CQA – Subjective Video Codec Quality Analyser

Tim Morris, Kevin Angus, Richard Butt, Andrew Chilton,
Paul Dettman, Steven McCoy

Department of Computation
UMIST
Manchester M60 1QD, UK

`t.morris@co.umist.ac.uk`

Abstract

A subjective video codec quality analyser (CQA) has been designed and implemented. It was specified to take two video streams as input: the original, uncompressed, stream and the compressed/decompressed stream. The CQA compared the subjective quality of the coded stream with respect to the uncoded one by computing the values of a number of factors designed to assess artefacts that were subjectively important. The values were then combined using a neural network to deliver, as output, a single quality measure that correlated with the subjective assessment of video data.

1. Introduction

In assessing the effectiveness of an image or video-coding algorithm, two factors must be considered. One is how much compression is achieved and the other is the quality of the subsequently decoded data. The degree of compression is easily measured by comparing the volumes of the coded and raw data. The result is expressed either as a compression ratio or as the number of bits required to store each pixel of the coded data. The quality of the decoded data could be measured by having an expert rate the data, but this would be an expensive solution and is not guaranteed to be free from bias. Software for automatically assessing the quality of the decoded data is the subject of this paper.

Many authors have quantified image quality by computing the differences between the original and the reconstructed (coded/decoded) data. The maximum difference or the average difference was quoted, much as one would quote peak or mean signal to noise ratios. However, such a measurement can do little more than hint at the quality of the reconstructed data; it is not difficult to imagine situations where the subjective quality differs from what one would expect from these difference measures. It is also argued that these difference measurements are not relevant in this situation: we are coding information that is to be viewed by humans, a measurement that assesses the data's subjective quality is therefore required.

Hosaka [1] suggested that the quality of a reconstructed image could be measured by dividing the original image and its reconstruction into non-overlapping blocks. The differences between the means and standard deviations of grey values in each block were plotted on a scatter diagram whose area and shape were descriptive of the quality of the reconstructed image. The method did not, however, take any account of the perceptually important block structure of the coding, nor other artefacts found to be annoying to human observers.

Other authors [2-4] have attempted to incorporate models of the human visual system into quality measures. The intention was that such measurements would be sensitive to those artifacts that detract from the perceived quality of an image. For example, Xu and Hauske [5], suggested that the perceptually important regions of an image were edges, and uniformly coloured and textured regions. Bock et al. [6] presented a system in which images were divided into uniformly coloured and textured, and edge regions and differences in these regions between the original and reconstructed images were accumulated. The accumulated differences were weighted and summed. No attempt was made to derive a score that related to an observer's assessment of the image's quality.

Algazi et al [7] described an early attempt at quantifying the subjective quality of reconstructed images. They derived a number of factors from still images: two weighted differences between source and reconstructed images, a factor relating to the codec's block structure and a factor relating to edges. These factors were found to be strongly correlated. They were decorrelated using a principal component analysis. A picture quality score (PQS) was derived by taking a linear combination of the components, such that the resulting score correlated linearly with the mean opinion score (MOS) estimated by nine observers. The correlation coefficient between the PQS and the MOS was 0.88. The PQS was then used to assess the quality of the images reconstructed from a number of different wavelet based encoders.

Subsequently, Cireddu et al. [8] performed similar work deriving distortion measures from an image and combining them to give a single quality measure. They derived the distortion measure from an analysis of the human visual system, but do not seem to have studied how the measures should be combined.

This paper describes an experimental codec quality analyser (called CQA) that we have developed. It was designed to give a subjective measure of the degradations introduced into a video stream by the coding process – specifically block based codecs such as MPEG. This was achieved by measuring subjectively important differences between the raw and coded/decoded (reconstructed) video streams. The measurements were then combined using a neural network to give a single value that reflected the subjective quality of the video stream. The remainder of the paper is organised in five sections. The first section describes the experiments that were performed in order to obtain subjective quality scores for a range of image sequences. We then describe the individual quality measures that we have derived or adopted. The third section describes the neural network that was derived to combine the individual quality measures to give a number similar to (the same as) the previously derived subjective assessment. We then present and discuss our results. The final section evaluates the CQA and outlines its future development.

2. Experimental Methods

Our intention was to derive a method of assessing the subjective quality of samples of video data from a wide range of subject matter. It was therefore important to gather a representative sample of video clips, and encode them to varying degrees of compression – and hence varying degrees of subjective quality. Ten video clips were used, sufficient to train and test the neural network we used and to have a wide range of subject matter, but few enough to prevent the test sessions being over long and tiring the observers. The video clips were captured from VHS tape. Their length was standardised to about ten seconds, long enough for an observer to rate the clip but not so long that the observer became familiar with the degradations in it and therefore able to compensate for them. The clips themselves covered a range of broadcast material, they were taken from cartoons (to give samples of images having regions of uniform, saturated colour) commercials and soap operas (typical tv program material), weather forecast (low motion, bright colours and small details) and the news (very little motion). We considered it important that the clips were of subject matter that would be familiar to the test subjects. Each clip was encoded to 60, 90, 120 and 200 kbs^{-1} and MJPEG, which was used as a control clip. The samples' resolutions were 352 by 288 pixels by 24 bit colour. The soundtrack was removed.

For assessment, the clips were displayed with a height of 15 cm to observers who sat at a distance of about 60 cm from the monitor. The observers were not trained in assessing the quality of video clips. Each of four versions of the ten clips was shown to the observers (either the 90 or 200 kbs^{-1} clip was omitted). The order of the presentation was randomised for each observer in order to mask any tendancy for the observer to rate a clip in relation to the previous one. In addition, five clips were randomly chosen and shown a second time, which allowed us to evaluate the consistency of the observers. The observers were asked to rate each clip on a five point scale, ranging from very high quality to very poor quality. The evaluation was translated to a numerical value for processing. A total of eighteen observers evaluated the clips; although this was a small number, the results appear to be repeatable and self consistent. The result of this experiment was a mean observer score (MOS) for each version of the video clips. The results of this will be presented below.

3. Quality Measures

Figure 1 outlines the architecture of the CQA. Conceptually it is very similar to other systems that measure the quality of an image or image sequence, in that it consists of a component to compute the values of particular measures, and a component that will combine the values into a quality score. This section of the paper will describe the quality measurements we have extracted from the image. The following section will describe the methods used to combine the measurements. The aim of this was to be able to automatically derive a score that was the same as, or at least correlated with the MOS.

The features we have used to assess the quality of the image sequences are tabulated in table 1 and discussed below. Deriving the values of these features was a two phase process. The first computed values of certain parameters for each image in the sequence. The second phase then combined these measures to give the equivalent

measurement for the entire sequence. For example, M0 was a measure of the difference between raw and reconstructed data, its value was computed by summing the difference

Original
Image
Coded
Image

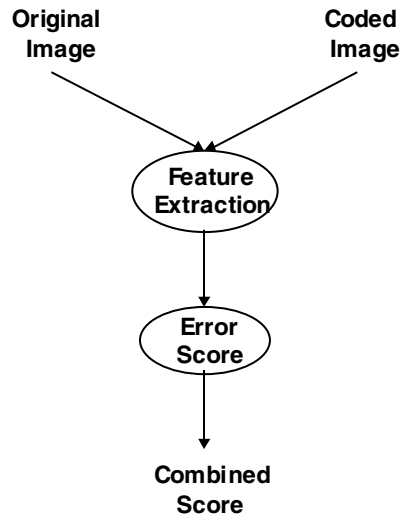Feature
Extraction

Error
Score

Combined
Score

Figure 1: Outline architecture of CQA.

between equivalent frames in the sequences (first phase) and then computing the average of these values (second phase).

In the assessment, two quantities, TI (temporal information) and SI (spatial information) were used. SI was a measure suggested by ITS and was the variance of the Sobel-enhanced image. Within an image sequence, each image had an SI value, we therefore had two sets of SI values – for the raw and reconstructed sequences. A single, normalised value was computed by differencing the pair and dividing by the SI value of the raw image. Similarly, the TI measure was computed by taking the ratio of the root mean square values of the absolute differences between temporally adjacent images in the raw and reconstructed sequences. Features characteristic of particular image sequence degradations were derived from these measures.

Quality Measure	Description
M0	Difference in Y values
M1	Maximum of TI
M2	RMS of TI
M3	Range of TI values
M4	Difference between positive and negative means of TI values

M5	RMS of TI error ratios
M6	RMS of positive TI error ratios
M7	Maximum SI
M8	RMS of SI
M9	Difference in U values
M10	Difference in V values
M11	Difference in Hue
M12	Difference in Saturation
M13	Difference in Lightness
M14	Difference in Value
M15	Range of mean square errors
M16	RMS of errors in Y values
M17	Number of edge pixels

Table 1: Quality measurements derived from the image sequence.

The measurements derived from the image sequence may be grouped according to the type of distortion they were designed to assess.

Measures M0, M15 and M16 are traditional error measurements: the mean difference between the sequences, the mean squared difference and the root mean squared difference. These measurements provided objective measures of the error in the reconstructed data, but were insufficient to measure the objective quality of the sequence.

Cirreddu et al [8] and others have suggested that major causes of degradation are blockiness (distortion of the image characterised by the appearance of an underlying block encoding structure), blurring (reduced sharpness of edges and spatial detail) and jerkiness (motion that was originally smooth and continuous is perceived as a series of distinct snapshots). The remaining groups of measurements were designed to assess these factors.

Measures M7, M8 and M17 were designed to evaluate the blockiness of the reconstructed data. M7 returned the maximum SI value and M8 returned the RMS SI value. M17 was designed to provide a measure proportional to the edges introduced by the block-structured nature of the codec. The exactly horizontal and exactly vertical edges of the reconstructed image were therefore enhanced (using the Sobel operator) and subtracted from the original. This left just those horizontal and vertical edges introduced by the coding process. Figure 2 is a sample derived from a single frame of a

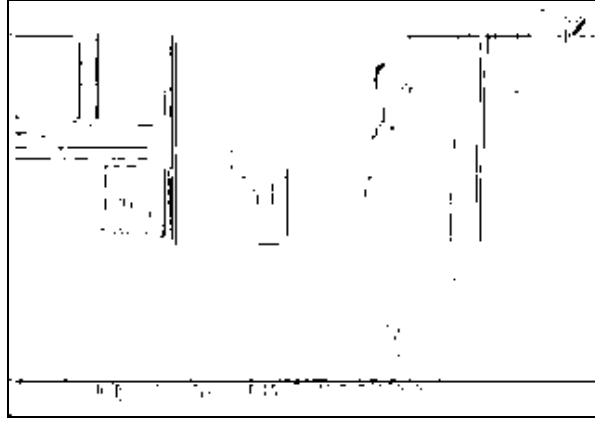typical video sequence. The length of the lines in each image of the sequence was

Figure 2: Sample image, horizontal and vertical edges enhanced.

summed, and the mean length assigned to M17.

Features M1 to M4 were chosen to reflect the jerkiness of the sequence, each used the TI measure described above in a manner stated in table 1. In a similar manner, M5 and M6 returned slightly different combinations of the TI values derived from the raw and reconstructed image sequences.

M9 to M14 were defined to reflect the fidelity of the colour information. Differences between the three colour values of two colour models were computed in phase one and in phase two the means were taken. HSV and Y, Cb and Cr colour models were used.

In summary, eighteen measures of image quality were derived, which assessed the colour fidelity, the edge and motion artefacts as well as the traditional difference measures. Our intention was to combine the measures in a manner that yielded a single quality score per image sequence that correlated with a mean operator score. The neural network derived for this purpose is the subject of the following section.

4. Combining Quality Measures

The quality measurement modules computed the values of a number of features believed to be important in ascertaining the subjective quality of an image. The major requirement of our system was that it should deliver a single value that reflected the subjective quality of the video data; large values should equate to "good looking" data and low values to unacceptable data. This section of the paper will describe the neural network that we have used to achieve this goal.

A feedforward network trained by the backpropagation method was used. The network's inputs were the eighteen quality features, the network therefore had to have

eighteen input nodes. Its output was to be an integer quality score in the range 1 to 5; therefore five output nodes were used. Experimentation suggested that 25 hidden nodes gave a sufficiently accurate result. Network experiments were performed using the Slug neural network simulator.

Training and testing of the network was achieved by selecting about three quarters of the data to train the network and using the remaining clips for testing. The test data was selected in two ways: first by taking all coded versions of seven clips for training and all versions of the remaining clips for testing (Method 1 of the following section); and secondly by taking ¾ of the coded versions of each clip for training and using the remaining ones for testing (Method 2). The training was repeated eight times in each case. Results are presented below.

5. Results and Discussion

This section will present the results of our investigations. Specifically, we shall demonstrate how the original subjective assessments (MOS) varied with respect to the compression of the video data. We shall present the accuracy results of the neural network and compare the network's scoring of a video clip with the MOS. Finally we shall examine and discuss which of the features contributed most to the quality measure.

Firstly, the Mean Observer Scores of all of the clips at all compression levels were avaraged, giving a value of 3.48. It is supposed that for an "ideal" set of clips observed by an "ideal" set of observers this average MOS would be 3.00 indicating a balance between "good" and "bad" clips. The proximity of our avarage to the expected is indicative that the clips have not been badly chosen.

The MOSs of the "reliability" clips (those clips shown twice to each observer) were examined. The mean absolute difference between the scores of the two clips was 0.53 – indicating that the observers were being consistent.
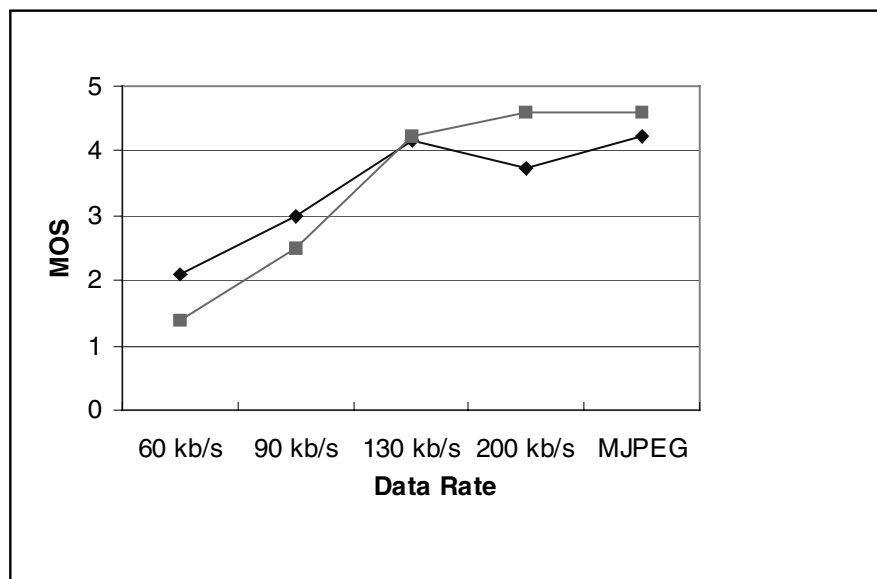

Figure 3: MOS values vs data rate for two representative clips.

General trends were noted for all of the MOS, typified by the curves of figure 3 which presents the Mean Observer Score with respect to data rate for two of the clips we have used. It should first be noted that the MOS increases as the data rate increases, as would be expected, and is a maximum for the MJEP coded clips, which may be regarded as the lossless coded data. Secondly, it is observed that the gradient of the curves is reduced at data rates greater than about 130 kbs^{-1}, the exact breakpoint varied. The origins of this phenomenon lie with the information content of the clip, above some data rate a barely perceptible amount of information has been lost.

It was also noted that clips with high mean brightnesses consistently gave a lower MOS, possibly because the high brightness made many of the coding errors are more apparent. The reverse was also true, clips darker than the average had lower than average MOS.

Finally, it must be noted that merely quoting the data rate is not an accurate means of quantifying the quality of digital video, even though this is a simple measurement to make.

Table 3 presents the accuracy rates achieved by the network using the two methods of dividing data between training and testing sets (i.e. what proportion of the test set were given the correct MOS). Eight networks were trained and assessed.

Test Number	Accuracy, Method 1	Accuracy, Method 2
1	25%	78%
2	50%	78%
3	50%	78%
4	87%	33%
5	31%	33%
6	87%	78%
7	-	56%
8	25%	56%

Table 3: Neural Network Accuracy Rates.

The second training method produced superior results, probably because it incorporated a better spread of content in the clips used for training, although the accuracy was extremely variable. This suggests that ten clips was really too small a number to do any more than demonstrate that this approach to quality assessment is viable. Network number 1 trained using method 2 was used for the remaining tests. Figure 4 compares the results of the subjective MOS and the output of the best of the trained networks. The horizontal axis represents the video clips ordered roughly according to their expected quality result, hence the positive correlation between median of the MOS values for a

clip and the (arbitrary) clip number. It is apparent that over most of the clips, the neural network has given a value in agreement with the subjective score, as we would expect.
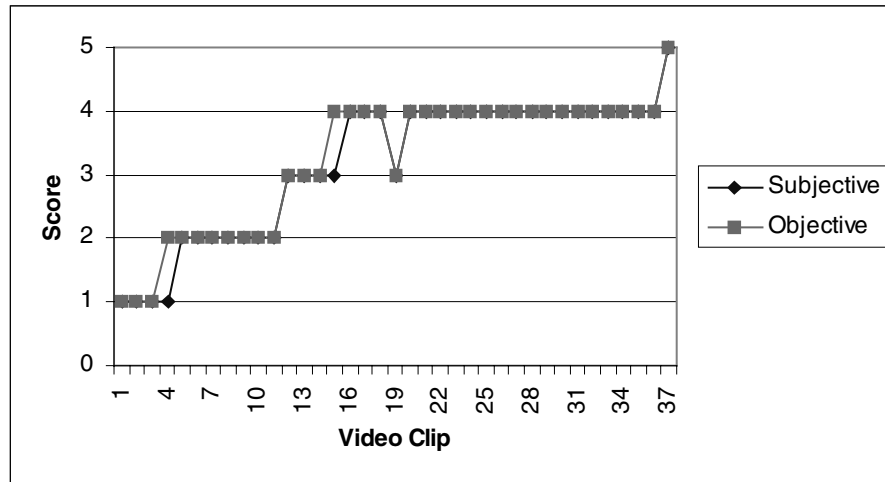

Figure 4: Comparison of MOS and Neural Network output.

Finally, is it possible to identify which features contribute most to the objective result? An analysis of the correlations between the median MOS and the individual measurement values over all of the clips, indicated that those measurements relating to block boundaries (M17), motion (M1, M2, M3, M4, M6) and colour (M10, M11, M14) demonstrated the most significant correlations. The traditional difference measures demonstrated weak correlations with the subjective quality measures, although they did relate strongly to the data rates.

6. Conclusions and Future Work

This paper has presented a system designed to compute a measure of the quality of a coded and subsequently decoded video clip that mimics the subjective assessment of quality by an untrained human observer. The quality measure is derived from a comparison of the original clip and the reconstructed one. The system functions in two stages, a number of quality measures are derived from the clips which are then fed into a neural network trained to associate these values with the equivalent subjective score.

Within the limitations imposed by the rather small number of clips and observers in this initial study, we claim that the system is successful in that the neural network's output is correct for 78% of the test data. It is supposed that this success rate would improve with more test data.

The study has indicated that the traditional measures of image quality, based on the differences between raw and reconstructed image sequences are inadeqate. Rather, measures derived from blockiness, jerkiness and colour relate more closely to the subjective quality of a clip.

In future, we intend to extend the range of clips and the number of observers assessing those clips so as to improve the agreement of the neural network system and the subjective assessments. We will also investigate the effectiveness of reducing the number of measures input to the neural network – it may well be possible that ineffective measures are adding noise and obscuring the relevant information.

We have deliberately ignored the audio data for the purposes of this study. Some authors believe that the audio data is as important or even more important that the visual data in determining the overall quality of a video clip. Measuring the quality of an audio track ought to be an exercise that could be performed in the same manner as the present study. However, measuring the interactions between the audio and visual data and their joint effects on video quality is a more involved problem.

References

[1] K. Hosaka, *A new picture quality evaluation method,* PCS'86, Tokyo, Japan, April 1986.

[2] N. Jayant, J. Johnston, R. Safranek, *Signal compression based on models of human perception,* IEEE Proc 81(10), pp 1383-1422, 1993.

[3] J.A. Saghri, *Image quality measure based on a human visual system model,* Optical Engineering 28(7), July 1989.

[4] X. Ran, N. Farvardin, *A perceptually motivated three component image model,* IEEE Trans Image Proc 4(6), pp 401-415, April 1995.

[5] W. Xu, G. Hauske, *Perceptually relevant error classification in the context of picture coding*, IEEE Conference on Picture Coding and Its Applications, pp 589-593, 1995.

[6] F. Bock, H. Walter, M. Wilde, *A new distortion measure for the assessment of decoded images adapted to human perception*, Proc IWISP'97, pp 215-218, November 1997.

[7] V.R. Algazi, Y. Kato, M. Miyahara, K. Kotani*, Comparison of image coding techniques with a picture quality scale*, Proc SPIE Applications of Digital Image Processing XV 1771, 1992.

[8] M Cireddu, F.G.B. de Natale, D.D. Giusto, P. Pes, *Blockness Distortion Evaluation in Block-Coded Pictures*, Proceedings IWISP 1996, Elsevier Science, 1996.