Recompression of Still Images

Internal Report
March 1992

Syin Chan
Computing Laboratory
University of Kent, Canterbury
Kent CT2 7NF , United Kingdom

Abstract

The effects of subjecting a digital image to more than one step of the lossy JPEG baseline data compression are studied. The performance in terms of the signal-to-noise ratio as well as the visual artifacts are observed and presented. Two types of visual artifacts have been identified and they are found to correlate closely to the two types of errors introduced when the Discrete Cosine Transform coefficients undergo repeated quantisation as part of the JPEG compression algorithm. A method is proposed to estimate the amount of errors generated in the process, which may be used to determine if any special processing is required to reduce these errors. Several methods to overcome the more severe artifacts in the images are investigated and the experimental results are presented.

1 Introduction and Objective

Along with the emergence of multimedia applications, there is a rapid growth in the usage of pictorial information on computers. Much of this pictorial information is acquired with devices such as scanners and cameras, and are subsequently digitised into a numerical representation suitable for input into a digital computer. For example, a colour picture or image may be digitised to a size of 640-by-640 pixels, with each of the colours red, green and blue represented by eight bits, thus generating a total of 9.8 Mbits ($640 \times 640 \times 3 \times 8$), equivalent to 1.2 Mbytes, of data. In the case of digital video pictures, the amount of data is even greater. In order to capture the video image sequences without causing noticeable jerkiness during playback, typically, a frame rate of 25 frames per second is required. Therefore, for the same image size and colour representation, the rate of data being generated is 30 Mbytes per second. Such huge amount of data poses problems in both storage and transmission, thus much research has been carried out in the area of image data compression.

In an ideal situation, where both storage capacity and device speed allow, images may be captured and digitised with full resolution, colour information and frame rate. This original image data may subsequently be retrieved from storage, and compressed to a wide range of desired image data size, subjected to the user's requirement of the image quality. There are, however, situations where the retrieved image is in the compressed form, and the uncompressed original image data is simply not available. Such situations arise because of two reasons. Firstly, due to difficulties in handling the huge amount of data generated in digitising an image, very often data compression is incorporated in the image capture and digitisation system. The images captured and compressed may be stored for local usage, or be distributed to other locations through computer networks. As such, after the image has been captured and compressed in the digitisation process, it is not always possible to capture the same image again but without applying any data compression. If the image data needs to be further reduced for certain applications, the operation will involve compressing an already-compressed image.

Secondly, with the increasing popularity of image data compression/decompression facilities, most images, whether digitised with or without compression, are likely to be stored on disks or transmitted through the various networks in a compressed format. For different applications, a remote recipient of one of these compressed images may find that a lower quality of the image is sufficient, it is then desirable for this recipient to have the capability to further compress the image data. Such an operation, referred to as *recompression* in this study, will not pose any problem when one is working with lossless compression techniques, but will lead to important consequences if lossy compression techniques are used.

The objective of this study is to investigate the feasibility of applying repeated lossy type compression on still images to obtain smaller image file size in the absence of the original image data. This is achieved by studying the kind of distortions introduced in a recompressed image and by identifying some methods which can reduce such distortions.

2 Scope

This study is carried out by applying the ISO (International Standard Organisation) JPEG (Joint Photographic Experts Group) baseline compression technique on monochrome eight-bit greyscale images of natural scenes. The JPEG technique is designed for compressing natural scene images and is not expected to perform as well for cartoon and animation type of images. Experiments were conducted to examine the artifacts in the recompressed images, and several methods were proposed and evaluated for their effectiveness in removing the artifacts.

3 Background

3.1 The JPEG Compression Algorithm

For details of the ISO JPEG image compression standard, reference may be made to a draft document on its technical specifications [1]. The development of the JPEG standard may be referenced from [2], [3] and [4]. The JPEG standard proposes a four-part algorithm definition:

- Baseline sequential encoding
- Progressive encoding
- Lossless encoding
- Hierarchical encoding

In *lossless* encoding the image data is compressed such that the original image can be reconstructed from this compressed form without any distortion. The other three algorithms are all *lossy*, that is, the reconstructed image closely resembles but is not identical to the original image. The JPEG baseline sequential encoding algorithm is briefly described here, more detailed descriptions may be found in [5] and [6]. An image is first divided into non-overlapping blocks of eight-by-eight pixels. Within each block, the spatial redundancy is removed by performing a two-dimensional *Discrete Cosine Transform (DCT)* [7] on the pixels to obtain a corresponding block of eight-by-eight DCT coefficients. These coefficients represent the spatial frequency components of the image block and are arranged such that the coefficient in the upper-left hand corner of the block is the DC coefficient, which measures the energy of the zero-frequency term. The other 63 AC coefficients represent the strengths of the components with increasing horizontal frequency from left to right, and of components with increasing vertical frequency from top to bottom. This is illustrated in figure 1.

In the normal ordering, the 64 DCT coefficients, numbered from 0 to 63, are ordered from left to right and from top to bottom. Each of the 64 coefficients is then quantised using one of 64 corresponding values from a quantisation table. JPEG does not specify any quantisation table but does include a good example of a set of uniform quantisers for the 64 coefficients, this set of quantisers makes use of the human visual frequency response property. This property allows the higher frequency components to be quantised to a greater extent than the lower frequency components without affecting the image's appearance to the human eye. The quantised AC coefficients are then scanned in a zig-zag manner as shown in figure 2, such that the lower frequency components are scanned before the higher frequency components. This has the advantage of exploiting the fact

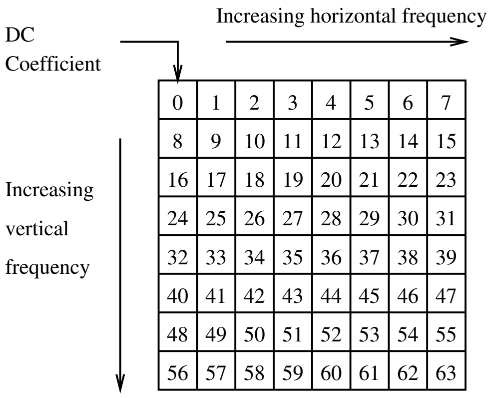

Figure 1: A block of eight-by-eight DCT coefficients numbered in normal ordering

that most of the high frequency coefficients will be quantised to zero, thus making the subsequent *run-length* [8] and *Huffman coding* [9] more efficient. The DCT coefficients in the zig-zag scan path are re-numbered accordingly as shown in figure 3.

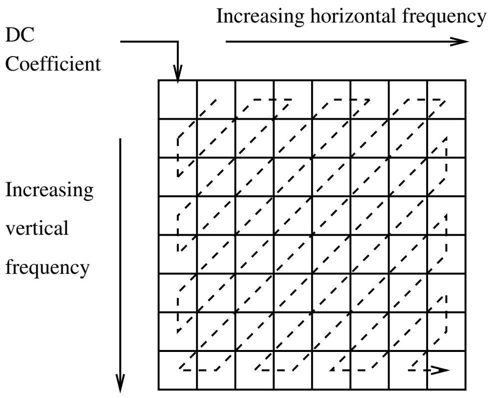

Figure 2: Zigzag scanning of an eight-by-eight block of DCT ceofficients

The DC coefficients are encoded in a slightly different manner. To exploit the high spatial correlation in an image, the DC coefficients are coded using the *Differential Pulse Code Modulation (DPCM)* [10], that is, the difference between the current block's quantised DC coefficient and the previous block's quantised DC coefficient, rather than the DC coefficient itself, is coded. The AC coefficients are first run-length coded and then entropy coded using the Huffman coding. In run-length coding, apart from coding the symbols, the number of consecutive zeros between these symbols are also coded. This method is very efficient since many of the high frequency DCT coefficients are quantised to zero, and many of these zero coefficients occur consecutively in the

Increasing horizontal frequency

DC Coefficient

Increasing vertical frequency

0	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Figure 3: A block of DCT ceofficients numbered in zigzag order

zig-zag scan. The DPCM coded DC coefficients and the run-length coded AC coefficients are then entropy coded using Huffman coding. The Huffman coding is a variable length coding method that allocates shorter code words to frequently occurring symbols and longer code words to less frequently occurring symbols. The coded bit-streams together with appropriate markers specified in the JPEG standard form the compressed image data. To decompress the image, the whole process is simply reversed.

3.2 The Free JPEG Software Implementation

This study has been carried out using a free JPEG software developed by the Independent JPEG Group. This software is copyright (C) 1991, Thomas G. Lane. The software implements JPEG baseline and extended-sequential compression processes. Both the compression and decompression programs are provided. To compress an image, the user needs only to select the amount of compression by specifying a *quality factor*, also known as the *Q factor*. In this implementation, if the input value of Q is 50, the quantisation table provided in the JPEG draft document is used. For smaller values of Q, that is lower quality, the quantisation steps in the table are linearly scaled by Q to obtain bigger steps. For larger values of Q, that is higher quality, the quantisation steps are linearly scaled by Q to obtain smaller steps. This user interface is specific to the free JPEG software implementation and is not part of the JPEG requirement.

3.3 Definition Of Terms

This section defines some of the terms used in this study:

- The *original* or *uncompressed image* is a digitised image that has not undergone any compression. It is usually kept in the *Portable Pixel Map (ppm)* format in this study.
- The *compressed image* refers to an image which has undergone a compression and is kept in a compressed format, in this case, the *JFIF* format which is based on the interchange format specified in the JPEG draft document.
- The *decompressed image* is obtained after decompressing a compressed image, the image is usually kept in the ppm format in this study.
- *Recompression* refers to further compressing a compressed image using the same compression algorithm. Usually it involves compressing a compressed image to a lower quality than before.
- The *quantised coefficients* in a DCT block refers to DCT coefficients that have been quantised in the compression process.
- The *unquantised coefficients* refer to the DCT coefficients which have not been quantised.
- The *dequantised coefficients* are obtained after reconstructing the DCT coefficients from the quantised form using the quantisation table.
- *Requantisation* means further quantising an already quantised DCT coefficient in a recompression operation.
- The *signal-to-noise ratio (SNR)* measures the amount of distortion in a decompressed image, it is defined by:

$$\text{SNR} = 10 \times \log_{10} \frac{(\text{peak-to-peak value of original image data})^2}{\text{mean-square error}}$$

  The SNR is not an ideal fidelity measure of an image meant for human vision, as the human eye's sensitivity to distortions is non-linear, also, the SNR is not likely to reveal localised distortions. To determine the quality of an image, it is necessary to complement the SNR figure with visual inspection. However, it is still a useful measure since it serves as an objective indicator to the average amount of distortion introduced in the compression/decompression process. For example, figures A.1 through A.3 are three versions of the image Roses, each of them has been compressed to a different quality and therefore has different SNR values. Perceptually, figure A.2 is not much worse than figure A.1 although there is a large difference in the SNR values (53.09 dB - 33.38 dB = 19.71 dB), but figure A.3 appears significantly poorer than figure A.2 even though the difference in SNR is merely (33.38 dB - 28.85 dB) 4.53 dB.

- The *bit rate* of an image provides a figure of merit for the efficiency of the compression. In an uncompressed eight-bit greyscale image, the bit rate is simply 8 bits/pixel. For a compressed image, the bit rate is calculated by:

$$\text{bit rate} = \frac{\text{(compressed image file size in bits)}}{\text{(total number of pixels in image)}} \quad \text{bits/pixel}$$

4 Artifacts In Recompression

Some experiments were carried out on three monochrome eight-bit greyscale images to investigate the problems when an image is subjected to more than one JPEG baseline compression. The following test images were used:

		Image size (width × height)
1)	Roses	640 pixels × 480 pixels
2)	Mandrill	512 pixels × 512 pixels
3)	London	512 pixels × 432 pixels

The images are shown in figures A.4, A.5 and A.6 respectively.

4.1 Recompression To The Same Quality Factor

This set of experiments was carried out on one of the images, Roses, to observe the accumulation of distortions when an image is repeatedly compressed to the same Q factor. When an image is compressed using the JPEG algorithm, some information, particularly the higher spatial frequency information, in the image is lost due to the irreversible nature of the quantisation process. An exception is when Q is chosen to be 100 in the free JPEG implementation, all the quantisation steps are set to unity, thus the DCT coefficients will not be quantised at all. In such a situation the decompressed image should be identical to the original image since theoretically the DCT is reversible. However, both the DCT and the inverse DCT (IDCT) are implemented with finite resolution, therefore round-off errors will be generated in the transform process and consequently the decompressed image is distorted. In situations where Q is less than 100, it is expected that the amount of information lost is greatest in the very first compression. This is because the compression process is similar to a low pass filtering operation, once the high frequency information has been removed, subsequent filtering is not likely to remove much further information. However, the distortion may still get accumulated due to round-off errors in the DCT and IDCT process.

Figure B.1 summarises the experimental results when the image Roses was compressed to a quality factor, decompressed, then compressed to the same quality factor and decompressed again, and so on for up to seven times. From the results, it can be seen that for higher Q factor, such as

Q=100 which corresponds to a bit rate of about 6.5 bits/pixel, and Q=95 which corresponds to about 4 bits/pixel, the SNR deteriorated significantly as the image went through the repeated compression, but it is also seen that the amount of deterioration began to saturate after about seven times of JPEG compression. For values of Q factor below 95, there was hardly any deterioration in the SNR even as the image got compressed repeatedly up to seven times.

In any case, in terms of visual perception, there was no discernible difference in the pictures that had gone through many times of compression and the one that went through only one time of compression. This was also verified from the fact that the maximum difference in the pixel grey-levels between the original image and the recompressed image only increased slightly as the number of repeated compression was increased. It is interesting to note that for high Q values such as 100 and 95, the image file size actually increased slightly as the image got repeatedly compressed. For smaller Q values, the file size fluctuated slightly and stabilised rather quickly.

The above observations are not further pursued in this study because in practical applications it is quite unlikely that a user needs to recompress an image to the same image quality.

4.2 Recompression To A Lower Quality Factor

All the three images were subjected to recompressions from an original quality factor Q1 to a lower quality factor Q2. Visual examinations of the one-time JPEG compressed images shows that the useful range of Q values is between 100 and 25. For Q factors below 25 the images became too blocky to be considered acceptable for normal viewing, although they might serve well as image search indices or icons. An assumption made in this study is that the value of Q1 is in the range of 40 to 80 in order to obtain reasonably sufficient compression and yet has potential for further file size reduction in a recompression. The values of Q2 were selected such that the difference between Q1 and Q2 was not less than 15, in order to justify the recompression effort in achieving sufficient file size reduction.

Two methods of recompression were used in this part of the study. The first method is simply to perform the compressions and decompressions in this sequence: compress, decompress, compress and decompress. This method is referred to as the *"D+C"* method, where D stands for decompress and C for compress. The second method involved modifying the free JPEG software by combining the decompression and the compression programs into a single recompression program called *rjpeg_norm*. It merely uncompresses a compressed file to obtain the DCT coefficients and requantises them accordingly as specified by the quality factor Q2. Therefore it eliminates the steps of IDCT and DCT operations in the D+C method as well as the round-off errors associated with these operations. This program takes a JFIF format file and recompresses it to another JFIF format file with the specified quality factor Q2.

For both methods of recompression, experiments were repeated on the three images for various combinations of Q1 and Q2 values, and measurements were taken for the compressed file size and the SNR. The variation of the SNR with the different combinations of Q1 and Q2 values are depicted in figures B.2 through B.7. It can be seen from the graphs that the variations in all three images follow a well-defined profile. Upon careful examination of the results it can be seen that the variation of the SNR is closely related to the ratio of the quantisation steps QS2/QS1, where QS1 is the DC coefficient quantisation step corresponding to the quantiser specified by Q1, and QS2 is the DC coefficient quantisation step corresponding to the quantiser specified by Q2. This observation is summarised in figure B.8, where a scatter graph of the SNR for the three images is plotted against QS2/QS1. From this graph it can be seen that the SNR is worst when QS2/QS1 is about two and best when it is about three.

In terms of visual perception, two types of artifacts were observed in these sets of recompressed images. The more severe of the two was the *grainy* effect which was most obvious in the relatively smooth regions in the images, this artifact appeared in the form of noise spots scattered all over in an area of the images. The other artifact appeared as a loss of sharpness in the images, which was less objectionable to the eye compared to the first type of artifact.

To arrive at an explanation for the observed artifacts, the mechanisms involved in a recompression were examined. During recompression, every DCT coefficient previously quantised in the first compression operation needs to be requantised using a second quantiser. Depending on the relative step size of the two quantisers, two types of errors may occur. The first type of error, referred to as a *positive error*, occurs when the requantised coefficient is larger in terms of magnitude than it would have been if directly quantised to quality Q2, this is illustrated in figure 4.

Suppose point A represents the unquantised DCT coefficient in the original image, after the first quantisation, its reconstruction value is r_1. Now if r_1 is subjected to requantisation specified by quality Q2, it will be quantised to 1. If the unquantised DCT coefficient A was directly quantised to the quality specified by Q2, it would have been quantised to zero. Due to this difference, the dequantised coefficient will have an error of QS2, or simply that the requantised coefficient has a positive error of unity.

The second type of error, referred to as a *negative error*, occurs when the requantised coefficient is smaller in terms of magnitude than it would have been if directly quantised to quality Q2. Again, referring to figure 4, if point C represents the unquantised DCT coefficient in the original image, its reconstruction value is r_2. Quantiser 2 will then requantise this value to 1, which upon dequantisation will become R_1. However, if coefficient C was directly quantised by quantiser 2, the quantised coefficient would be 2 and the dequantised value would be R_2. Hence there is a

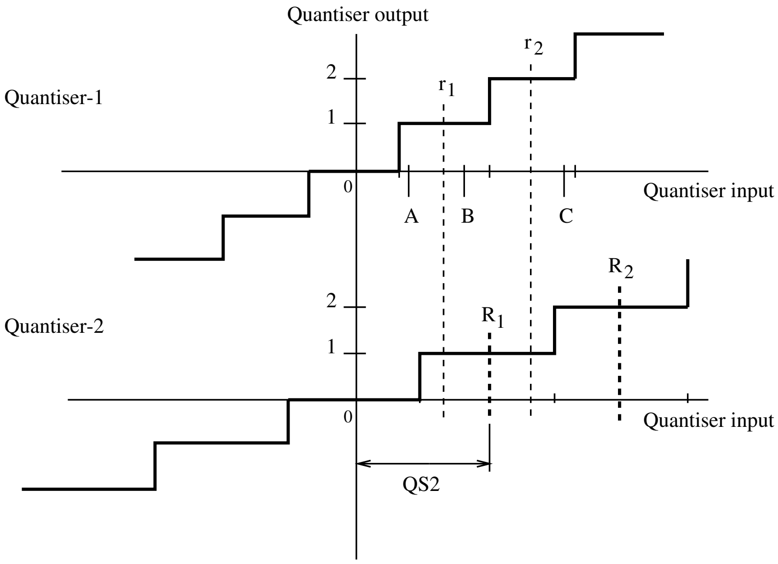

Figure 4: Positive and negative requantisation errors

negative error of unity in the requantised coefficient. It is worthwhile to note that the requantisation error, whether positive or negative, will have a magnitude of one. Therefore, if the location and the polarity of the error is known, a simple correction may be made to the coefficient to eliminate the error, the recompressed image will then be closer to the one obtained by a single step compression of quality Q2.

There are of course situations where there is no error in the requantised coefficient, this corresponds to, for example, the case when point B is the unquantised DCT coefficient in the original image.

To verify the above analysis, the percentages of positive and negative errors in the recompressed images obtained by the rjpeg_norm method were measured. A comparison between the variations of the percentage of positive errors and the variations of the SNR as a function of Q2 show that the two bore close relationship with each other, where there was a peak in the percentage of positive errors, there was a dip in the SNR, and vice-versa. This relationship is best summarised in figure B.9 where SNR is plotted against the percentage of positive errors in the recompressed image.

The percentage of negative errors, however, did not have similar effect on the SNR. In fact, in many cases the SNR was high when there was a high percentage of negative errors.

The variations of the percentages of positive and negative errors with the quantisation step ratio are illustrated in figures B.10 and B.11, respectively. It can be easily seen that the peak positive errors occurred when QS2/QS1 was around two and four; and the peak negative errors occurred when QS2/QS1 was slightly more than two. In general, the 64 DCT coefficients do not have the same quantisation step and therefore are not likely to have the same quantisation steps ratio QS2/QS1 as the DC coefficient. However, in the free JPEG software implementation, the 64 quantisation steps are linearly scaled by the quality factor, thus the AC coefficients will have quantisation steps ratios approximately equal to QS2/QS1.

In terms of compressed image file size, the results show that for the same value of Q2, it was possible to obtain different file sizes depending on the value of Q1, also in certain cases the SNR actually decreased with an increase in the file size, which was quite contrary to the normal rate-distortion relation. A careful investigation of the data reveals that the file size tends to vary directly with the percentage of positive errors, and vary inversly with the the percentage of negative errors in the recompressed image. This is probably because the positive errors have occurred in coefficients which should have been zero, and the negative errors have forced the requantised coefficients to zero. The former situation will worsen the efficiency of the run-length coding, whereas the latter will improve it, thus leading to the corresponding variations in compressed image file size.

The two types of visual artifacts observed in the recompressed images are directly related to the requantisation errors. In cases where there were high percentage of positive errors, the image appeared grainy, many fine spots were spread across the image, especially in the relatively smooth areas. Thus it may be concluded that the positive errors constitute an addition of noise to the image, and this noise is more objectionable when it is of a higher frequency than the components present in the original image. This artifact was usually seen to be accompanied by a poor SNR figure. In cases where there were high percentage of negative errors, the image lost its sharpness particularly around the edges, also certain areas of the image became *blocky*, that is, the area was seen to be made up of square blocks of pixels. The negative errors correspond to an attenuation in the component strength and are more objectionable to the eyes when they occur in the high frequency components.

However, the visual impact of the positive errors also depends on the contents of an image. In the case of image Roses, the grainy effect was most objectionable as it occurred mainly on the human face in the image, which was supposed to be smooth. For image Mandrill, the grainy effect and the addition of high frequency noise actually enhanced the rough texture and made the image appear sharper, even though the SNR performance was rather poor, and the errors only became noticeable when it was compared with the original image. In image London, the positive errors mainly appeared as spots of noise around the outline of the buildings. The images with substantial amount of such artifacts are shown in figures A.7, A.8 and A.9.

4.3 Estimation Of The Probabilities Of Errors

As explained in the previous section, the requantisation process introduces additional distortion to an image which undergoes a recompression operation. In the spatial frequency domain, these distortions are reflected in the DCT coefficients. In comparison with an image directly compressed from its original data, these DCT coefficients may be larger or smaller by one unit. If it is possible to know the locations of the positive and negative errors, a correction may be made to restore the image to be as closed as to the intended image. However, in the situation where the original image data is not available, it is not possible to determine exactly the coefficients in which the requantisation errors occur. It is therefore proposed to perform a calculation of the probabilities of the occurrence of positive and negative errors in the coefficients, subsequently these probabilities may be used for determining the correction to be made. A simple method has been implemented in this study to perform such an estimation.

4.3.1 Description of Method

First of all, given two uniform quantisers with known quantisation steps QS1 and QS2, the transition levels and the reconstruction levels of the quantisers can be calculated easily. Referring to figure 5, the transition levels of quantiser 1 and quantiser 2 are t_1, t_2,... and T_1, T_2,..., respectively; and the reconstruction levels are r_1, r_2,... and R_1, R_2,..., respectively.

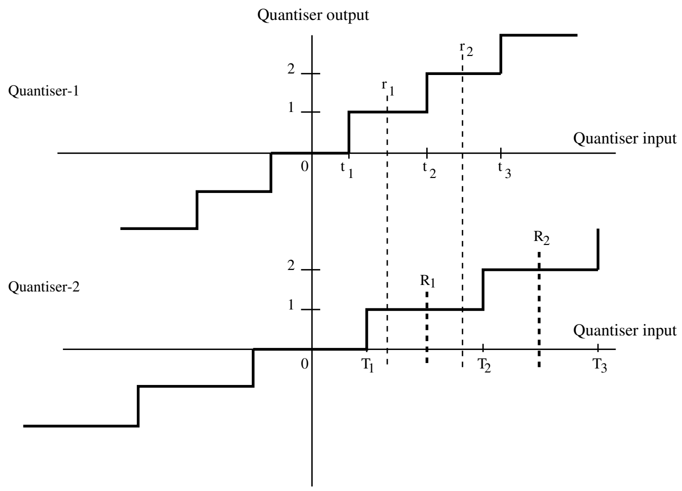

Figure 5: Estimation of the probabilities of positive and negative requantisation errors

It can be seen that, given that a dequantised coefficient has the reconstruction level r_1, the probability of the requantised coefficient being larger than it would have been had it gone through the second quantiser directly is given by:

$$\frac{T_1 - t_1}{t_2 - t_1}$$

and the probability of the requantised coefficient being smaller than it would have been is simply zero. These two probabilities are referred to as the probability of *coefficient enlargement* and the probability of *coefficient reduction*, respectively.

Similarly, if the dequantised coefficient has a value of r_2, the probability of coefficient enlargement is zero, and the probability of coefficient reduction is:

$$\frac{t_3 - T_2}{t_3 - t_2}$$

Therefore, for any given quantised DCT coefficient, if both the previous and the present quantisation step sizes are known, both the probability of coefficient enlargement and the probability of coefficient reduction can be calculated. It can be easily shown that the probability of coefficient enlargement is highest when the quantisation steps ratio is slightly lower or equal to two, in which case $(T_1 - t_1)$ is almost or equal to half of $(t_2 - t_1)$. The probability of coefficient reduction is highest when the quantisation steps ratio is slightly over two, in which case the probability is given by $\frac{t_2 - T_1}{t_2 - t_1}$, and $(t_2 - T_1)$ is almost half of $(t_2 - t_1)$. When the quantisation steps ratio is an exact odd integer, both the probability of coefficient enlargement and the probability of coefficient reduction are zero.

By going through the full range of the quantised coefficient values, a table of probability of coefficient enlargement can be constructed, this table is indexed by the quantised coefficient value. A similar table can be computed for the probability of coefficient reduction. It is noted that, due to the symmetry of the uniform quantiser, the probabilities of coefficient enlargement are the same for two coefficients of opposite polarity (positive and negative) as long as they share the same magnitude. The same may be stated for the probability of coefficient reduction.

As pointed out earlier, each of the 64 DCT coefficients has a different set of quantisation steps, the 63 AC coefficients' quantisation steps ratios are only approximately equal to that of the DC coefficient. Theoretically a set of the tables of probability of coefficient enlargement and reduction may be calculated for each of the 64 coefficients, but that would involve too much computing as well as table indexing effort. Thus the probabilities of coefficient enlargement and reduction for the AC coefficients are approximated by those of the DC coefficient.

The purpose of constructing these tables is two-fold. First of all, in the process of recompression, if it is known that the coefficient has a sufficiently high probability of enlargement, the error may be compensated by suppressing the requantised coefficient by one unit. Similarly, if the coefficient has a sufficiently high probability of being reduced, the error may be compensated by increasing the magnitude of the requantised coefficient by one unit. However, since the method is based on probability, there will be times that a suppression or an addition is done where there is no error at all. Therefore this method is very sensitive to the proper selection of a threshold against which the probabilities of enlargement and reduction are compared.

Secondly, with the table of probability of coefficient enlargement, it is possible to calculate the expectation of the number of positive errors in the recompressed image if the probability density function of the quantised DCT coefficients is known. Similarly the expectation of the percentage of negative errors may be calculated. These two estimated values serve to predict the amount of distortions to be introduced by the recompression operation, a decision may thus be made to select the appropriate correction for the entire image, or to leave out the correction totally if the errors are likely to be small.

Graphical results of the estimated probabilities have been obtained for the three images at various levels of compression as specified by the quality factors and are included in figures B.12 through B.23. A graph to illustrate the probability density function of the quantised DCT coefficient values is in figure B.24. From these graphs, it can be seen that in most cases the estimated percentages of positive and negative errors follow very closely to the profiles of the actual errors, even though there are some differences in the actual values.

5 Experimental Techniques To Overcome The Recompression Artifacts

5.1 General Approach

A few methods were experimented to reduce some of the artifacts produced by a recompression. The main objective of these methods is to suppress the amount of high frequency noise generated from the positive errors. The loss of sharpness caused by the negative errors may be dealt with in a similar manner but is not treated at the present moment, partly because this may be overcome by recompressing the image to a not-too-low quality factor. A probabilistic approach has been taken to reduce this type of high frequency noise which create grainy effects in the picture, this is achieved by predicting the locations of the positive errors and suppressing the corresponding coefficients.

The main difference between the following methods lies in the mechanism used for determining whether a coefficient needs to be suppressed. Once a decision of suppression is made, there are two methods to perform the actual suppression which are described below.

5.1.1 Suppression Method A

This method makes use of the fact that the positive errors are most severe in the case when the dequantised DCT coefficient lies exactly on the transition level of the new quantiser, this is illustrated in figure 6.

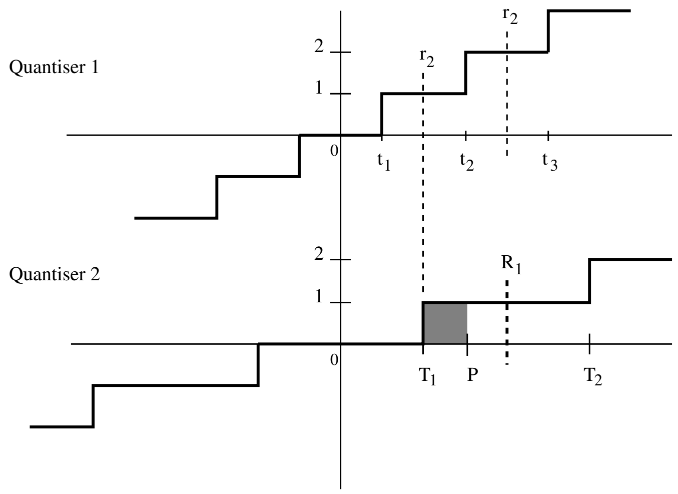

Figure 6: Coefficient suppression method A

It can be seen that half of the dequantised coefficients lying between t_1 and t_2 will suffer an enlargement in the requantisation process. In such a case, the usual requantisation operation implementation:

temp = dequantised_coefficient + (new_quantisation_step/2)

requantised_coefficient = temp/new_quantisation_step

will cause the magnitude of the requantised DCT coefficient to be larger than it would have been by exactly one unit. To alleviate this problem, a simple suppression scheme is implemented by modifying the requantisation operation to:

temp = dequantised_coefficient + (new_quantisation_step/4)

requantised_coefficient = temp/quantisation_step

In this case, dequantised DCT coefficients which lie in the shaded area between the transition level T_1 and the quarter mark P (figure 6, pg. 15) will be suppressed in the requantisation process, where $(P - T_1) = (T_2 - T_1)/4$.

5.1.2 Suppression Method B

This method makes use of the fact that any requantisation error, if present, will either increase the magnitude of the coefficient by one unit (in the case of a positive error) or decrease the magnitude by one unit (in the case of a negative error). Here, the positive errors are the issue of concern, so when it is decided that a coefficient needs to be suppressed, the requantisation operation becomes this:

temp = dequantised_coefficient + (new_quantisation_step/2)
requantised_coefficient = temp/new_quantisation_step
if (|requantised_coefficient| != 0)
    decrement |requantised_coefficient| by 1 unit

where *|requantised_coefficient|* denotes the magnitude or absolute value of the requantised coefficient.

The main difference between method A and method B is: In method A, even when a coefficient is subjected to the suppression operation, it is only actually suppressed if its dequantised value lies within a distance of (quantisation step/4) from the second quantiser's transition level, for example in the shaded region depicted in figure 6. Thus, this method depends on the actual value of the coefficient concerned. Whereas in method B, once it is decided that the coefficient is to be suppressed, the magnitude of its requantised value will be decreased by one unit, subject to the lower bound of zero. Therefore, given the same decision criterion for suppression, it is more likely that additional negative errors are introduced when method B is used instead of method A. The methods experimented so far are described below.

5.2 Simple Adaptive Method - Method I

This is a simple method that attempts to suppress DCT coefficients enlargement in the higher frequency components.

5.2.1 Description Of The Method

For each block of 8x8 quantised DCT coefficients in a given compressed image file, the total number of non-zero DCT coefficients in the zig-zag scan path bounded by the coefficient numbers LOW_LIMIT (inclusive) and UPP_LIMIT (exclusive) is computed. For instance, if LOW_LIMIT = 10 and UPP_LIMIT = 15, then the coefficients numbered 10, 11, 12, 13 and 14 (as in figure 3, pg. 5) are checked for non-zero coefficient values. This number is used as a crude indicator to the amount of information in that particular block. Typically, a pixel block which contains substantial amount of information has more non- zero high frequency components than a block which is relatively flat; also, the former is more likely to have non-zero coefficients in a band defined by higher values of LOW_LIMIT and UPP_LIMIT (which correspond to a higher spatial frequency band) than the latter pixel block. If the number of non-zero coefficients is more than a number THRESHOLD, a higher value is selected for FREQ_CUT, otherwise a lower value is selected for FREQ_CUT.

The DCT coefficients are then processed in the zig-zag order as depicted in figure 2 (pg. 4), for components below and equal to FREQ_CUT, the normal requantisation is carried out. For components above FREQ_CUT, the suppression method A described in 5.1.1 is used. To illustrate, if FREQ_CUT = 6, then the coefficients numbered from 0 to 6 (as in figure 3, pg. 5) are quantised as per normal, whereas the coefficients numbered from 7 onwards will go through the suppression process.

5.2.2 Experimental Results

An experiment was carried out on the image Roses with the following selection of parameter values:

LOW_LIMIT = 10, UPP_LIMIT = 15 , THRESHOLD = 2

FREQ_CUT = 6 or 10.

This method, due to its simplicity, was not expected to produce very good results but was used mainly to observe the effect of coefficient suppression on a recompressed image. The results show that, in recompressing the image Roses from various values of Q1 to the value of 25 for Q2, the compressed file sizes were smaller than those obtained either from the D+C or the rjpeg_norm recompression operation. The reduction in bit rate could be up to 0.24 bit/pixel. In two cases, where Q1=50 and Q1=45, the SNR increased by 0.9 and 0.7 dB respectively, even though there were reductions in the bit rate. Also, visually the images did not have the grainy effects observed in the D+C and rjpeg_norm recompressed images, although they did suffer from some loss of sharpness. These suggested it is possible to remove such artifacts without having to sacrifice the bit rate, and better results may be obtained by more refined method of coefficient suppression.

5.3 Adaptive Method With Six Frequency Bands - Method II-A and Method II-B

5.3.1 Description Of The Methods

These methods are similar to, but slightly more refined than method I, the total number of non-zero DCT coefficients in each block is compared to six thresholds to categorise the block's degree of business into one of the six bands, then a corresponding value for FREQ_CUT is selected. In method

II-A, the suppression method A is used, whereas in method II-B, the suppression method B is used.

5.3.2 Experimental Results

The experiments based on method II-A were carried out on the image Roses with the following sets of parameter values:

THRESHOLD[NUM_OF_BANDS] = 4, 8, 12, 16, 25, 64

FREQ_CUT[NUM_OF_BANDS] = 1, 3, 6, 10, 15, 21

In recompressing versions of the image Roses of various values of Q1 to the value of 25 for Q2, the results show that in comparison with method I, the SNR in all cases improved as less negative errors were introduced in the recompression process. The file sizes also increased except in two cases: Q1=40, Q2=25 and Q1=45, Q2=25. Comparing to the compressed files obtained from the D+C and rjpeg_norm recompression methods, the compressed file sizes obtained in method II-A were still smaller. The more important results were in the cases where:

- Q1=50, Q2=25: the SNR was 1.02 dB better than the D+C recompressed image, and the bit rate was smaller by 0.178 bit/pixel.
- Q1=45, Q2=25: the SNR was 0.83 dB better, and the bit rate was smaller by 0.265 bit/pixel.
- Q1=40, Q2=25: the SNRs were the same, but the bit rate was smaller by 0.237 bit/pixel.

In all three cases, the images obtained using method II-A did not exhibit the grainy effects present in the images obtained by the D+C and the rjpeg_norm methods. In some of the other cases, the SNR actually deteriorated compared to those obtained by the D+C and rjpeg_norm recompression methods, but the amounts of deterioration were less than 0.5 dB.

A similar set of experiments was carried out on the same images employing method II-B. The results obtained here are worse than those using method II-A, mainly because the coefficients above FREQ_CUT were suppressed indiscriminately, resulting in additional negative errors without much further elimination of the positive errors. Along with a decrease in compressed file size, the SNR values also declined. Visually the images appeared significantly less sharp than those obtained by the D+C and the rjpeg_norm methods.

5.4 Adaptive Method Using Probability Of Coefficient Enlargement - Method III

5.4.1 Description Of The Method

This method uses a table of probability of enlarging a DCT coefficient computed from the two quantisation steps, as described in 4.3. The six-band classification of block business mentioned in

5.3.1 is used to first select the FREQ_CUT. For components below or equal to FREQ_CUT in the zig-zag ordering, no special processing is made. For components above FREQ_CUT, the coefficient is checked against the table of probability of coefficient enlargement. If the probability is greater than PROB_LIMIT, the coefficient is suppressed using method B.

5.4.2 Experimental Results

Experiments similar to those described in the previous methods were carried out on the image Roses, with the following set of parameters:

THRESHOLD[NUM_OF_BANDS] = 4, 8, 12, 16, 25, 64
FREQ_CUT[NUM_OF_BANDS] = 1, 3, 6, 10, 15, 21
PROB_LIMIT = 0.24

The results obtained from this method are very similar to those obtained in method II-A. The file sizes tended to be smaller in this method but the differences were slight. This method was very effective in reducing the amount of positive errors particularly in the cases where Q2 was 75, 50, 45 and 40. However, it generated more negative errors and caused the image to appear less sharp. The SNR figures also dropped because of these additional negative errors.

5.5 Detection Of Highest Frequency Components - Method IV-A

5.5.1 Description Of The Method

In this method, a simple edge-detection algorithm is used to locate the highest spatial frequency components in a block of dequantised DCT coefficients, as shown in figure 7 (pg. 20). The highest spatial frequency components in the block are then suppressed in the recompression process by using method A. The main reason for doing this is to concentrate on suppressing only the unwanted noise in the highest frequency band as they are more objectionable. This also helps to reduce the introduction of negative errors in the DCT coefficients by the suppression method, which tend to worsen the SNR values as well as make the image lose significant sharpness.

5.5.2 Experimental Results

The image Roses were recompressed from various values of Q1 to the value of 25 for Q2. Comparing with the results obtained in the D+C and rjpeg_norm methods, the most significant improvements occurred in the cases of Q1=50 and Q1=45, where the SNR improved by 1.02 dB and 0.86 dB respectively, and the bit rate was reduced by 0.16 and 0.24 bit/pixel respectively. Visually the images also appeared much better as the grainy artifacts had been removed.

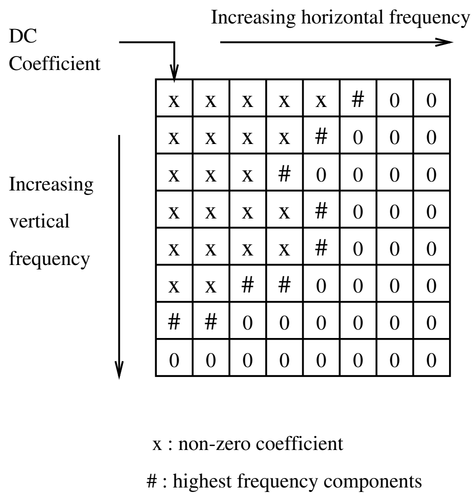

Figure 7: Highest frequency components in a DCT block

5.6 Probability Of Coefficient Enlargement Method - Method IV-B

5.6.1 Description of the method

This method is similar to method IV-A with the exception that each of the highest frequency coefficient is checked against the table of probability of coefficient enlargement to determine if it needs to be suppressed using method B.

5.6.2 Experimental Results

A more elaborate set of experiments were conducted on the images Roses, Mandrill and London using method IV-B. Firstly, the image Roses was recompressed from various values of Q1(in steps of 5 in the range from 80 to 40) down to the values of 25, 35 and 45 for Q2. The recompressions were repeated with different values of PROB_LIMIT between 0.0 and 0.30 to observe the effects on the recompressed images. The results obtained show that if PROB_LIMIT was too low, a large amount of negative errors were generated even though the positive errors were effectively suppressed, resulting in images which lost significant sharpness. On the other hand, if PROB_LIMIT was too high, there was not sufficient suppression of the high frequency noise and the images appeared grainy. The results obtained for values of PROB_LIMIT = 0.20, 0.24, 0.27 and 0.30 did not differ much, thus the value of PROB_LIMIT was set to 0.24 in the subsequent experiments. The SNR results for the image Roses are graphically presented in figures B.25 through B.27, for the image Mandrill they are in figures B.28 through B.30, and for the image London they are in figures B.31

through B.33.

From these three sets of graphs, it can be clearly seen that the variations of the SNR with the various combinations of Q1,Q2 were similar in all three images. It is also observed that the significant improvements in SNR occurred when the ratio of the quantisation steps QS2/QS1 was approximately 2.0, as reflected in the following cases:

Q1 = 50, Q2 = 25, QS2/QS1 = 2.0

Q1 = 45, Q2 = 25, QS2/QS1 = 1.78

Q1 = 60, Q2 = 35, QS2/QS1 = 1.77

Q1 = 70, Q2 = 45, QS2/QS1 = 1.80

This is also consistent with the results depicted in figure B.10 where the percentage of positive errors in recompression were highest when QS2/QS1 was slightly less than or equal to 2.0, which was expected since the method that had been experimented with was designed to reduce the positive errors. However, the method was not effective where the amount of positive errors were less significant. In terms of recompressed image file size, in almost all cases the file sizes were smaller than those obtained by the D+C or the rjpeg_norm method, the difference in bit rate could be as high as 3.60 bits/pixel. Thus the recompression technique did not only provide an improvement in the SNR over the straight forward D+C or the rjpeg_norm method, it also produced a recompressed image of smaller file size. This is best summarised in tables 1 through 3 where the improvement in SNR and bit rate of the three images recompressed by method IV-B over those recompressed by the D+C and the rjpeg_norm methods are shown.

From the visual perception point of view, the recompression technique had successfully removed the grainy artifacts in the recompressed images in most cases. However, a problem remains that the images were not as sharp as it should have been if it had only gone through a single step JPEG compression, afterall the coefficient suppression was based on a probabilistic decision and in many cases it redistributed the positive errors to negative errors. This may be improved by sacrificing the bit rate, that is, only recompress the image by a small degree so that not too much additional loss is incurred. Alternatively, a similar scheme to coefficient suppression, but designed to expand rather than suppress the DCT coefficients, may be added to overcome some of the negative errors introduced in the recompression process.

Comparing the results obtained by method I, II-A, II-B, III, IV-A and IV-B, in terms of SNR performance, method II-A was significantly better than method I, method I was efficient in removing the positive errors but tended to introduce excesssive negative errors. Method II-A also produced better SNR results than method II-B, which was even worse than method I in terms of producing negative errors. Method III performed slightly worse in some cases than method II-A, due to the fact that it was capable of generating large negative errors in the recompression process, otherwise the results were similar. Method IV-B produced slightly better SNR results than IV-A, which in turn produced slightly better SNR results than II-A. The images obtained by this method and have Q1=50, Q2= 25 are shown in figures A.10 through A.12. These three methods performed similarly in terms of redistributing the positive errors to negative errors.

6 Conclusions and Further Work

The above experimental results show that it is possible to reduce the artifacts in the form of grainy effects in recompressed image by performing some coefficient suppression. However, the major problem lies in achieving a good balance of coefficient suppression and normal processing in order that the image does not lose significant sharpness. In the case of method IV-B recompression, this may be achieved by first investigating the probabilities of coefficient enlargement and reduction before selecting an appropriate value for PROB_LIMIT. The loss of sharpness may also be overcome by including a scheme to expand the coefficients in the presence of negative errors. The results also show that the artifacts are worse in some cases than others, therefore if we can in some way determine whether special processing is required to perform a particular recompression, for example by first estimating the percentage of positive and negative errors introduced, then the additional processing efforts in recompression can be saved and at the same time unneccesary loss of sharpness is avoided.

The estimated percentages of positive and negative errors, though in most cases correlate well with the actual amount of errors, do show some discrepancies in the profile in a few cases, it would be of interest to investigate into the cause of such discrepancies. In estimating the percentages of positive and negative errors, the probability density function of the quantised DCT coefficients are computed from the compressed images. As there has been a previous study on the distribution of DCT coefficients [11] it may be worthwhile performing the estimation based on the statistical model of the quantised DCT coefficients probability density function to save the computation efforts. So far the study has only investigated the problems of subjecting a compressed image to a second time compression, in the next stage of the study, the images will be subjected to a multiple number of repeated compression to further explore the problems and solutions.

References

[1] JPEG Committee Draft, "Digital Compression and Coding of Continuous-tone Still Images; Part 1 : Requirements and Guidelines," Reference no. ISO/IEC JTC 1 /SC 2, March 1991.

[2] Hudson, G.P., et al, "The International Standardisation of a Still Picture Compression Technique," *Globecom '88 - Florida*, Vol. II, pp. 1016-1021.

[3] Wallace, G., et al, "Subjective Testing Results for Still Picture Compression Algorithms for International Standardization," *Globecom '88 - Florida*, Vol. II, pp. 1022-1027.

[4] Leger, A., et al, "Still Picture Compression Algorithms Evaluated for International Standardisation," *Globecom '88 - Florida*, Vol. II, pp. 1028-1032.

[5] Wallace, G., "The JPEG Still Picture Compression Standard," *CACM*, Vol. 34, No. 4, pp. 31-44, April 1991.

[6] Hung, A., "Image Compression: The Emerging Standard for Color Images," *Computing Futures*, 1989.

[7] Clarke, R.J., *Transform Coding of Images*, Academic Press, London, 1985, ch. 3.

[8] Jayant, N.S. and Noll, P., *Digital Coding of Waveform: Principles and Applications to Speech and Video*, Prentice Hall, New Jersey, 1984, ch. 10.

[9] Huffman, D.A., "A Method for the Construction of Minimum Redundancy Codes," *Proceedings of the IRE*, Vol. 40, pp. 1098-1101, Sep. 1952.

[10] Jayant, N.S. and Noll, P., *Digital Coding of Waveform: Principles and Applications to Speech and Video*, Prentice Hall, New Jersey, 1984, ch. 6.

[11] Reininger, R.C. and Gibson, J.D., "Distributions of the Two-Dimensional DCT Coefficients for Images," *IEEE Transactions on Communications*, Vol. COM-31, No. 6, pp. 835-839, June 1983.

Table 1: Improvement in SNR and bit rate using method IV-B

Roses		Improvement in SNR (dB)		Improvement in bit rate (bit/pixel)	
Q1	Q2	Over D+C	Over rjpeg_norm	Over D+C	Over rjpeg_norm
50	25	1.02	1.56	0.160	0.266
45	25	0.95	0.90	0.230	0.244
60	35	0.96	0.91	0.283	0.298
70	45	1.27	1.30	0.316	0.360

Table 2: Improvement in SNR and bit rate using method IV-B

Mandrill		Improvement in SNR (dB)		Improvement in bit rate (bit/pixel)	
Q1	Q2	Over D+C	Over rjpeg_norm	Over D+C	Over rjpeg_norm
50	25	1.65	2.12	0.248	0.317
45	25	1.26	1.26	0.295	0.296
60	35	1.28	1.28	0.323	0.324
70	45	1.70	1.71	0.340	0.348

Table 3: Improvement in SNR and bit rate using method IV-B

London		Improvement in SNR (dB)		Improvement in bit rate (bit/pixel)	
Q1	Q2	Over D+C	Over rjpeg_norm	Over D+C	Over rjpeg_norm
50	25	1.66	2.25	0.166	0.224
45	25	1.41	1.41	0.207	0.207
60	35	1.40	1.40	0.237	0.238
70	45	1.66	1.68	0.252	0.269

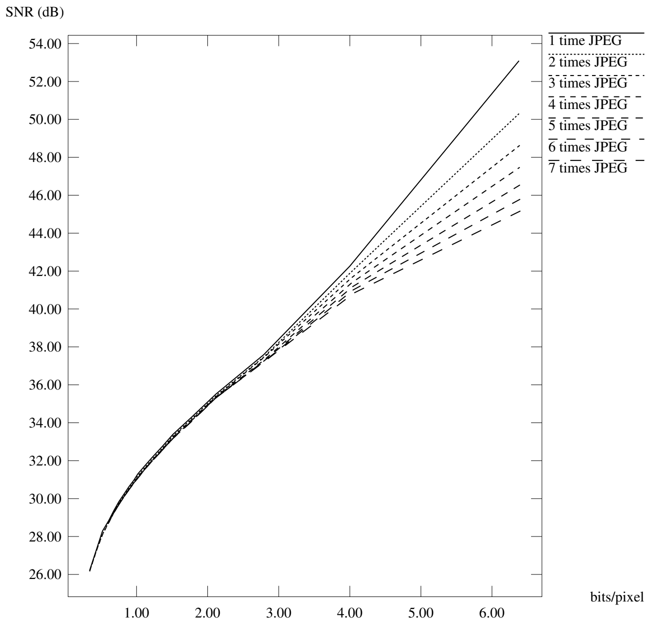

Figure B.1: SNR versus bit rate for image Roses when recompressed to same Q-factor

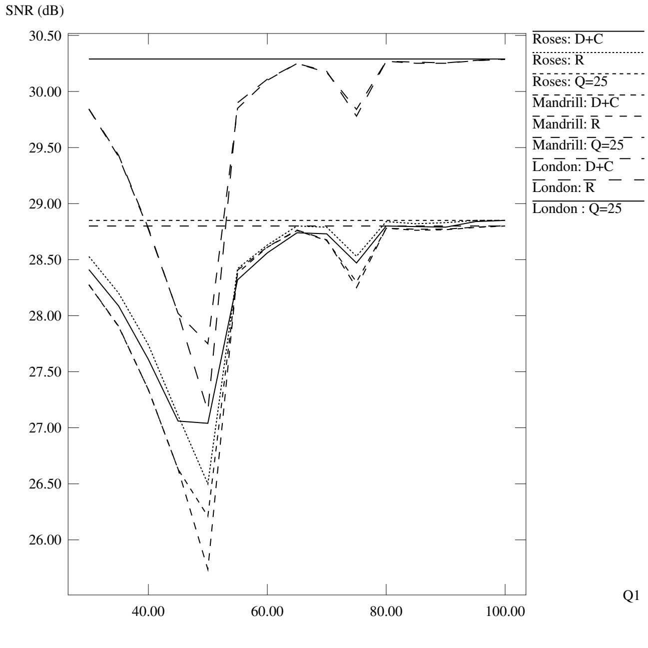

Figure B.2: SNR of recompressed images for Q2 = 25

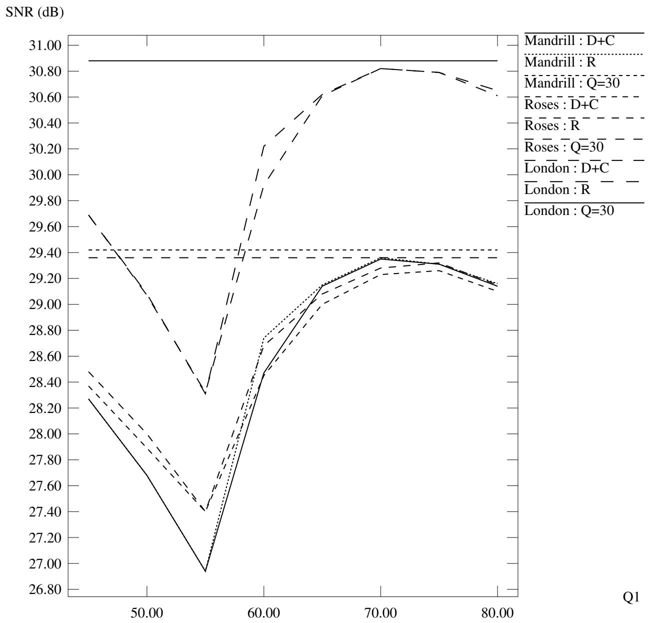

Figure B.3: SNR of recompressed images for Q2 = 30

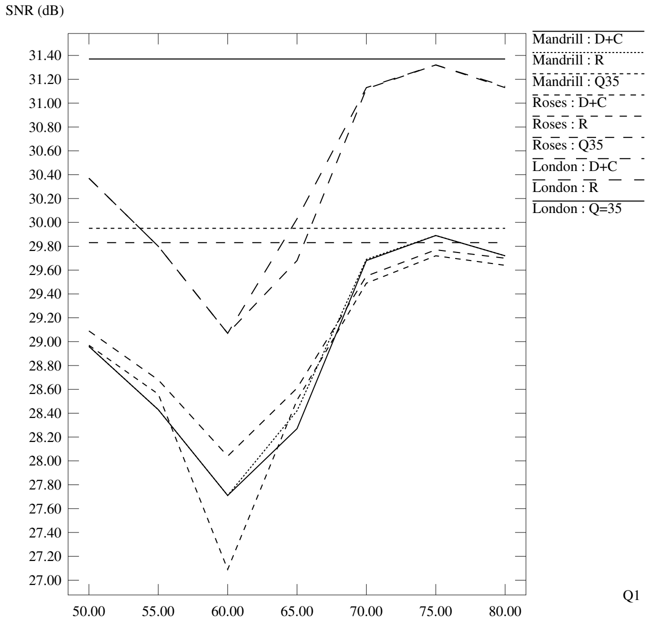

Figure B.4: SNR of recompressed images for Q2 = 35

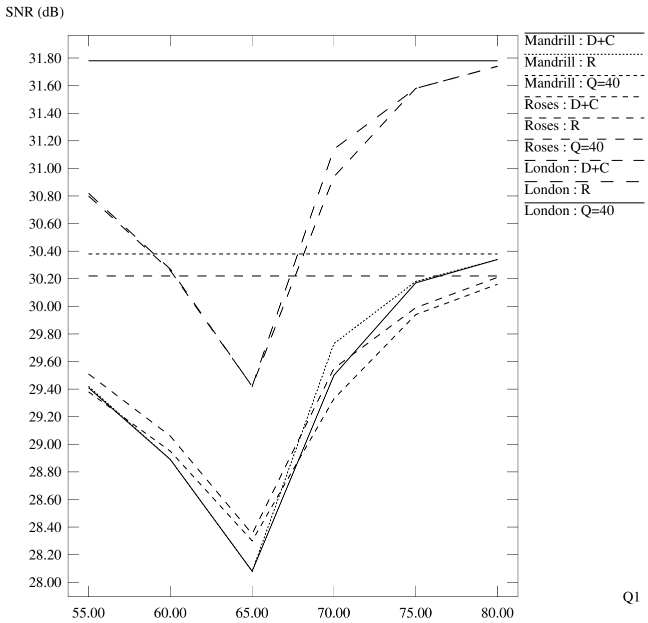

Figure B.5: SNR of recompressed images for Q2 = 40

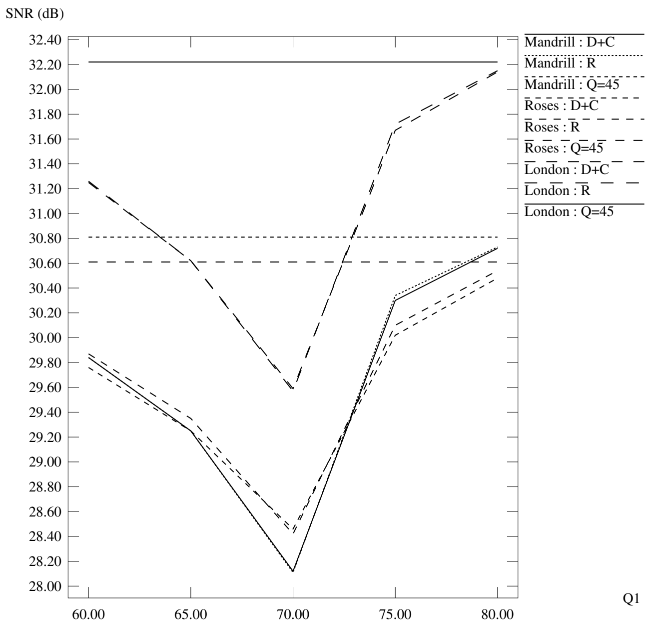

Figure B.6: SNR of recompressed images for Q2 = 45

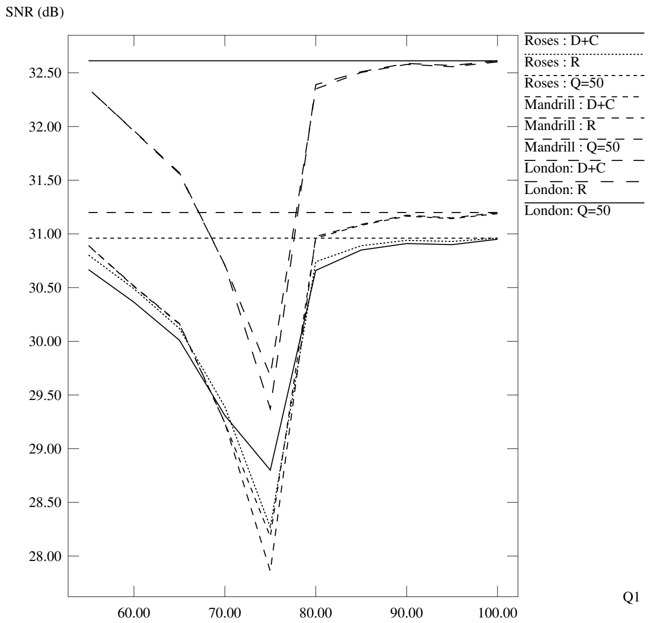

Figure B.7: SNR of recompressed images for Q2 = 50

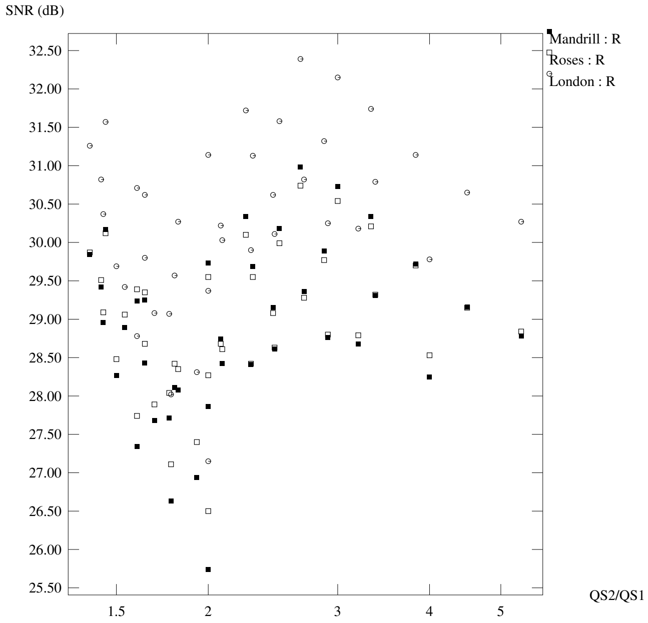

Figure B.8: SNR of recompressed images versus quantisation step ratio

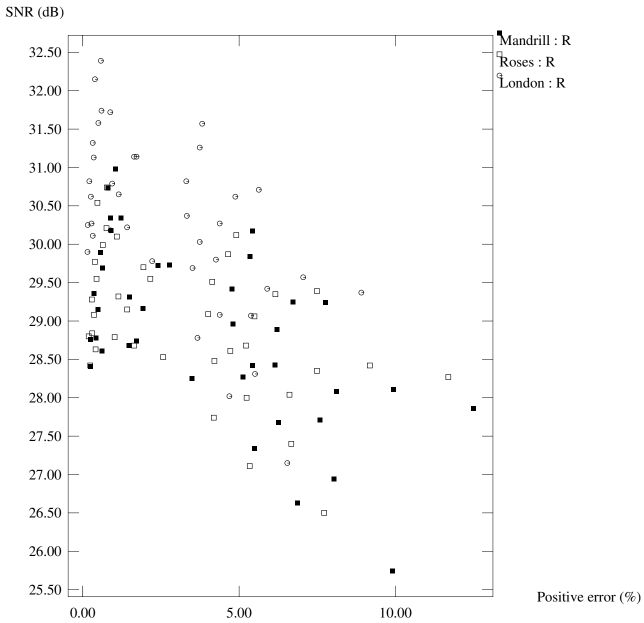

Figure B.9: SNR versus percentage positive error in recompression

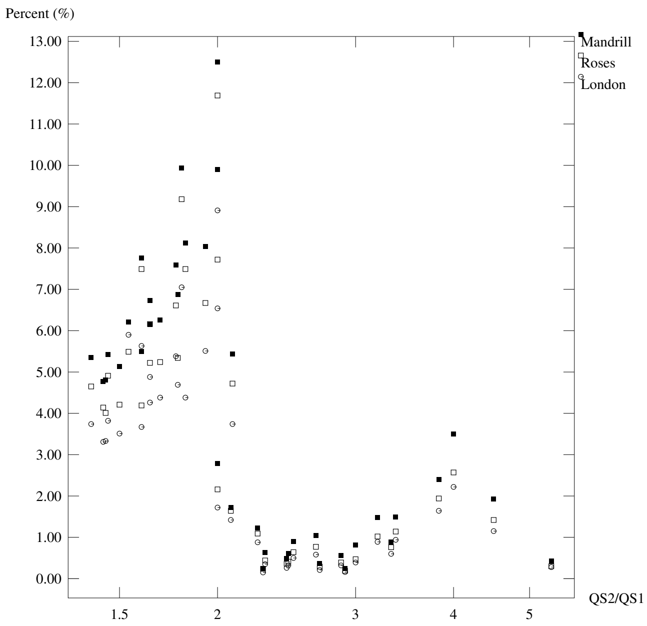

Figure B.10: Percentage positive error in recompression versus quantisation step ratio

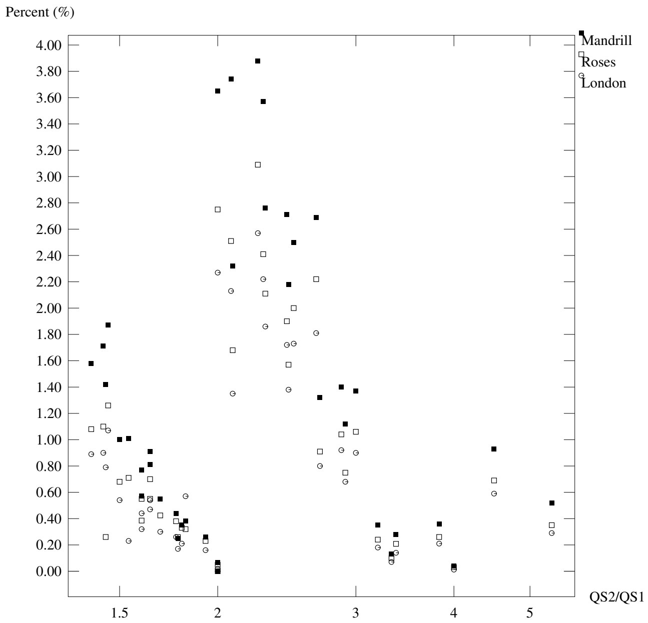

Figure B.11: Percentage negative error in recompression versus quantisation step ratio

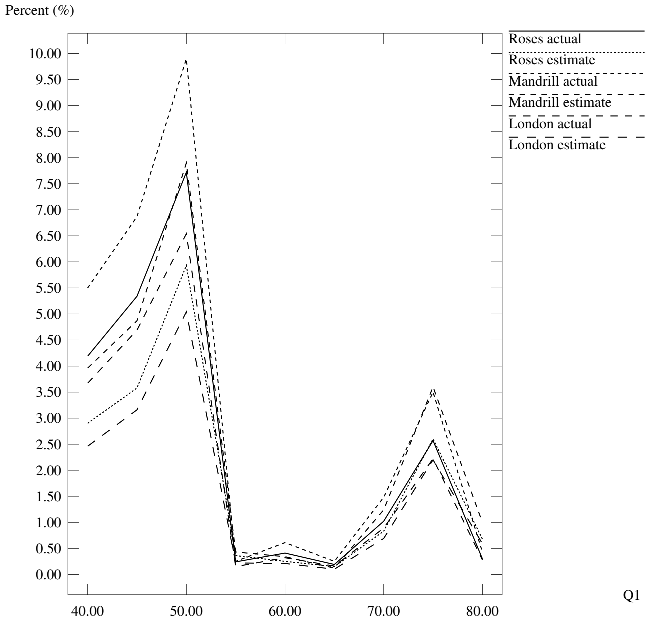

Figure B.12: Estimated and actual percentage positive error in recompression for Q2 = 25

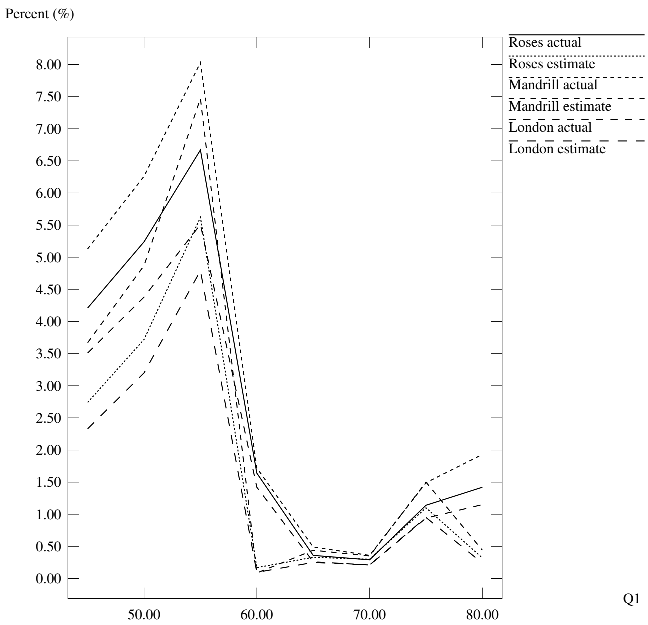

Figure B.13: Estimated and actual percentage positive error in recompression for Q2 = 30

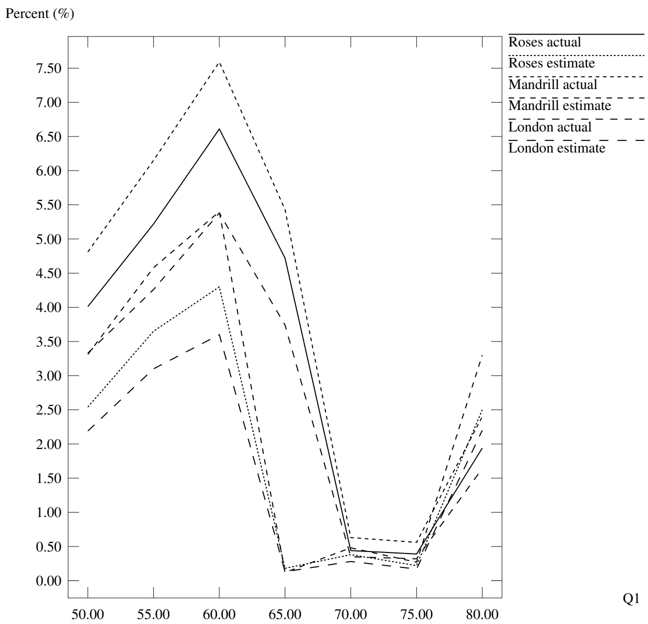

Figure B.14: Estimated and actual percentage positive error in recompression for Q2 = 35

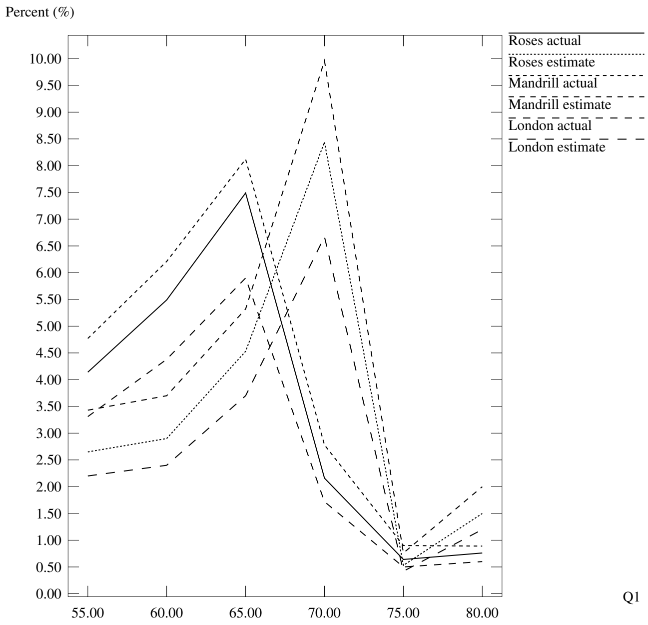

Figure B.15: Estimated and actual percentage positive error in recompression for Q2 = 40

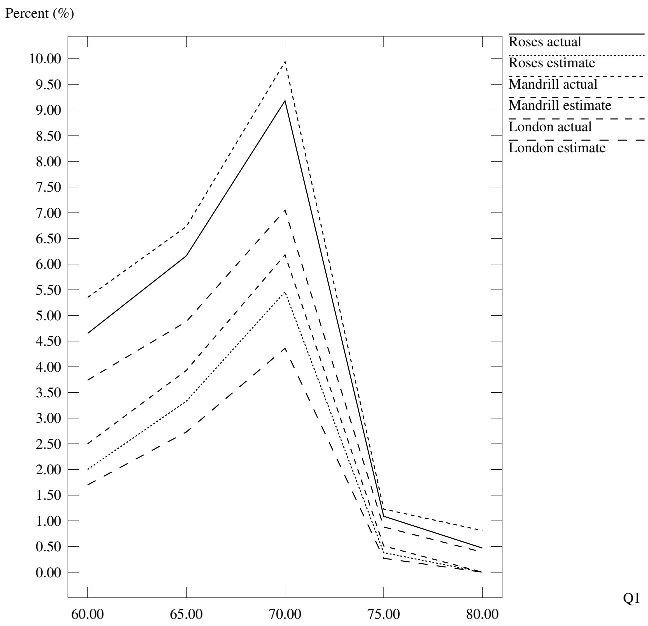

Figure B.16: Estimated and actual percentage positive error in recompression for Q2 = 45

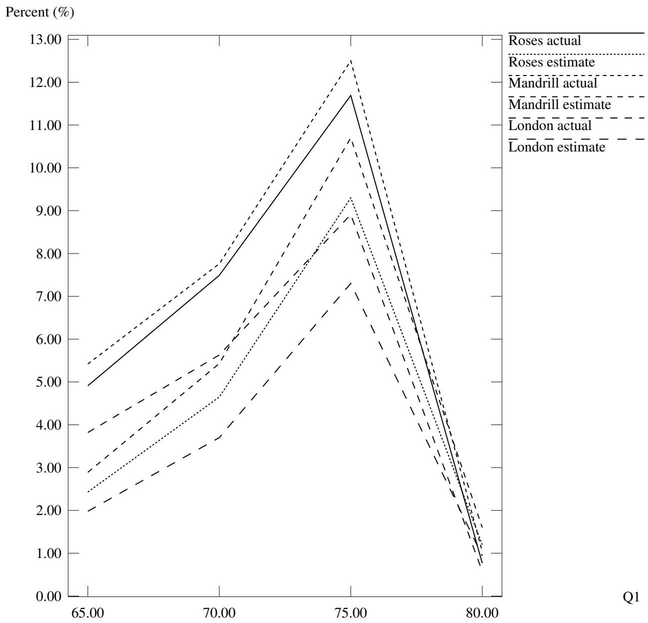

Figure B.17: Estimated and actual percentage positive error in recompression for Q2 = 50

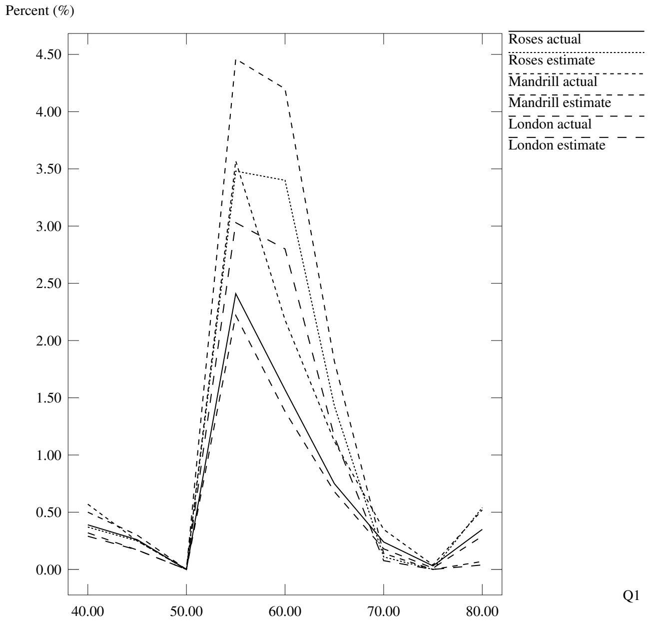

Figure B.18: Estimated and actual percentage negative error in recompression for Q2 = 25

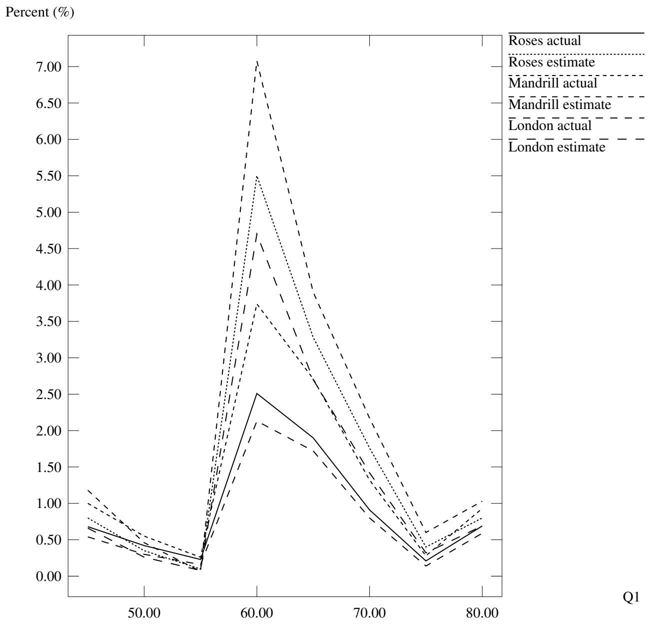

Figure B.19: Estimated and actual percentage negative error in recompression for Q2 = 30

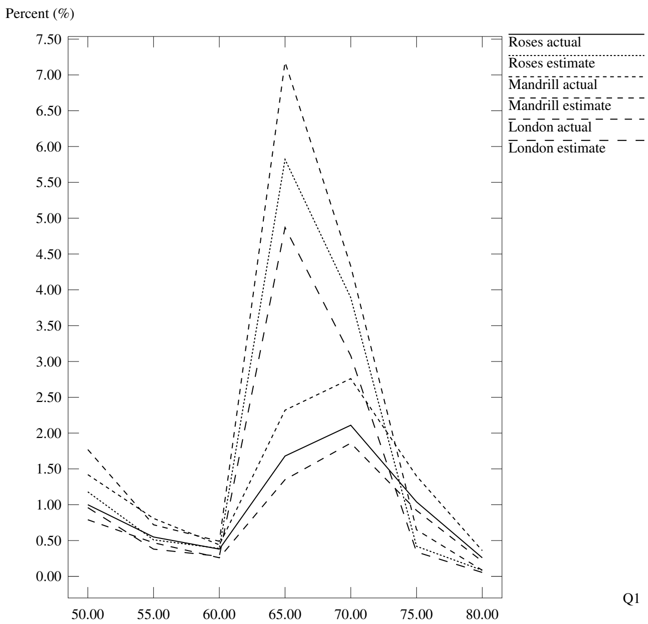

Figure B.20: Estimated and actual percentage negative error in recompression for Q2 = 35

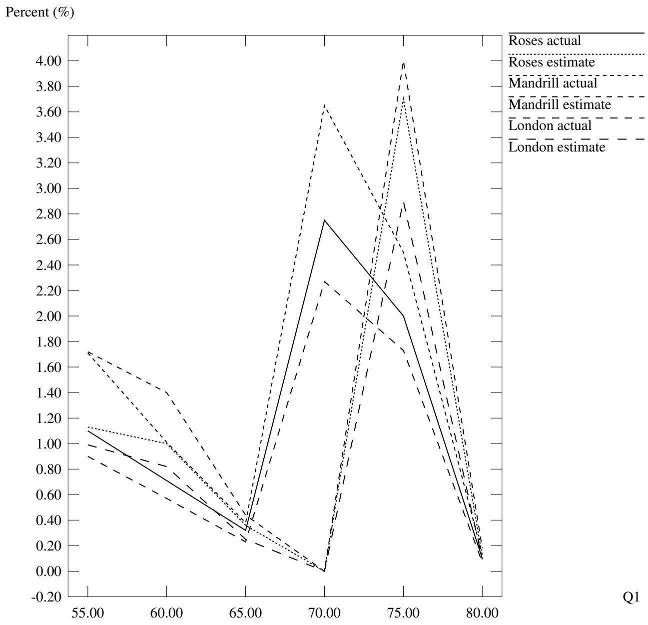

Figure B.21: Estimated and actual percentage negative error in recompression for Q2 = 40

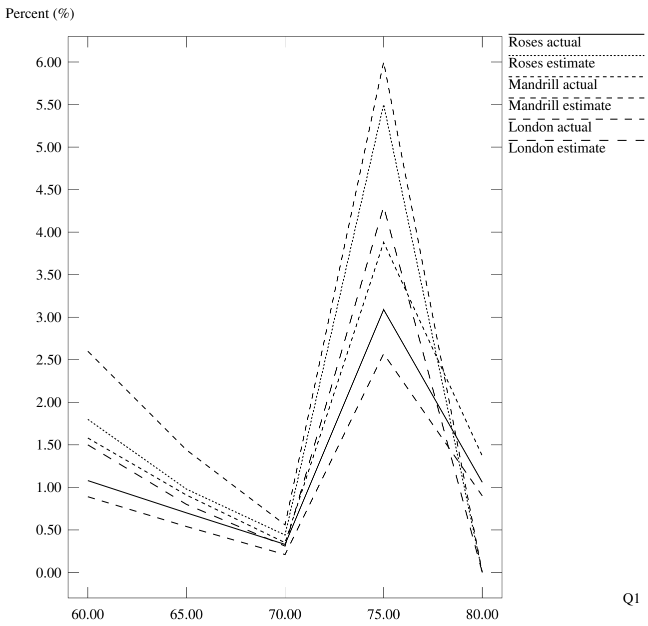

Figure B.22: Estimated and actual percentage negative error in recompression for Q2 = 45

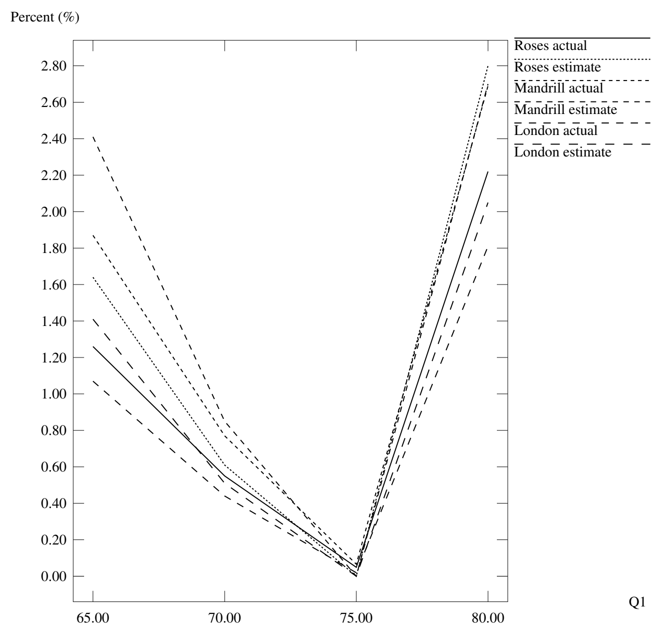

Figure B.23: Estimated and actual percentage negative error in recompression for Q2 = 50

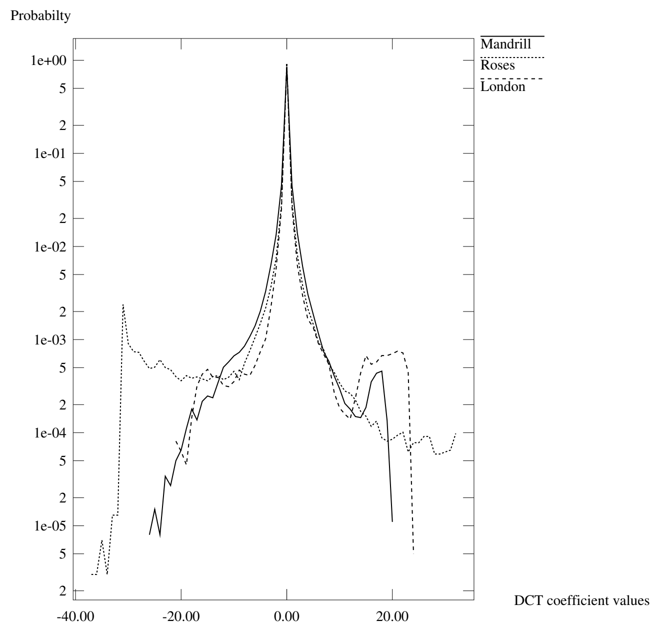

Figure B.24: Probability density function of quantised DCT coefficients, Q = 25

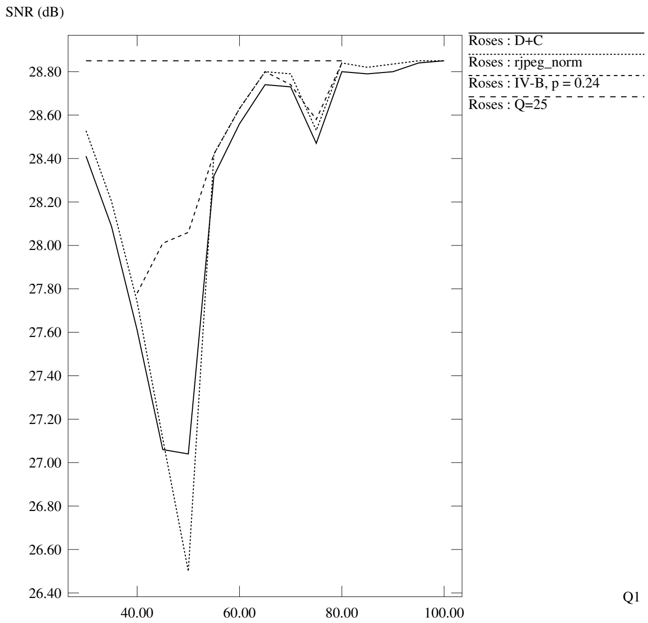

Figure B.25: SNR of recompressed image Roses for Q2 = 25

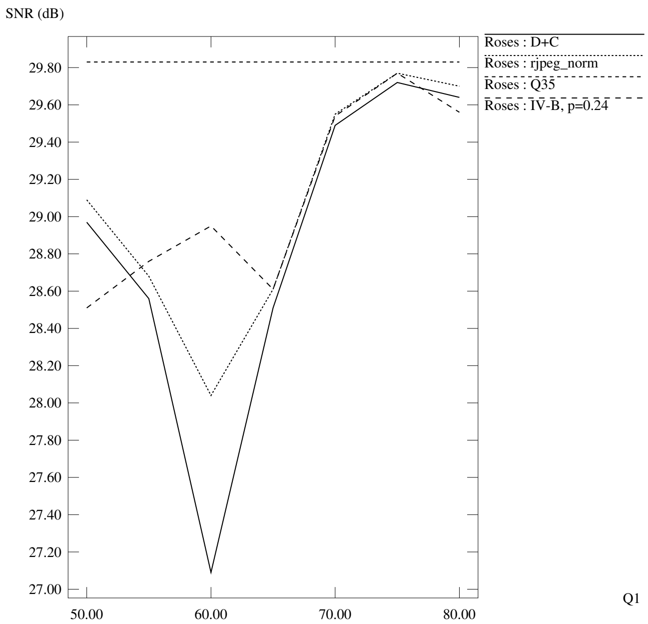

Figure B.26: SNR of recompressed image Roses for Q2 = 35

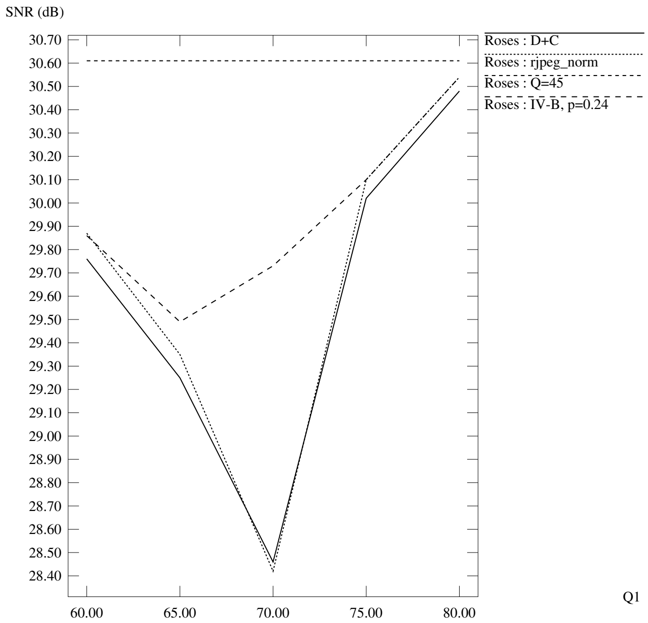

Figure B.27: SNR of recompressed image Roses for Q2 = 45

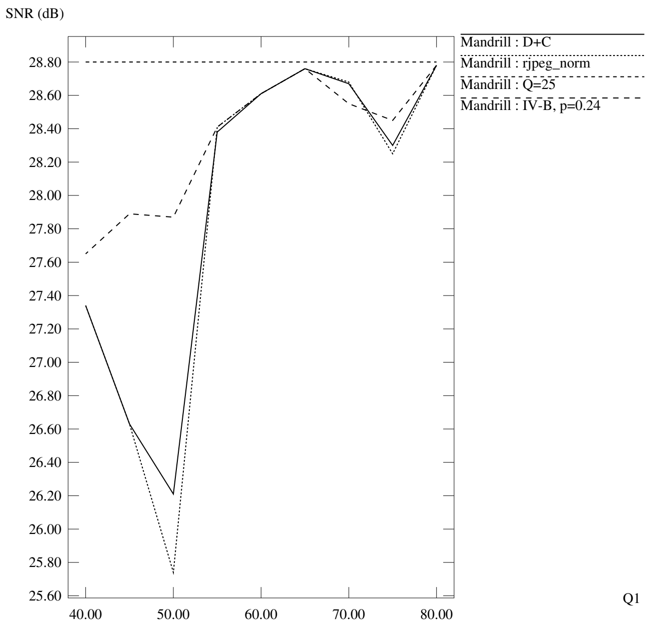

Figure B.28: SNR of recompressed image Mandrill for Q2 = 25

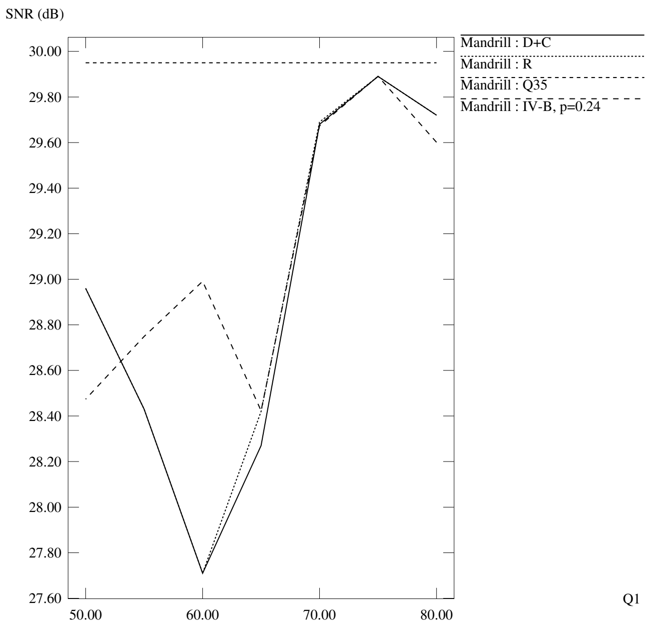

Figure B.29: SNR of recompressed image Mandrill for Q2 = 35

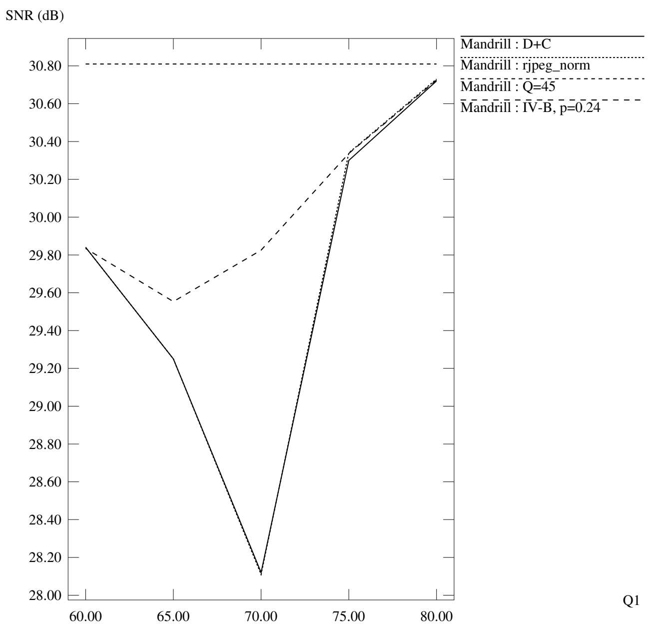

Figure B.30: SNR of recompressed image Mandrill for Q2 = 45

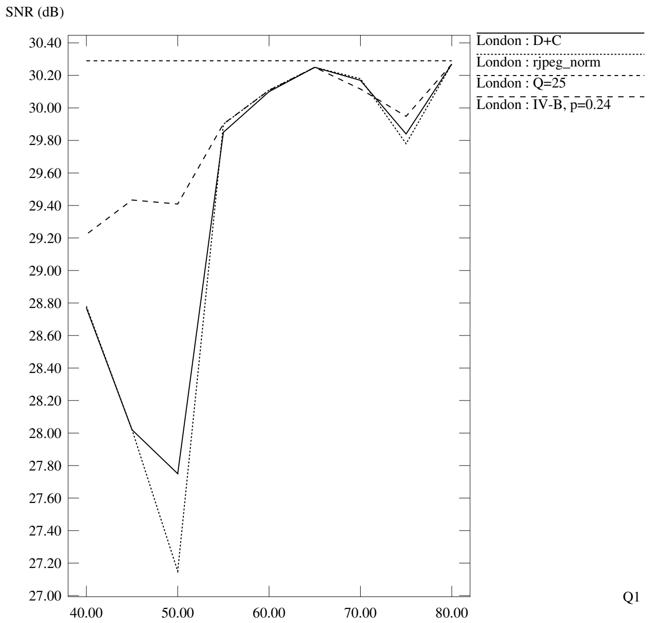

Figure B.31: SNR of recompressed image London for Q2 = 25

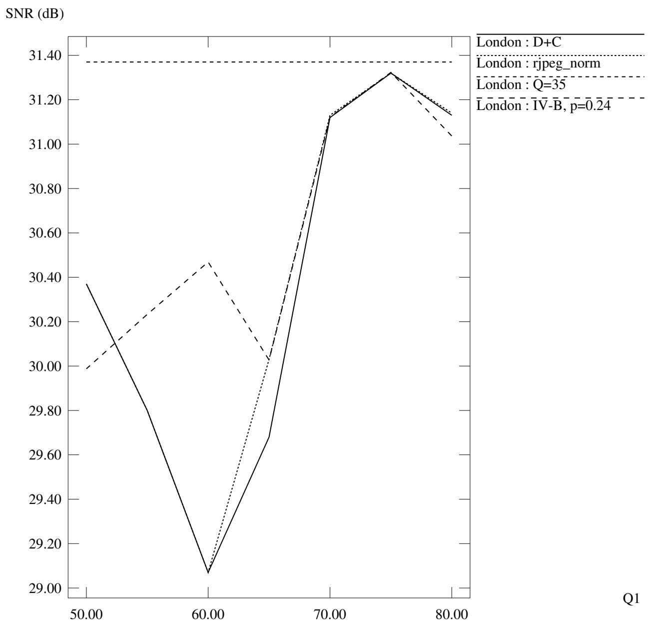

Figure B.32: SNR of recompressed image London for Q2 = 35

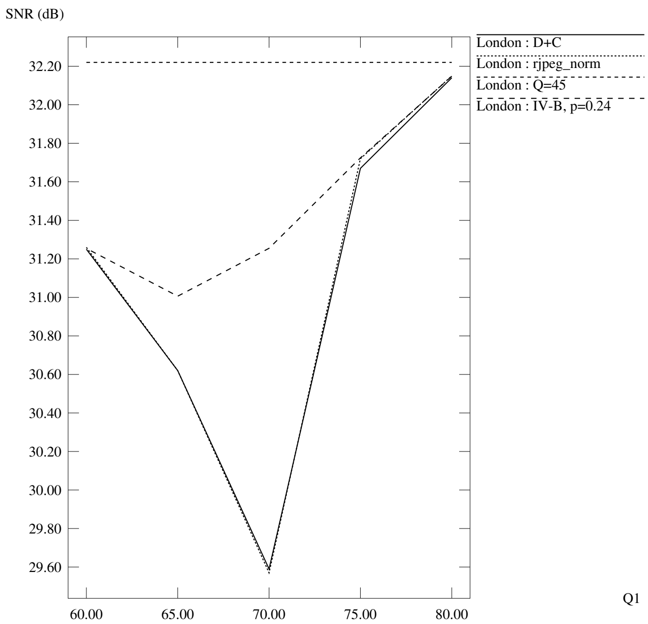

Figure B.33: SNR of recompressed image London for Q2 = 45