METHODS FOR PICTURE QUALITY ASSESSMENT IN RELATION TO IMPAIRMENTS FROM DIGITAL CODING OF TELEVISION SIGNALS

(Question 3/11 and Study Programme 3B/11)

(1990)

1. Introduction

With the increasing application of digital coding and bit-rate reduced transmission, the assessment of coding impairments are 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 characterised with traditional television test signals or patterns.

Studies in connection with Question 3/11 and Study Programme 3B/11 indicate the desirability of establishing relationships between objective measurements of signals impaired by digital coding, and the subjective assessments of the quality of the picture thus obtained. This Report gives progress towards this end, which is proving more difficult to achieve as codec complexity increases.

Subjective methods for conventional resolution television picture quality and impairment assessment are given in Recommendation 500 and for HDTV, are given in Recommendation 710. ——The application of these methods to television codec assessment is considered in this report.

Recently, considerable experience has been gained in the assessment of the performance of high quality codecs for 4:2:2 component television at 34, 45 and 140 Mbit/s [CCIR, 1986-90a]. In these trials, codec performance was examined in terms of basic decoded picture quality, quality after studio post-processes (chromakey and slow motion) applied to the decoded pictures, and the decoded picture impairment associated with the presence of a range of channel error rates. Parts of this report draw upon these experiences.

For distribution applications quality specifications can be expressed in terms of the subjective judgement of observers. Such codecs can in theory therefore be assessed subjectively

or objectively against these specifications. The quality of a codec designed for contribution applications however, could in theory be specified in terms of objective performance parameters because its ouput 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 Report can also be used to assess them.

Throughout this Report the importance of choosing critical test picture sequences, mostly of natural scenes, is stressed and some guidelines on how such sequences may be generated or chosen is given.

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 sections 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 140Mbit/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 quantisation 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 2 or 3 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. Examples of these measurements are given in [Kobayashi, 1977] for intrafield codecs and [Hishiyama & Inoue, 1984] for interframe codecs. 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. A method for classification in such cases is presented in [CCIR, 1982-86]. In that 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 I.

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.

TABLE I

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-p} to minimum quantizing error p-p
- Granular noise	- Equivalent analogue signal-to-noise ratio expressed in terms of $S_{p-p}Nrms$
- "Dirty window" effect	- Maximum noise amplitude
- Temporary image blur	- Rise time of a moving edge
- Jerkiness	- Field or frame difference in terms of moving edge position
- Mosaic-like impairment	- To be studied
- Bit error	- To be studied

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 [Roufs et al., 1989]. 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 section 9) 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 normalised 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 normalised with respect to (the square of) the full amplitude range of the picture samples. Sometimes the normalised 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 (ie 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 recognise 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 optimise 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 section 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 characterisation 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 characterise these distortions (see reservations in section 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 optimise 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 utilisation 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 section 9). 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 normalised mean square error method discussed in section 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 (ie 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 recognise the strong influence which the image content itself has on the sensitivity of an observer to impairments. As was mentioned in section 3.1.4, codec design optimisation involves the use of subjective experiments and psychovisual data to match the distortion-tolerance of the human observer 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 (section 3.1.1) along the lines discussed in [Inoue and Hishiyama, 1984].

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, which is adequate for high quality television applications (see for example, [Sakrison, 1977]). Little is known about the modelling of the response to larger distortions.

A significant study detailing the design of a visual model for picture quality prediction was made by [Lukas & Budrikis, 1982]. Their paper examines the development of this 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 quantisation, 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 another visual model has been developed [Girod, 1988] for application to digital picture coding and [Zetzsche & Hauske, 1989].

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. Unfortunately it has received little attention and more contributions to this topic are encouraged.

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 error rates.

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 (see for example [Jones & Pullum, 1981]). 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 on the decoded image.

[CCIR, 1986-90b]——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 characterised. 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. Subjective Assessment of Codec Picture Quality

Although progress is being made, there is currently insufficient experience to give details of objective picture quality assessment methods for codecs. In the area of subjective assessment, where much experience exists, test conditions and methodologies can be recommended. It must be remembered however when specifying quality or impairment targets, that existing methods cannot give absolute subjective ratings but rather results which are influenced to some extent by the choice of the reference and/or anchor conditions. The same methodologies may be adopted for both fixed and variable word-length codecs, and for intrafield and interframe codecs although the choice of test images sequences may be influenced (see section 9).

At the present time, the most completely reliable method of evaluating the ranking order of high-quality codecs is to assess all the candidate systems at the same time under identical conditions. Tests made independently, where fine differences of quality are involved, should be used for guidance rather than as indisputable evidence of superiority.

5.1 Basic Quality Assessment

Where a codec is being assessed for distribution applications this quality refers to pictures decoded after a single pass through a codec pair. For contribution codecs, basic quality may be assessed after several codecs in series, in order to simulate a typical contribution application.

5.1.1 Viewing conditions and choice of observers

It is recommended that these should be as in section 2.4 of Recommendation 500——————for conventional resolution television and as in————Recommendation 710 ————for HDTV codecs.

5.1.2 Use of test picture sequences

It is recommended that at least six picture sequences be used in the assessment, plus an additional one to be used for demonstration purposes prior to the start of the trial. The sequences should be of the order of 10s in duration but it should be noted that test viewers may prefer a duration of 15-30s [Inoue, 1988] [CCIR, 1986-90c]. They should range between moderately critical and critical in the context of the bit-rate reduction application being considered (see section 9).

5.1.3 Test methodology

Where the range of quality to be assessed is small, as will normally be the case for television codecs, the testing methodology to be used is the double-stimulus continuous quality-scale described in Recommendation 500. The original source sequence will be used as the reference condition. Discussion on the duration of presentation sequences is continuing in IWP 11/4 [CCIR, 1986-90 d,e]. In the recent tests by IWP 11/7 on codecs for 4:2:2 component video [CCIR, 1986-90a, f] with the results given in [CCIR, 1986-90g], it was considered advantageous to modify the presentation from that given in Rec. 500. Composite pictures were used as an additional reference to provide a lower quality level against which to judge the codec performance.

5.2 Post-processed Quality Assessment

This assessment is intended to permit judgement to be made on the suitability of a codec for contribution applications with respect to a particular post-process eg chromakey, slow motion, electronic zoom. The minimum arrangement of equipment for such an assessment is a single pass through the codec under test, followed by the post-process of interest, followed by the viewer. It may however be more respresentative of a contribution application to employ further codecs after the post-process.

5.2.1 Viewing conditions and choice of observers

See section 5.1.1.

5.2.2 Use of test picture sequences

Because of the practical constraints of possibly having to assess a codec with several post-processes, the number of test picture sequences used may be a minimum of three with an additional one available for demonstration purposes. The nature of the sequences will be dependent upon the post-processing task being studied but should range between moderately critical and critical in the context of television bit rate reduction and for the process under consideration. The sequences should be of the order of 10s in duration but it should be noted that test viewers may prefer a duration of 15-30s [Inoue, 1988] [CCIR, 1986-90c]. For slow motion assessment a display rate of 1/10th of the source rate may be suitable.

5.2.3 Test methodology

The test methodology to be used is the double-stimulus continuous quality-scale method described in Recommendation 500. Here however the reference condition will be the source subjected to the same post-processing as the decoded pictures. If inclusion of a lower quality reference is considered to be advantageous then it too should be subjected to the same post-process. In the tests described in [CCIR, 1986-90f] slight modification was made to the presentation given in Rec. 500.

6. Subjective Assessment of Codec Picture Impairment due to Transmission Errors

Section 4 presented some discussion of the way in which transmission errors are handled by a digital decoder, with a view to considering how objective picture quality analysis could be approached. A useful subjective measure may be impairment determined as a function of the rate at which transmission errors occur in the link between coder and decoder. At present there is insufficient experimental knowledge of true transmission error statistics to recommend parameters for a model which accounts for error clustering or bursts. Until this information becomes available Poisson-distributed errors may be used. Some details of data corruptors for application to the 34, 45 and 140Mbit/s hierachical transmission levels are given in [CCIR, 1986-90f].

6.1 Use of test picture sequences

Because of the need to explore codec performance over a range of transmission error rates, practical constraints suggest that 3 test picture sequences with an additional demonstration sequence will probably be adequate. Each sequence should be of the order of 10s in duration but it should be noted that test viewers may prefer a duration of 15-30s [Inoue, 1988] [CCIR, 1986-90c]. It should range between moderately critical and critical in the context of television bit rate reduction (see section 9).

6.2 Choice of error rates

A minimum of 5, but preferably more, error rates should be chosen, approximately logarithmically spaced and spanning the range which gives rise to codec impairments from "imperceptible" to "very annoying".

6.3 Test methodology

As the tests will span the full range of impairment, the double-stimulus impairment scale (EBU) method is appropriate and should be used. The method is described in Recommendation 500.

6.4 A note on the use of very low error rates

It is possible that codec assessments could be required at transmission error rates which result in visible transients so infrequent that they may not be expected to occur during a 10s test sequence period. The presentation timing suggested here is clearly not suitable for such tests.

If recordings of a codec output under fairly low error rate conditions (resulting in a small number of visible transients within a 10s period) are to be made for later editing into subjective assessment presentations, care should be taken to ensure that the recording used is typical of the codec output viewed over a longer time-span.

7. Subjective Comparisons Between Codecs

Where a judgement of absolute codec quality or impairment is not required, but only the ranking order, or where confirmation of the ranking order found from double-stimulus results is desired, the method of paired-stimulus comparisons should be used. This is given in section 4.2 of Recommendation 500.

As it is described, the method provides a sensitive comparison and a means of determining a measure of the relation between pairs of systems. An extension of this method, to ranking the quality or impairment of more than two systems, is given in [CCIR, 1986-90h]. In this approach overall ranking order is derived from the ranking of all possible pairs of picture sequences by the observers.

The analysis is complicated by the fact that an observer can rank, for example, picture A better than picture B, and picture B better than picture C, but also picture C better than picture A. This is termed an "intransitive triad". The statistical treatment of transitivity for each observer is dealt with in [CCIR, 1986-90h, i] as is the important aspect that statistically significant systematic agreement between observers exists.

A problem with the method is that the number of presentations required increases as the square of the number of test picture sequences and codecs, and can become impractical.

8. 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 section 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 be 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 CCIR 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 [Inoue, 1987]. In this paper 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.

9. The Choice of Test Picture Material for Digital Codec Assessment

Throughout this Report, the importance has been stressed of testing digital codecs with picture sequences which are critical in the context of television bit-rate reduction. It is therefore reasonable to ask how critical a particular image sequence is for a particular bit-rate reduction task, or whether one sequence is more critical than another. A simple but not especially helpful answer is that "criticality" means very different things to different codecs. For example, to an intrafield codec a still picture containing much detail could well be critical, while to an interframe codec which is capable of exploiting frame-to-frame similarities, this same scene would present no difficulty at all. Some sequences employing moving texture and complex motion will be critical to all classes of codec so these are most useful to generate or identify. Complex motion may take the form of movements which are predictable to an oberver but not to coding algorithms, such as tortuous periodic motion.

A consideration of possible statistical measures of image criticality [CCIR, 1986-90j]————such as by correlative methods, spectral methods, conditional entropy methods etc. has revealed a simple but useful measure based on an intrafield/interframe adative entropy measurement. This method was used to "calibrate" picture sequences proposed for use in the IWP 11/7 trials of codecs for 34, 45 and 140Mbit/s and proved useful for the selection of the sequences used. The making of such measurements on picture sequences is most easily accomplished by transferring them to image processing computers and subjecting them to analysis by software.

Where access to these techniques is not available, the following presents some general guidelines on how to choose critical material.

Fixed word-length intrafield codecs: while it is possible and valid to assess these codecs on still images, the use of moving sequences is recommended since coding noise processes are easier to observe and this is more realistic of television applications. If still images are used in computer simulations of codecs, processing should be performed over the entire assessment sequence in order to preserve temporal aspects of any source noise, for example. The scenes chosen should contain as many as possible of the following details: static and moving textured areas (some with coloured texture); static and moving objects with sharp high contrast edges at various orientation (some with colour); static plain mid-grey areas. At least one sequence in the ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic (ie computer generated) so that it is free from camera imperfections such as scanning aperture and lag.

Fixed word-length interframe codecs: the test scenes chosen should all contain movement and as many as possible of the following details: moving textured areas (some coloured); objects with sharp, high contrast edges moving in a direction perpendicular to these edges and at various orientations (some coloured). At least one sequence in the ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic.

Variable word-length intrafield codecs: it is recommended that these codecs be tested with moving image sequence material for the same reasons as the fixed word-length codecs. It should be noted that by virtue of its variable word-length coding and associated buffer store, these codecs can dynamically distribute coding bit-capacity throughout the image. Thus, for example, if half of a picture consists of a featureless sky which does not require many bits to code, capacity is saved for the other parts of the picture which can therefore be reproduced with high quality even if they are critical. The important conclusion from this is that if a picture sequence is to be critical for such a codec, the content of every part of the screen should be detailed. It should be filled with moving and static texture, as much colour variation as possible and objects with sharp, high contrast edges. At least one sequence in the text ensemble should exhibit just perceptible source noise and at least one sequence should be synthetic.

Variable word-length interframe codecs: this is the most sophisticated class of codec and the kind which requires the most demanding material to stress it. Not only should every part of the scene be filled with detail as in the intrafield variable word-length case, but this detail should also exhibit motion. Furthermore, since many codecs employ motion compensation methods, the motion throughout the sequence should be complex. Examples of complex motion are: scenes employing simultaneous zooming and panning of a camera; a scene which has as a background a textured or detailed curtain

blowing in the wind; a scene containing objects which are rotating in the three dimensional world; scenes containing detailed objects which accelerate across the screen. All scenes should contain substantial motion of objects with different velocities, textures and high contrast edges as well as a varied colour content. At least one sequence in the test ensemble should exhibit just perceptible source noise, at least one sequence should have complex computer generated camera motion from a natural still picture (so that it is free from noise and camera lag), and at least one sequence should be entirely computer generated.

Test sequences required for post-processing assessments are subject to exactly the same criticality criteria. This may be difficult to achieve however in chromakey foreground sequences because they usually have a significant proportion of featureless blue background.

A comprehensive library of test sequence material has been prepared by IWP 11/7 in 4:2:2 component format and is held on D1 tape. Details of these sequences, together with the criteria by which they were prepared (which may apply to other imaging standards), are given in ——— Report 1213.

REFERENCES

Girod, B [1988] A Model of Human Visual Perception for the Reduction in Redundancy of Television Luminance Signals, PhD. Thesis University of Hannover 1988 (in German).

Hishiyama, K & Inoue, M [September, 1984] Physical Measures for Interframe Coded Picture Quality, Review of ECL, Vol 32, No 5.

INOUE, M. [January, 1988] - The influence of picture presentation period on subjective evaluation, Tech. Rep. of IEICE Japan, IE 87-105 (in Japanese).

Inoue, M [May, [1987] The Proposed Method for Noise Specifications to Mixed Analogue-Digital Video Transmission Systems, J ITEJ, Vol 41, No 5 (in Japanese).

INOUE, M. and HISHIYAMA, K. [1984] - Trade-off between information suppression effect and picture quality degradation with interframe coding, Review of the Electrical Communication Laboratories (Japan), Vol. 32, No. 5.

Jones, W J & Pullum, G G [March, 1981] Error Performance Objectives for Integrated Services Digital Networks Exhibiting Error Clustering, Proc. IEE Conf. on Telecom Transmission, London, Publ no 193.

Kobayashi, Y [May, 1977] Picture Quality Evaluation Method for Digitally Encoded Video Signals, Trans. Inst. Elect. Com. Engrs. Japan, Vol J60-B, No 5 (in Japanese).

Lukas, F X J & Budrikis, Z L [July, 1982] Picture Quality Prediction Based on a Visual Model, IEEE Trans. Com., Vol COM-30, No 7.

ROUFS, J., DE RIDDER, H. and WESTERINK, J. [1989] - Perpetual image quality metrics, Institute for Perception Research, Eindhoven, Manuscript MS 692.

Sakrison, D J [November, 1977] On the Role of the Observer and a Distortion Measure in Image Transmission, IEEE Trans. Com., Vol COM-25, No 11.

Zetzsche, C & Hauske, G [July, 1989] Principal Features of Human Vision in the Context of Image Quality Models, Proc. IEE Int. Conf. on Image Processing and Applications, Warwick, Publ no 307.

CCIR Documents

[1982-86] : 11/26 (CMTT/31) (Japan)

[1986-90]: a. 11/498 (IWP 11/7); b. IWP 11/7-192 (Japan); c. 11/422 (Japan); d. IWP 11/4-154 (Japan); e. IWP 11/4-158 (France); f. IWP 11/4-182(Rev.2) (IWP 11/7); g. 11/498 (IWP 11/7); h. IWP 11/4-175 (Federal Republic of Germany); i. IWP 11/4-156 (Federal Republic of Germany); j. IWP 11/7-261 (United Kingdom).