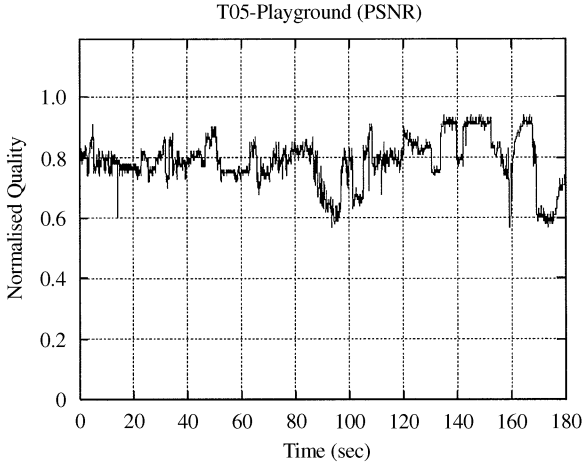
Fig. 14(a–d) shows the PSNR, IPQ, video distortion meter output with noise weighting, and video distortion meter output without noise weighting, together with associated SSCQE results in Fig. 14(e). Conventional analysis might suggest the computation of mean-square error between the SSCQE and the other approximations to it as a performance indicator. The problem with this approach is the variable delay between as subject seeing the distortion and moving the lever which records his or her SSCQE response. In the absence of a reliable metric, we simply present the graphs here for visual inspection and interpretation.

Comparing the five charts, it is obvious that the video distortion meter output with noise weighting provides the closest approximation to the SSCQE graph. As expected, PSNR does not correlate well with SSCQE. The IPQ (Fig. 14(b)), output of the distortion weighting stage, shows some slight improvement compared to PSNR, but yet the correlation with SSCQE is unsatisfactory. By passing the IPQ through the cognitive emulator for further processing, the data have been transformed to closely follow the SSCQE result, as illustrated in Fig. 14(c).

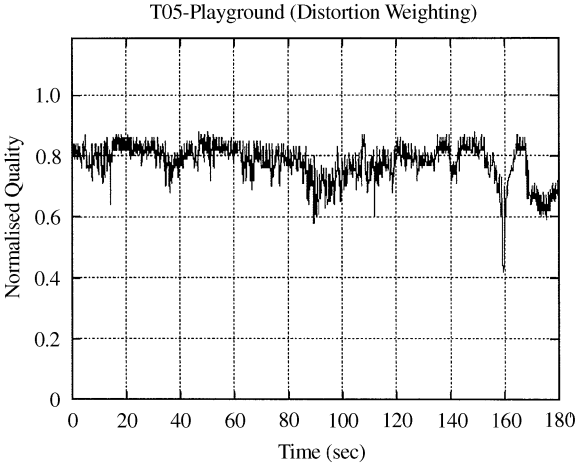
Using PSNR computed from unweighted noise as the input to the cognitive emulator produces the objective picture quality curve given in Fig. 14(d). Although it does not follow the SSCQE curve as well as in Fig. 14(c) (in which weighted noise was used to compute the PSNR), its tracking is, nevertheless, much better than PSNR (Fig. 14(a)) alone. This demonstrates the effectiveness of the cognitive emulator in transforming raw difference into a signal that approximates SSCQE.

Similar observations can be made from Fig. 15. Comparing Fig. 15(a,b,e) we can see the advantage of having the distortion weighting stage. Despite of the high-frequency variation, the outline of the graph in Fig. 15(b) resembles the SSCQE result in Fig. 15(e) much better. After the cognitive emulation, the video distortion meter tracks the SSCQE data even more accurately.

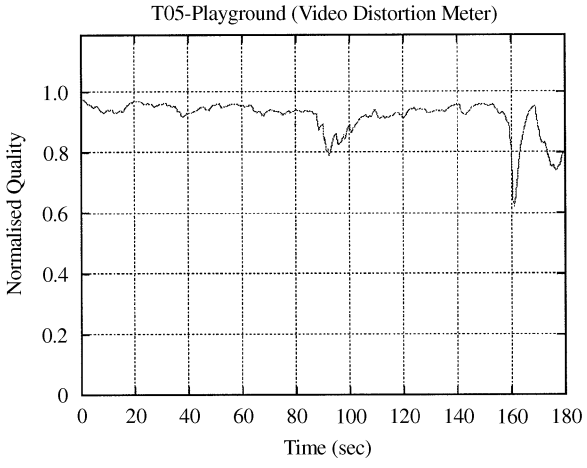
Fig. 16(a–e) shows the results from the most stringent test sequence. The picture quality varies



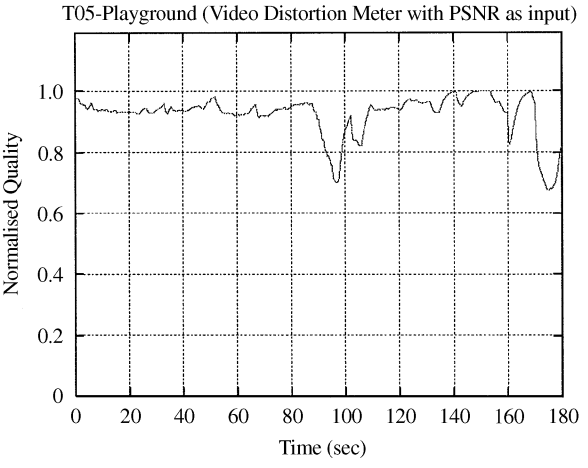
(a) PSNR



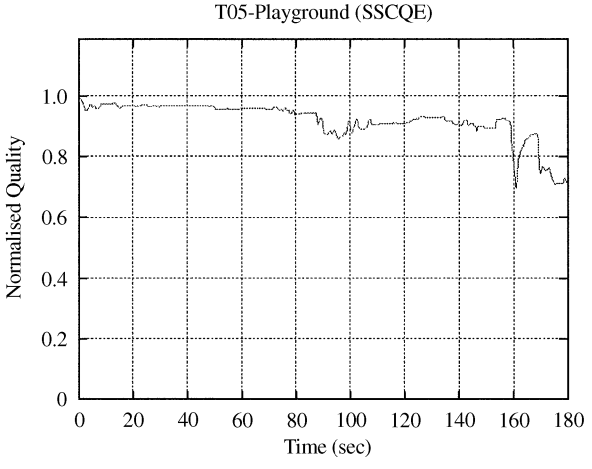
(b) Distortion Weighting Only



(c) Video Distortion Meter

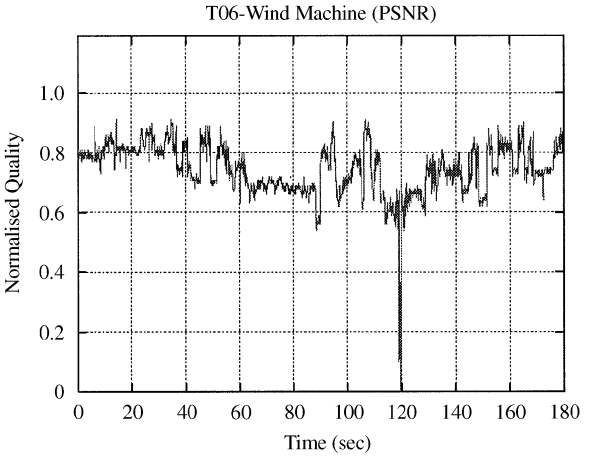


(d) Video Distortion Meter with PSNR as input to Cognitive Emulator

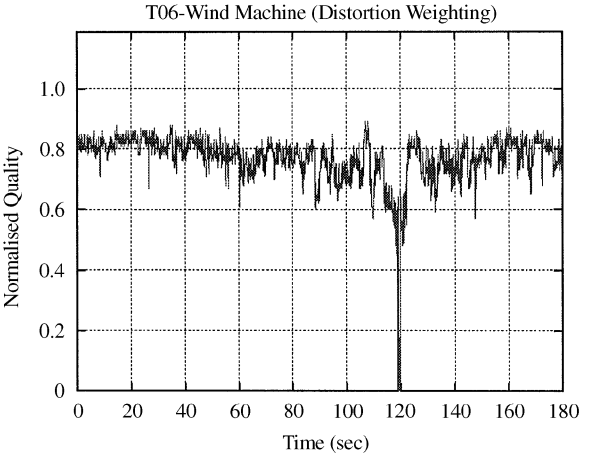


(e) SSCQE

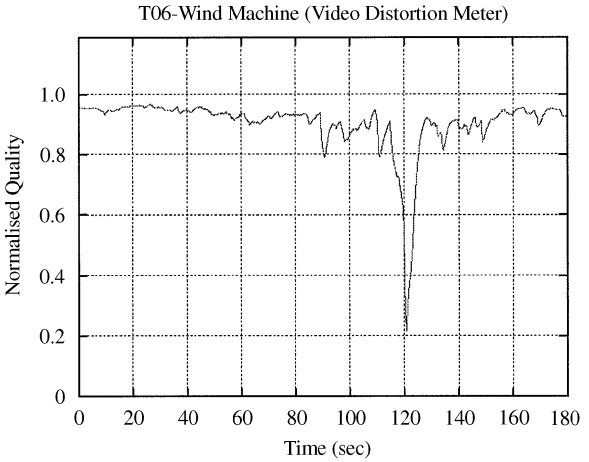
Fig. 14. Results for test sequence “Playground”.



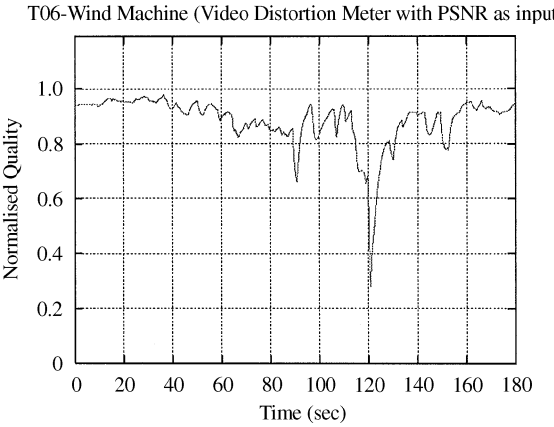
(a) PSNR



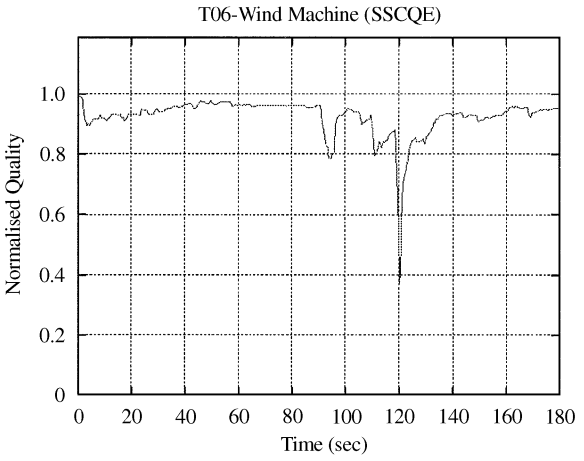
(b) Distortion Weighting Only



(c) Video Distortion Meter



(d) Video Distortion Meter with PSNR as input to Cognitive Emulator



(e) SSCQE

Fig. 15. Results for test sequence “Wind Machine”.

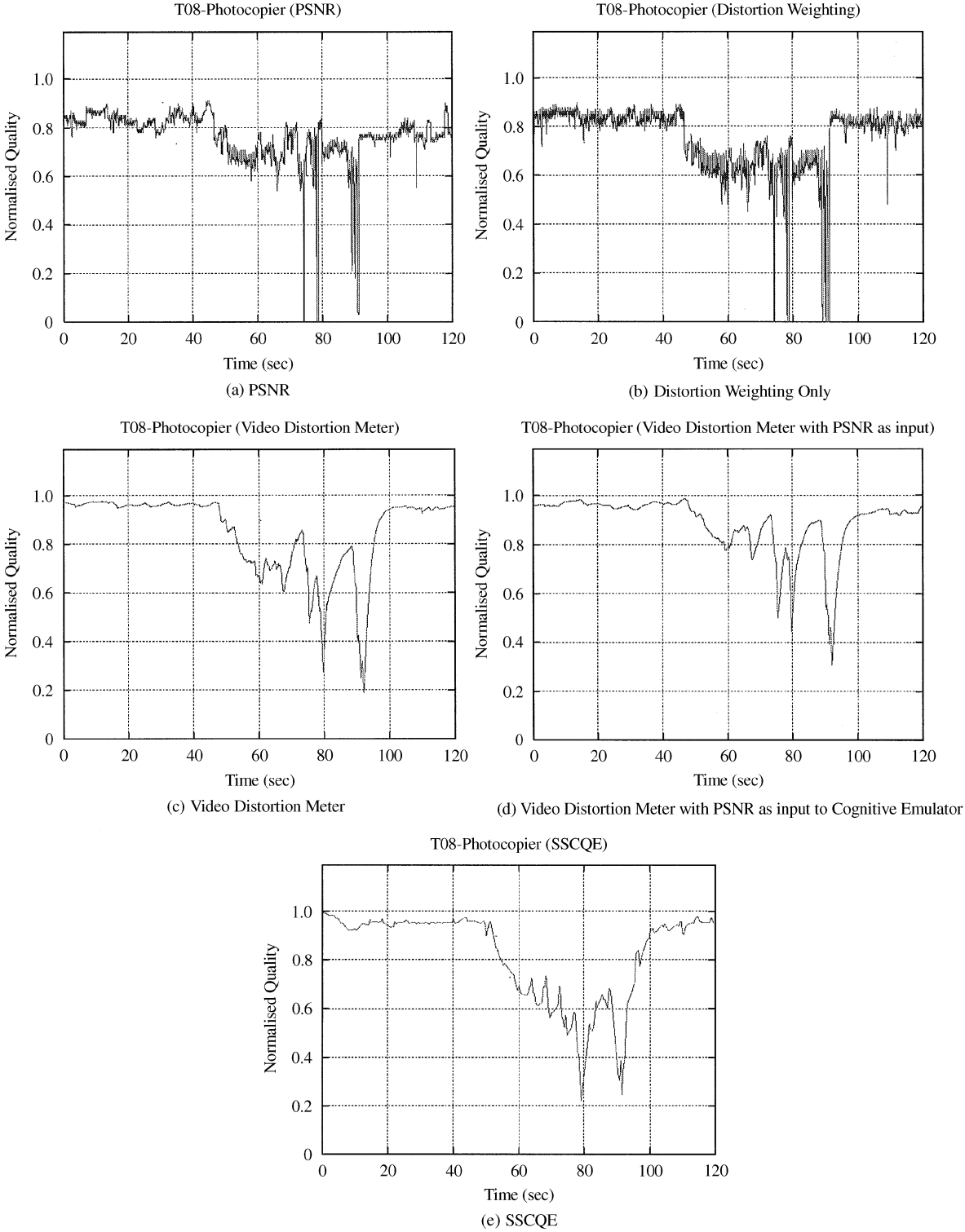


Fig. 16. Results for test sequence “Photocopier”.

violently, due to the low coding bit-rate (1 Mbits/s) during the 46th to the 90th sec.

Cascaded to the cognitive emulator, both the PSNR and distortion weighting stage produce an output trace closely following the SSCQE curve for most of the 120 s period. The disagreement between the SSCQE data and the video distortion meter during the 65–75 s could be due to the rapid picture quality fluctuation, together with frequent scene cuts, giving the subjects great difficulty in synchronising the slider movement with the picture quality variation. It could be also due to the domination of tiling effect in the coded pictures as a result of the low bit-rate and violent motions, causing the distortion weighting stage to underestimate the subjective picture quality. Both the possibilities remain to be investigated, prompting the area for future improvement.

6. Conclusions

We have presented an objective measurement model suitable for assessing the temporal quality variations in long MPEG-2 video sequences. The model has been tested using a variety of coded video sequences of 2 and 3 min duration, and the performance of the complete two-stage meter was compared with the PSNR and SSCQE graphs. It is clear that the meter provides a closer approximation to SSCQE than either weighted or unweighted PSNR.

These results highlight the importance of having a cognitive emulation stage to simulate the decision-making process of humans during subjective assessment, an aspect which has been neglected in many other models. Although the parameter settings are not yet optimum pending further experimental investigations, the second stage of the model does illustrate the potential to enhance the performance of low-level models which give a single rating for overall video quality. Our method provides important details of the picture quality variations, hence making it a more suitable tool for assessing moving picture quality.

Acknowledgements

The authors acknowledge with gratitude the support of the UK Independent Television Commission for this work, and the contributions of European partners to the TAPESTRIES project reported in this paper.

References

[1] R. Aldridge, Continuous quality assessment of digitally-coded television pictures, Ph.D. Thesis, University of Essex, UK, 1997.

[2] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, D.E. Pearson, Recency effect in the subjective assessment of digitally-coded television pictures, in: 5th Internat. Conf. on Image Processing and its Applications, 4–6 July 1995, Edinburgh, UK, No. 410, pp. 336–339.

[3] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, D.E. Pearson, Subjective assessment of time-varying coding distortions, in: PCS '96, Melbourne, Australia, 13–15 March 1996, pp. 269–274.

[4] J.W Allnatt, Transmitted-Picture Assessment, Wiley, London, 1983.

[5] J.W. Allnatt, J.M. Corbett, Adaptation in observers during television quality-grating tests, Ergonomics 15 (1972) 353–356.

[6] L. Boch, S. Fragola, Lancini, M. Visca, Extracting spatial and motion information from video sequences to correlate objective measures and human observer scores, in: PCS '97, Berlin, Germany, 10–12 September 1997, pp. 183–187.

[7] J.C. van den Branden Lambrecht, A working spatio-temporal model of the human visual system for image restoration and quality assessment applications, in: Proc. ICASSP, Atlanta, GA, 7–10 May 1996.

[8] P.N. Gardiner, M. Ghanbari, D.E. Pearson, K.T. Tan, Development of a perceptual distortion meter for digital video, in: IBC, 1997, pp. 493–497.

[9] R. Hamberg, H. de Ridder, Continuous assesment of perceptual image quality, J. Opt. Soc. Amer. 12 (12) (December 1995) 2573–2577

[10] D. Hands, Mental processes in the evaluation of digitally-coded television pictures, Ph.D. Thesis, Dept. of Psychology, University of Essex, UK, 1998.

[11] Y. Horita, M. Katayama, T. Murai, M. Miyahara, Objective picture quality scale for video coding, in: ICIP-96, 16–19 September 1996, Lausanne, Switzerland, Vol. 3, pp. 319–322.

[12] D. Kahneman, A. Tversky, Prospect theory: An analysis of decisions under risk, Econometrica 47 (1979) 263–291.

[13] N.K. Lodge, Interpolative coding methods for the digital transmission of conventional and high definition television, Ph.D. Thesis, Heriot-Watt University, Edinburgh.

[14] J. Lubin, A visual discrimination mode for image system design and evaluation , in: E. Peli (Ed.), Visual Models for

Target Detection and Recognition, World Scientific Publishers, Singapore, 1995.
[15] N. Narita, Subjective-evaluation method for quality of coded images, IEEE Trans. Broadcasting 40 (1) (March 1994) 7–13.
[16] A.N Netravali, B.G Haskell, Digital Pictures, Representation and Compression, Plenum Press, New York, 1988.
[17] J. Okamoto, S. Hangai, K. Miyauchi, A study on subjective and objective evaluation method for coded moving picture quality, in: PCS, Melbourne, 3–15 March 1996, pp. 519–523.
[18] Recommendation ITU-R BT.500-7 (Revised), 1996. Methodology for the subjective assessment of the quality of television pictures.
[19] K.T. Tan, M. Ghanbari, D.E. Pearson, A video distortion meter, in: PCS '97, Berlin, Germany, 10–12 September 1997, pp. 119–122.