RECOMMENDATION ITU-R BT.813*

Methods for objective picture quality assessment in relation to impairments from digital coding of television signals

(Question ITU-R 44/6)

(1992)

The ITU Radiocommunication Assembly,

considering

a) that with the increasing application of digital coding and bit-rate reduced transmission, the assessment of coding impairments is of critical importance;

b) that objective measurements are necessary for specific evaluations as routine supervisions or system optimizations;

c) that several types of objective measurements may be defined for different digital systems and applications;

d) that the adoption of standardized methods is of importance in the exchange of information between various laboratories,

recommends

1 that the general methods described in Annex 1 for obtaining objective measurements of digital picture quality should be used;

2 that the choice of the appropriate objective measurements for a system or an application should be made on the basis of the information given in Annex 1.

ANNEX 1

1 Introduction

With the increasing application of digital coding and bit-rate reduced transmission, the assessment of coding impairments is of critical importance. An understanding of these assessment methods is relevant not only to the performance of new coding equipment, but also to an interpretation of measurements made on such equipments and to specifications for target performance. Moreover digital codecs as with all adaptive or non-linear digital processes, cannot be fully characterized with traditional television test signals or patterns.

Quality of codecs for distribution application can be measured objectively, quality specifications being expressed against the subjective judgement of observers.

---

* Radiocommunication Study Group 6 made editorial amendments to this Recommendation in 2002 in accordance with Resolution ITU-R 44.

Studies indicate the desirability of establishing relationships between objective measurements of signals impaired by digital coding, and the visual quality of the picture thus obtained. This Annex gives progress towards this end, which is proving more difficult to achieve as codec complexity increases.

The quality of a codec designed for contribution applications however, could in theory be specified in terms of objective performance parameters because its output is destined not for immediate viewing, but for studio post-processing, storing and/or coding for further transmission. Because of the difficulty of defining this performance for a variety of post-processing operations, the approach preferred has been to specify the performance of a chain of equipment, including a post-processing function, which is thought to be representative of a practical contribution application. This chain might typically consist of a codec, followed by a studio post-processing function (or another codec in the case of basic contribution quality assessment), followed by yet another codec before the signal reaches the observer. Adoption of this strategy for the specification of codecs for contribution applications means that the measurement procedures given in this Annex can also be used to assess them.

2 Digital codec classification

The function of digital coding is to reduce the bit rate needed to represent a sequence of images while ensuring minimal loss in picture quality. Coding equipment does this, first by removing as much statistical redundancy from the images as possible (i.e. no loss in quality occurs as a result of this conceptual first stage). Then, if more bit-rate reduction is necessary, some distortion has to be introduced into the picture, although one of the objectives of codec design is to hide this distortion by exploiting certain perceptual insensitivities of the human visual system.

It is convenient to divide codecs into two classes, those using fixed word-length coding and those using variable word-length coding (see definitions in § 3.1 and 3.2 respectively). The latter class is more efficient and complex, and includes all recently proposed systems for coding 4:2:2 video to the range 30-45 Mbit/s. The former class is however sufficient to permit 4:2:2 video to be reduced to 140 Mbit/s while still preserving the quality demanded for contribution applications. A further sub-division of these classes is also useful, into intrafield (or spatial) codecs and interframe (including interfield) codecs, which contain frame (or field) stores permitting them to exploit the redundancy which exists between successive picture frames (or fields).

There is emerging a third class of codec which employs variable word-length coding but which is being designed for variable bit-rate networks. These codecs can in principle, preserve a constant decoded image quality subject to the bounds of peak network demand. The quality-testing of such codecs would have to take into account the nature of the network used and the statistics of the data injected by all of its users, and remains to be studied.

3 Objective assessments of codecs in terms of perceived picture impairments

3.1 Fixed word-length codecs

With fixed word-length codecs a fixed number of bits is used to represent a fixed number of source picture samples. For example in fixed word-length PCM or DPCM codecs, a fixed number of bits is

allocated to each picture sample, and in fixed word-length transform or vector quantization codecs, a fixed number of bits is allocated to each block of picture samples.

3.1.1 Methods based on the use of synthetic test signals

In these codecs the impairment introduced into each received picture sample of an image is dependent upon the values of those samples in the locality surrounding it, either in the same field (for an intrafield codec) or in the same and previous fields (for an interframe codec). It is therefore possible, using suitably chosen two or three dimensional digital test signals, to artificially provoke the degradations characteristic of digital image coding.

Some of these degradation factors have acquired names such as false contouring, granular noise, blur, blocking impairments, etc., relating to their interpretation by observers. Having provoked these distortions, their magnitudes can be objectively measured and, using experience gained from subjective assessments these measurements could then be related to some quantification of codec quality. Relating the degradation factors to their interpretation by observers may prove difficult in interframe coding systems or systems employing some adaptive processing because they can vary at any moment, with motion or adaptation of the coding algorithm. In a proposed method, the subjective assessment test first uses scales derived from pairs of opposite adjectives (the semantic differential method), and then the results are analysed by principal component analysis to extract the picture quality degradation factors. The classification results can be tested by applying multiple regression analysis which relates the factors to subjective judgements. A list of picture quality degradation factors is presented in Table 1.

TABLE 1

Examples of picture quality degradation factors for digital system and corresponding physical measures (units)

Picture quality degradation factor	Physical measure
Image blur	Step response rise time
Edge busyness	Step response jitter width
False contouring	$S_{p\text{-}p}$ to minimum quantizing *p-p*
Granular noise	Equivalent analogue signal-to-noise ratio expressed in terms of $S_{p\text{-}p} / N_{r.m.s.}$
“Dirty window” effect	Maximum noise amplitude
Movement blur	Rise time of a moving edge
Jerkiness	Field or frame difference in terms of moving edge position

While these methods appear to have conveniences for codec assessment and also to offer a tool to the codec designer, they are difficult to relate to the performance of a codec for real pictures for the following reasons:

– the complex composition of real picture sequences cannot be satisfactorily modelled by a practical number of synthetic test signals;

– degradations can be numerous in character and difficult to classify because of their subtle nature (for example, a particular distortion may be visible only in textured parts of an image moving in a particular way);

– meaningful objective measurements of degradations can be difficult to define (for example, for motion portrayal). It should be noted that the duration of the period in which objective measures are taken should correspond to the observation window provided by the duration of the presentation in subjective tests.

3.1.2 Methods based on natural picture material and coding error

Natural picture sequences can be thought of as being composed of a number of different regions, each with different local content and each exercising different fixed word-length codecs in different ways. Therefore the content of an image sequence will have a significant impact on the quality perceived by a viewer.

It is also possible, where a comparison is to be made between two codecs, for the image sequence content to determine which appears the better. Not only does this underline the importance of the choice of test images for subjective assessments (see Recommendation ITU-R BT.500) but also that an objective measure of the performance of a particular codec must consider image content, if there is to be a correlation between subjective and objective assessment results.

The most common forms of objective quality measurement are based on the coding error of a codec; that is, the difference between an input picture sequence and its decoded output. This difference signal (often amplified) can itself be displayed as an image sequence and this can provide a useful development aid to the codec specialist. It should not however be used as material for subjective assessments.

3.1.3 Methods based on normalized mean square error

A frequently used objective measure of decoded image quality is mean square coding error. This is the average, over every picture sample in a sequence, of the square of the coding error and is usually normalized with respect to (the square of) the full amplitude range of the picture samples. Sometimes the normalized mean square error (NMSE) is quoted as a coding noise figure evaluated as $-10 \log$ (NMSE). The popularity of the NMSE measure stems from its mathematical convenience but it must be regarded with caution as a measure of decoded quality. It cannot distinguish, for example, between a few large coding errors (which may be annoying to an observer) and a large number of small coding errors (which may be imperceptible). Weighting of the coding error signal (performed after a log operation) prior to the NMSE evaluation, with a filter derived from a visual model, has been attempted and has achieved improved correlation with subjective assessment results. The NMSE is a useful practical tool in codec development where it is often required to compare coding methods which are very similar (i.e. those which use minor variants of the same algorithm and where impairment processes can be assumed to be identical).

3.1.4 Methods based on visual models

The sensitivity of the human visual system to coding error in a particular region of an image is strongly influenced by the characteristics of the image material itself in that region. The inability to recognize this fact is the major failing of the NMSE measure. To give just one example of this influence: it is known that an observer's sensitivity to coding error noise is reduced when the

spectrum of that noise approximately coincides with the spectrum of the "background" image. These properties of the visual system are those which are being exploited in codec design when subjective experiments or psychovisual data are used to optimize system parameters.

In order to further the correlation between objective measures of picture quality and that judged by human observers it is necessary to develop a visual model which can interpret local coding error in the context of the background image and which can combine all these local assessments to form a global quality rating. This approach is applicable to both fixed and variable word-length codecs and is considered in § 3.2.3.

3.2 Variable word-length codecs

Television codecs which require to reduce their source image data by more than a factor of about two, use methods based upon variable word-length coding. These codecs have increased efficiency because they possess the flexibility to allocate dynamically coding bits to the parts of an image sequence where they are most effective in maintaining decoded image quality. There are several ways in which codecs can do this; the use of variable length entropy codes is not necessarily implied.

3.2.1 Methods based on the use of synthetic test signals

Because of the flexibility of these codecs, the impairment which they introduce into each coded sample is dependent not only on the values of samples in the same locality, but also on the history of previous samples extending a frame or more into the past. This means that for either intrafield or interframe variable word-length codecs it is not meaningful to attempt codec characterization by trying to provoke local distortions with local test signals and making objective measurements on them. If, however, the adaptation modes of a variable word-length codec can be artificially held (requiring access to its internal workings), each mode may be characterized separately. Knowledge of the codec's adaptation switching, when it is presented with natural scenes, could then be used to objectively determine its performance.

It is possible to contrive moving synthetic test sequences which take a codec to the point where it produces visible distortion, but even if objective measurements could be defined to characterize these distortions (see reservations in § 3.1.1), their interpretation could only be made in the context of that entire test sequence. This raises questions about how typical of natural scenes it is, and whether a codec designer would have the opportunity to optimize its performance to suit known test material.

3.2.2 Methods based on natural picture material and coding error

It is important in any assessment of variable word-length codecs that natural picture sequences be used. Bearing in mind the ability of these codecs to direct the utilization of coding bits throughout the image, careful consideration should be given to the content of every part of the image sequence when judging its criticality (see Recommendation ITU-R BT.500). It is recommended that any objective assessments be based on the coding error of a codec where the inputs are a number of natural test pictures. The normalized mean square error method discussed in § 3.1.3 may also be applied to the coding error from variable word-length codecs but such results should be for specialist interpretation only and even then, only as a supplement to subjective assessments. Similarly objective comparisons between codecs based on the NMSE should only be undertaken by

specialists in codec design and only where techniques to be compared have very minor differences (i.e. are variants of the same algorithm) and where impairment processes can be assumed to be identical.

3.2.3 Methods based on visual models

The major disadvantage of measures based upon the NMSE is that they do not recognize the strong influence which the image content itself has on the sensitivity of an observer to impairments. As was mentioned in § 3.1.4, codec design optimization involves the use of subjective experiments and psychovisual data to match it to the distortion-tolerance of the human observer and to the characteristics of local image regions. This ensures that when a variable word-length codec apportions coding bit-capacity (and therefore also apportions the magnitudes of coding errors) throughout an image it can do so in a manner which is also matched to visual characteristics. Any objective assessment method must therefore encompass properties of the human visual system if it is to yield results which correlate well with subjectively-determined quality ratings. It is the function of a visual model to interpret coding error in the context of the source image in which it occurs.

The assumption in the following text assumes that access to the internal workings of a codec is not available. If information on adaptation modes can be obtained, variable word-length codecs can also be assessed using the method of degradation factors (see § 3.1.1).

In the development of a visual model, two levels of knowledge must be incorporated. The first concerns how visible any arbitrary impairment is, given its location in the image, and the second determines how the visibility of all the impairments should be combined to yield an overall quality rating. It is, however, only necessary to concentrate on models which account for the impairments characteristic of digital coding methods; distortions of a geometric or semantic nature, for example, need not be considered. Models of the response of the human visual system to distortions arising from image transmissions have concentrated on phenomena at or near the threshold of visibility adequate for high quality television applications. Little is known about the modelling of the response to larger distortions.

A study detailing the design of a visual model for picture quality prediction has been made. This study examines the development of the model and its performance as a predictor of subjective quality, from a simple estimator based on raw error measures, through one which models (non-linear) visual filtering, to one which can account for the spatial and temporal masking properties of vision. As vehicles for this study, the distortion processes of uniform quantization, DPCM coding, additive Gaussian noise, and low-pass filtering were used. Particularly noteworthy in the derivation of an overall quality measure for an image sequence, was the modelling of the observation that viewers tend to grade pictures according to the level of distortion present in the most impaired locality of the image and not as an average over all the image. More recently, other visual models have been developed for application to digital picture coding.

The use of visual models for the objective determination of picture quality in the presence of not only digital coding impairments but also impairments arising from other non-linear or adaptive processes, is an area of great promise and more studies on this topic are required.

4 Objective assessment of codec picture quality in the presence of transmission errors

In a practical transmission environment, the link between coder and decoder will be subject to influences which can corrupt the data being conveyed, so an important characteristic of a decoder is its response to the presence of these transmission errors. In a carefully designed codec this response will be of the form of local transient distortions within the decoded image, where the number of these transients is related to the channel error statistics, and their nature is related to the picture coding algorithm employed and the criticality of the image sequence being displayed. Typically, the aim of assessments involving transmission errors is to derive, for a codec, a graphical representation of the impairment perceived by the viewer over a range of bit error ratios.

There are several levels of processing within a decoder which determine its response to transmission errors, some of which may be analysed mathematically (or simulated by computer), while others require either some degree of subjective assessment or an objective model of the viewer's response to transient distortions.

The first stage in an objective analysis is to describe as accurately as possible, the way in which errors occur in a practical link; this is usually expressed as a statistical model. In its simplest form such a model assumes that errors occur randomly and independently (Poisson distribution). However, it has long been known through practical observation that in reality errors appear in clusters or bursts. Several models have been proposed to account for this behaviour, the most popular being based on the Neyman type A distribution. Whereas the simple Poisson distribution is completely defined by a single parameter, the mean bit-error ratio, the Neyman A model requires a further two parameters to be quantified relating to the degree of clustering and the error density within each cluster. No Recommendation is yet available for realistic choices of these parameters.

Aware of the bursty nature of transmission errors, codec designers often incorporate a process of time-reordering of the transmitted bits before they enter the channel. This ensures that bursty channel error occurrences are spread by the inverse reordering mechanism in the decoder and are thus rendered in a form which is more amenable to processing by the subsequent error correction system. This error correcting system will be capable of completely correcting a number of errors using a redundant overhead of transmitted data capacity but there will remain some distribution of "residual" errors which will enter the picture decoding algorithm. The distribution of residual errors may be calculated for a particular codec and channel model but it remains to assess the effect that these errors will have on the decoded image.

ITU-R suggests that the performance of a particular codec in transmission errors be judged in two parts. First subjectively, in order to determine the impairment due to the distortion transient characteristic of that codec, and second objectively, taking into account the rate of residual errors obtained by computation from the above considerations. At present no experimental evidence is available to support this approach. It could, however, be the first step in a wholly objective measure, if the response of the viewer to different codec transients can be characterized. It is important to note that some transmitted bits are more sensitive to corruption than others, meaning that a codec's response to a single bit residual error can vary greatly and can also depend on the criticality of the source image sequence. In interframe codecs for example, the transient resulting from residual errors can remain in static parts of a picture sequence until provision is made to remove them by refreshing. Finally, a feature of some codecs employing variable word-length coding is that they can detect some violations of coding caused by transmission errors and use this knowledge to attempt to

conceal the distorting transients. While not successful for every error, this concealment process generally improves the subjective quality of the resulting image, a fact which must be accounted for in any objective codec assessment.

5 Distortions in mixed analogue and digital transmission

Until the present time, picture quality specification problems have been considered individually for analogue or digital systems. If the psychological independence of picture quality degradation phenomena mentioned in § 3.1 can be assumed, then the approach described in that section may also be applicable to mixed systems. That is, they can be classified into one of the following three groups from the viewpoint of psychological independence:

a) impairments caused only by the analogue section;

b) impairments caused only by the digital section;

c) impairments caused by both analogue and digital sections (which might be independent factors in each individual system).

Impairments belonging to group a) or b) will be dealt with as independent factors and a function has already been proposed to the ITU-R for estimating the overall picture quality in this case. This estimation function is applicable when certain mutually independent psychological factors exist at the same instance.

On the other hand, in group c), where picture quality degradation phenomena from both sections are so similar that they cannot be regarded as independent, it will be necessary to find a new way to allocate picture quality degradation to both the analogue and digital sections before applying the estimation equation mentioned above.

An example of the investigation results for such a case is reported in which a combination of random noise from the analogue system and granular noise from a fixed word-length intraframe DPCM coding system was investigated to show that it is possible to replace a physical measure in the analogue system with a corrected value based on visual sensitivity differences.