17.1.86

CCITT SG XV

DOCUMENT No. 65

Working Party on Visual Telephony

Specialist Group on Coding for Visual Telephony

SOURCE:-British Telecom

Title: Objective measures of quality in highly compressed images

This contribution briefly examines previous work on objective measures of quality and considers their efficacy for pictures with large impairments and concludes with a tentative suggestion for an improved method.

1 INTRODUCTION

There are two reasons why a reliable objective measure of picture quality would be valuable.

i) So that coding algorithms could be evaluated easily at each stage of development and a final choice amongst several contenders made independently of subjective influences.

ii) So that compression schemes in which choices of parameter are made by the algorithm on the basis of error measures could be improved.

In the case of i) costly exercises in subjective testing could be minimised. In the case of ii) better code-books for VQ could be generated and better quantisers for DPCM might be found.

2 DIFFICULTIES

The principal problem is the extremely complex processing which takes place in the human brain. Some papers, e.g Lukas(1) attempt to make use of aspects of the human visual system, but concentrate on the early processing in the eye which is well understood, and ignoring the later processing of the brain. This paper has been used in COST211 bis (ref.4) as part of a basis for a measure. The masking of impairments by high activity regions in the picture is an important component of this approach. In the past most subjective tests for picture quality have been attempts to determine the thresholds at which errors become visible. In COST211 bis on the other hand we are often concerned to choose between pictures where far from being invisible the impairments are very obvious, quite disturbing, and of completely different sorts in the compared pictures. This is a very different situation and we should not be surprised if results obtained by earlier workers are inappropriate for our purposes.

3 POSSIBLE REASONS FOR POOR RESULTS OF OBJECTIVE MEASURES AT LOW BITRATES

A numerical measure attempts to put a single figure of quality on a picture, but the impairments that may be seen are of many types. Each coding scheme has its own particular form of degradation and in a sense a vector quality measure would be more reasonable except that vectors do not lead to a quality ordering. There seem to be two kinds of impairments for high compression, those where the impairmant is rather evenly distributed, for example , blur and white noise, and those where the impairments form structures themselves, e.g block edges, contouring, ringing and stepping. A common feature of the latter type is that they usually form one-dimensional subspaces of the two dimensional array. As is mentioned below this fact has not been taken account of in error measures and may be very signficant.

There are other much more complex and obscure factors which influence our subjective judgement. They are best seen in terms of a few examples:

a) Picture material influences judgement: e.g, a slightly green cast can be completely acceptable in a landscape but totally unacceptable in a human face.

b) Positioning of impairment in picture is important: e.g a block edge crossing a face can render an otherwise good picture unacceptable.

c) Judgements of subjective quality may depend upon the distance an observer sits from the screen. Blur, becoming visible closer to the screen can be more objectionable than blocking since the eye feels uncomfortable if it cannot focus. Yet from further back the softness is not perceived whereas the blocks are still visible and become more annoying than the blur.

4 NUMERICAL INCOMPATIBILITY

Perhaps the most subtle difficulty is the "dimensionality" problem. As we all know the ideal point has no dimensions and the ideal line has only one dimension. Point-like and line-like artefacts may therefore consist of almost no points in the mathematical sense, relative to the total number of pels. When a point-like or line-like error is measured it may make very little contribution to the MSE or other objective error measure. But viewed subjectively a bright or black line, for example, may be considered a very gross impairment. It is very difficult to believe that 2 (a), (b) or (c) can be incorporated into a mathematical objective measure in the near future since they require that scenes and parts of pictures can be recognised without human interference or that account may be taken of seating arrangements. Numerical incompatibility, on the other hand, seems more susceptible of study since there is no reason why a machine could not recognise large isolated errors and lines of isolated errors. It may be supposed that measuring maximum errors rather than mean square errors might help but such an approach does not automatically recognise the high-visibility of errors in the near spatial or temporal presence of perfectly correct pels. The situation is even further complicated if one considers the effect of block edges. Although they are perceived as lines there are no real lines of errors, since the line in this case is a genuine one-dimensional

artifact with no width at all. It can be argued that there are errors of a two dimensional nature at either side of the block edge, but these may be very slight yet at the edge itself a far greater impairment is visible than would be the case with larger errors in the interior of the blocks. These interior errors would nevertheless make a greater contribution to any simple objective measure so that the subjective judgement would not be similar to the objective measure.

5 RESULTS

Several authors have considered the relationship between subjective testing and objective measures. For analog transmission of television signals Lewis(2) did studies using measurements of a test waveform (Pulse-and-bar) and comparing them to the results of subjective tests. Very useful results were obtained but they applied principally to the one-dimensional effects of transmission lines on fairly high-quality pictures. It may be that a good two-dimensional waveform- or test pattern- could be developed in which particular attention could be paid to the kinds of distortion which arise from DCT, VQ and other current forms of image data compression. Of course such a test pattern should never itself be used for making subjective judgements of quality as is sometimes mistakenly done with testcards where the frequencies beloved by engineers can so readily be seen.

Limb(3) has considered the effects of several forms of distortion on a set of five pictures and compared both MSE and several other error measures including higher degree means. He also considered amplitude weighted error measures where picture activity is taken into account. The forms of distortion he considered were analagous to rather low compressions -for example 4-bits per pel. His work was restricted to still pictures so that it was only partially relevant to the COST 211bis project. He came to the conclusion that MSE performed rather well and did not seem to strongly recommend the model he developed in his study, indeed he considers that local measures are only marginally better.

Measures based on the work of Limb and Lukas have been proposed within COST 211bis (4). They have been tested by workers at the University of Hannover(5) who report that for small impairments in broadcast images the new measure is better than MSE, but that when impairments become gross the advantage of the new method becomes neglegible. The reason for this may be that later processing of the brain becomes more and more significant as impairments become greater. Also artefacts begin to appear at greater compressions so that the comments of 2(d) above become relevant.

Minimax errors have been used at BTRL in attempting to improve the codebooks in VQ but if any thing gave worse results than MSE.

6 FUTURE WORK

The difficulty mentioned in 2(c) could possibly be tackled by looking for high values of the gradient of the error function as well as high values of the error itself. For example a block structure might be a very significant contribution to an error measure of this kind. I therefore propose the following form of objective error measure.

Let $I=I(i,j)$ be a typical image function and let a norm be defined as $NI=\|I(i,j)\|$ where any desired measure of $I(i,j)$ (mean square, minimax etc) can be used as the norm. Then if $S(i,j)$ is a source image

and C(i,j) is a coded image we let

$$E(i,j)=S(i,j)-C(i,j)$$

We define f and g as scalar thresholds and F(i,j) and G(i,j) as error functions such that :

$$F(i,j)=0 \text{ for } E(i,j)<f$$
$$F(i,j)=E(i,j) \text{ for } E(i,j)>=f$$

$$G(i,j)=0 \text{ for } *E(i,j)<g$$
$$G(i,j)=*E(i,j) \text{ for } *E(i,j)>=g$$

where * represents a gradient function.

We are then in a position to define a new objective measure M:

$$M=p\|F(i,j)\|+q\|G(i,j)\|$$

where p and q are suitable weighting factors.

The function G(i,j) should probably be filtered to pick out the coherent artefacts from the noise that would certainly be present in an error signal of this kind.

We note that M is a function of p,q,f,g and determining the optimum value of these parameters and testing their consistency would be a way of finding and testing a possible new objective measure.

It has also been observed in the work on RBN at BTRL that errors in the second derivative are offensive to the eye, especially in regions where there is no activity masking. If the first derivatives proved useful another term could be added to the measure to account for the second.

It should be noticed that the proposed new measure does not appear to have taken account of local effects due to masking, but it can be hoped that since high activity usually corresponds to high gradient values such an effect could tend to modify the value of q. There is also no reason why E(i,j) could not be obtained using ideas of masking, the resultes could be combined with those of this suggestion to obtain a more sophisticated result.

This suggestion applies to still picture assessment in its present form, but a time gradient could easily be added to account for time dependent impairments.

CONCLUSION

It appears that in the case of high-compression algorithms no objective measure has been found which is markedly superior to MSE even though such a measure would be very desirable. The reasons for the difficulties are not hard to see and there are probably more than those mentioned in this document. We have observed that the existence of artefacts begins to play an important part at high compressions and a possible way of measuring these kind of impairments together with more common degradations has been proposed.

REFERENCES

(1) F X J Lukas, Z L Budrikis,

´Picture Quality assessment based on Human Visual Model´, IEEE Trans. on Comm., Vol. COM-30, no.7,pp., July 1982

(2) N W Lewis, ´Waveform Responses of Television links´ Proc. IEE Paper No.1688R, July 1954(Vol 101, Part III, p 258.)

(3) J O Limb, ´Distortion Criteria of the Human Viewer´ IEEE Trans.Sys.Man.,Cybern.,vol SMC-9,no.12,pp 778-793, Dec 1979.

(4) COST 211 bis Document SIM 85/65

(5) B Girod, Discussion in COST211bis simulation subgroup.