VIII. Quantisation

Quantisation forms the kernel around which most of the compression revolves once the transform has been chosen. From the information theory guidelines this section is equivalent to solving (1), and sometimes both (1) and (2). The input of the image codec is assumed to be already digitised with a sufficient number of bits per pixel (6 bits or more for normal monochrome video systems) to avoid contouring effects. As a result of the mathematical processing in the transform, the dynamic range of the transform coefficients is much more than that of the spatial image. The larger the block size the bigger the dynamic range of the coefficient since there are more multiply and adds. The coefficients need to be quantised further in order to achieve the desired compression ratio. The quantisation of the coefficients will be investigated in this section. The first part of this section deals with the actual quantiser design while the second part is concerned with the optimal allocation of bits to each coefficient.

Since the dynamic range of the coefficients is bounded the distribution of the coefficients is always bounded between absolute maximum and minimum values. These values are determined by the dynamic range of the input signal and the type of transform used. For an orthonormal real transform the maximum and minimum values are half (since the input is positive and the basis-function norm is one) that

of the maximum value of the input signal times a constant that is proportional to the size of the transform.

Three different types of quantisation will be investigated. The first is the Lloyd-Max quantiser that has been frequently used and still is very popular for use in image coding systems. This quantiser makes use of the probability density function (pdf) of the coefficients to iteratively minimise the mean square error. The second set of quantisers are more simple and are classified as companders. As the name indicates they make use of a nonlinear function derived from the pdf to transform the coefficient to a uniform distribution where it can be quantised uniformly. Both of these quantisers operate on a sample by sample, i.e scalar, basis. The third method of quantisation that will be looked at is that of source coding. In source coding the coefficients are quantised uniformly and then coded via an optimal source encoding scheme like Huffman coding.

A. Lloyd-Max Quantisers

Quantisation [1,57] is the process of subdividing the range of a signal into non-overlapping regions. The amplitude of the signal is compared to a set of decision levels. If the sample amplitude falls between two decision levels, it is quantised to a fixed reconstruction level lying between the two decision levels. The quantiser as defined here is a memoryless non-linearity and is shown graphically in figure 16.

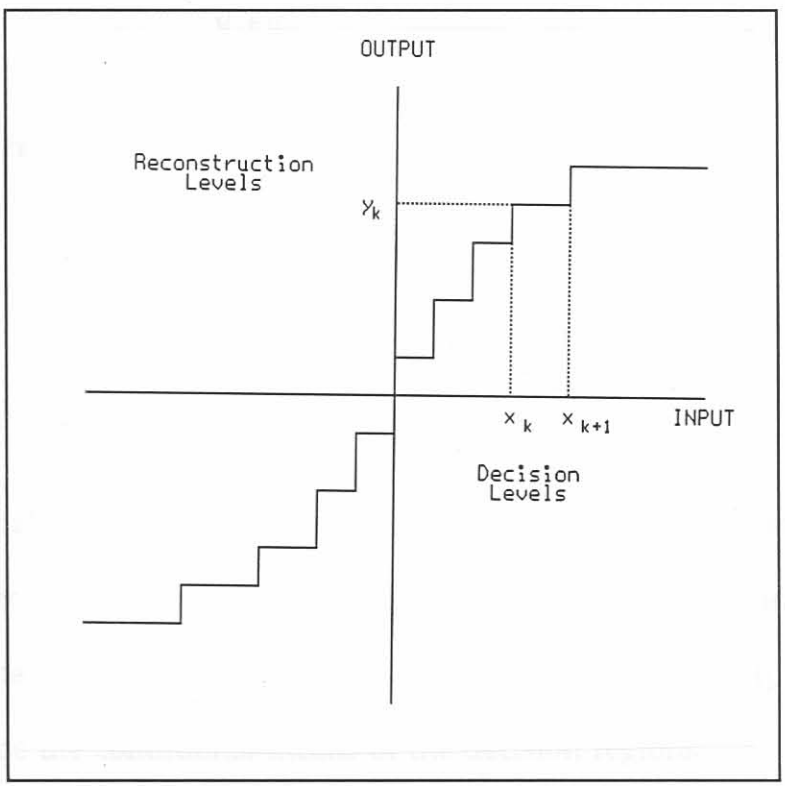

Figure 16 Quantiser input output characteristic.

Consider an N level quantiser with output levels $y_1, y_2, \ldots, y_N$. The output level y_k is associated with a decision region specified by its boundaries, the decision levels,

$$y_k \Leftrightarrow < \{x_{k-1} < x \leq x_k\}, \quad for \quad k = 1,2,\ldots,N. \tag{31}$$

For convenience, the x_i are in increasing order and the two extreme decision levels, x_0 and x_N, are chosen at infinity.

The total mean square error (mse) is

$$\overline{e^2} = \sum_{i=1}^{N} \int_{x_{i-1}}^{x_i} (x - y_i)^2 \, p(x) \, dx \tag{32}$$

Differentiating the mse with respect to x_k and y_k and setting equal to zero gives the decision levels as

$$x_k = \frac{y_k + y_{k+1}}{2}, \quad for \quad k = 1,2,\ldots,N-1 \tag{33}$$

and the reconstruction levels as

$$y_k = \frac{\int_{x_{k-1}}^{x_k} x \, p(x) \, dx}{\int_{x_{k-1}}^{x_k} p(x) \, dx}, \quad for \quad k = 1,2,\ldots,N \tag{34}$$

These conditions must be satisfied by a minimum mean square error quantiser. They can be interpreted to mean that the decision levels should be midway between the output levels and that the output levels should be the conditional means of the decision regions.

Recursive solution of equation (33) and (34) for a given pdf provides the values for the optimum decision and reconstruction levels. Two versions of the iteration that can be used is given in [57]:

Method I

In the first version, often termed Lloyd's method I, an initial guess is made for the output levels and a set of decision boundaries corresponding to these is determined using (33). Then (34) can be applied to determine a new set of output levels which is optimal for the decision boundaries just determined, completing one iteration. At the end of an iteration the mse has either decreased or remained unchanged.

A variation of this technique, introduced by Kabal [57], applies both halves of the iteration to each output level in turn. In this way the effect of changing an output level is allowed to propagate to other output levels. This modified version of method I, which uses the same number of integral evaluations as the original technique, often converges faster in practice. This method was used for the computation of the Max-Lloyd quantiser levels in this thesis, and no problems with convergence were experienced.

Method II

A variational technique, dubbed method II, proposed by both Lloyd and Max involves a one dimensional search. An initial guess is made as to the value of the first output level y_1. The

The Lloyd-Max quantiser was implemented for the Laplacian density function

$$p(x) = \frac{\alpha}{2} e^{(-\alpha|x|)} \qquad (35)$$

$$where\ \alpha = \sqrt{\frac{2}{\sigma^2}}\ ,\ \sigma^2\ is\ the\ variance\ of\ x$$

The reconstruction levels can be solved by replacing (35) in (34), i.e.,

$$y_k = \frac{(\alpha x_k + 1)\, e^{-\alpha x_k} - (\alpha x_{k-1} + 1)\, e^{-\alpha x_{k-1}}}{\alpha\, e^{-\alpha x_k} - \alpha\, e^{-\alpha x_{k-1}}} \qquad (36)$$

The mean square error can be computed by replacing (35) in (32), i.e.,

$$err_i^2 = \int_{x_{i-1}}^{x_i} (x^2 - 2xy_i + y_i^2)\, \frac{\alpha}{2}\, e^{(-\alpha x)}\, dx \qquad (37)$$

this reduces to

$$err_i^2 = f(x_i) - f(x_{i-1})$$

$$where \qquad (38)$$

$$f(x) = e^{-\alpha x} \left(\frac{y(x+1)}{\alpha} - \frac{y^2}{2} - \frac{(\alpha x)^2 + 2x + 2)}{2\alpha^2} \right)$$

The average mse is given by

$$\overline{err^2} = \sum_{i-1}^{N} err_i^2 \qquad (39)$$

This error function has been tabulated in the *bit-assignment* section, Table II. The recursive solution of method I has been used for computing the decision and reconstruction levels. It was found that this method converged quickly for quantiser sizes up to eight bits.

value of the decision level below this output level, in this case x_0, is known. The next decision level can be determined by finding the value of x_1 which satisfies (34). This step is normally carried out using iterative numerical techniques. The next output level can now be computed and the process repeated for all levels. The last output level will generally not be the conditional mean of the last interval. The difference between y_N and the conditional mean of the last interval can be used to determine an update for y_1 for the next iteration. The process of determining the output levels continues until sufficient precision has been achieved.

For the mse, a sufficient condition for uniqueness of the Lloyd-Max solution is log-concavity of the pdf [57]. The Gaussian and Laplace distributions have associated with them unique (and hence symmetrical) quantisers. It is also shown [57] that the Laplace distribution occupies a unique place in the continuum of generalized Gamma distributions in that it sits on the boundary between distributions that have unique optima and those which do not.

Conversion was slow for higher bit rates as a result of the increase in number of levels that had to be computed for every iteration.

All numerical implementations involving this kind of recursive structure with mathematical functions, were done using double precision mathematics. The decision and reconstruction levels generated were verified with those given by Pratt [1 p.144].

A different approach to nonlinear quantisation that achieves similar results is that of companding (compressing and expanding) a signal and using uniform quantisation. The compander quantiser is discussed in the next paragraph.

B. Companders

A compander (compressor-expander) is a uniform quantiser preceded and succeeded by nonlinear transformations as shown in figure 17. The random variable x is first passed through a nonlinear memoryless transformation $g(.)$ to yield another random variable w. This random variable is uniformly quantised to give $y\varepsilon\{yi\}$, which is non-linearly transformed by $h(.)$ to give the output z. The overall transformation from x to z is a nonuniform quantiser. The functions $h(.)$ and $g(.)$ can be easily derived:

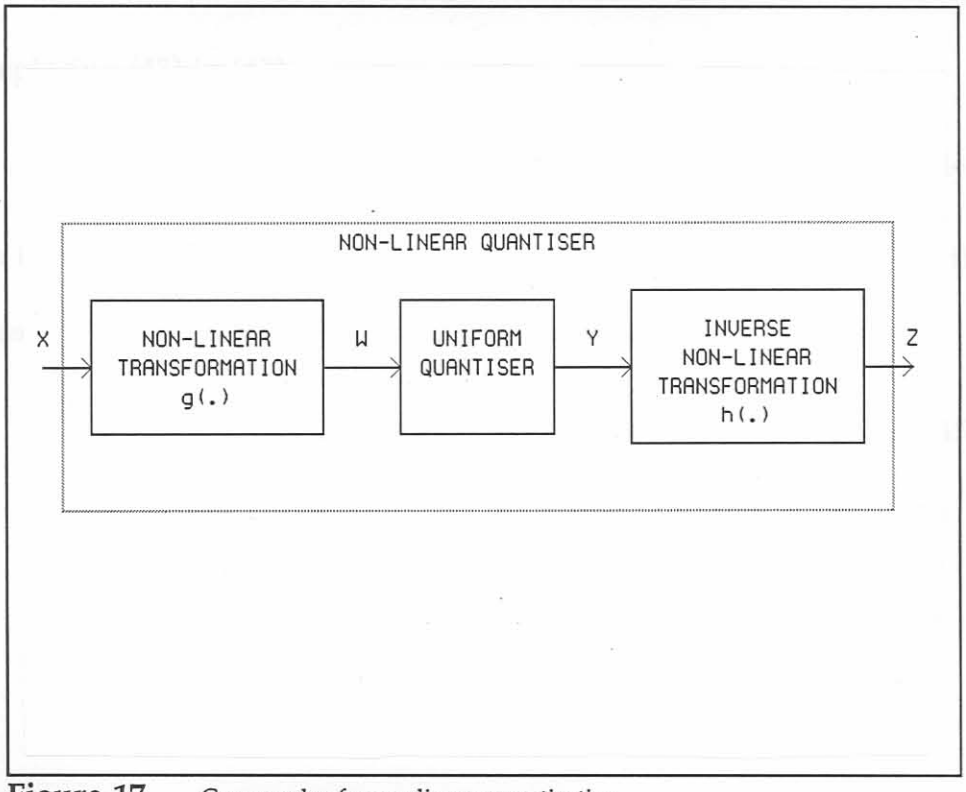

Figure 17 Compander for nonlinear quantisation

We wish to find a function $g(x)$ such that the random variable

$$Y = g(X) \tag{40}$$

transforms the input cumulative distribution function (cdf) $F_X[x]$ to a desired output cdf $F_Y[y]$. The cumulative distribution function is a monotone increasing function, that is

$$F_X[x_1] \leq F_X[x_2] \quad if \quad x_1 \leq x_2 \tag{41}$$

Since $Y=g[X]$, we have

$$P\{Y \leq y\} = P\{X \leq x\} \tag{42}$$

where P is the probability of occurrence.

From the left hand side in (42) and using (40) we have

$$F_Y[y] = P\{Y \leq y\} = P\{g(X) \leq g(x)\} \tag{43}$$

Replacing (43) in (42)

$$F_Y[g(x)] = F_X[x] \tag{44}$$

In (44) $f_X[x]$ is the pdf of the input to the quantiser and $f_Y[g(x)]$ is the desired pdf. The function $g(x)$ can be computed by inverting (44)

$$g(x) = F_Y^{-1}[F_X[x]] \tag{45}$$

For a uniform output pdf, between maximum and minimum values $f_Y(y)$ becomes

$$f_Y(y) = \frac{1}{y_{max} - y_{min}} \tag{46}$$

Substituting (46) into (45) gives

$$g(x) = (y_{max} - y_{min}) F_X[x] + y_{min} \tag{47}$$

The expansion is done by computing the inverse of this function, i.e.

$$h(z) = g^{-1}(x) \quad with \quad z = g(x) \tag{48}$$

The performance that can be achieved with the compander and the Max-Lloyd quantiser is similar [4]. However, the compander is simpler to compute and easier to implement for common distributions, i.e. the Laplacian distribution. As an example of the application of equations (47) and (48) the forward and inverse transformation functions were computed for the Laplacian distribution, the derivation is given in Appendix B. Other companders, i.e Gaussian and Rayleigh, are given by Pratt [1].

C. Source Encoding

For source coding the entropy of the output sequence is specified, if the distribution of the input is known this allows us to compute the step size of the quantiser. For ease of implementation the uniform quantiser is normally used [57]. The uniform quantiser with source coding is normally a sub-optimal quantiser, except for the Laplacian distribution, for which it is the optimum quantiser [53,57,59].

A method to compute the optimal bin width is given by Eggerton and Sirnath [56]. They determined the entropy of the quantised coefficients to be approximately given by, (see Appendix C)

$$H(x^*) = H(x) + \beta\Delta - \log_2\Delta$$

$$\textit{where}\quad \Delta = \textit{stepsize},$$

$$\beta \quad \textit{is a function of the variance},$$

$$H(x) \textit{ is the entropy of the source.} \tag{49}$$

If the coefficients are assumed to be Laplacian distributed, the variables of (49) was found to be [57]

$$H(x) = \log_2(\sqrt{2}\,\sigma e)\,\log_2 e \qquad (\textit{bits})$$

$$\beta \simeq \frac{0.096}{\sigma} \tag{50}$$

$$\textit{where } \sigma \textit{ is the standard deviation}$$

The bin width can now be obtained from (49) by equating $H(x^*)$ with the desired bit rate. Once the coefficients have been quantised, an optimal source encoding method, i.e. Huffman coding, can be used.

The uniform quantiser with source coding were used in simulations on the test image's. A comparison between the Lloyd-Max quantiser and the source encoding quantiser is given in figure 18, with the results for images coded with 2.0, 1.0 and 0.5 bits/pel given in figures 19-21. It is clear from the graph that the source coder performs better than the Lloyd-Max quantiser. This is in agreement with theoretical predictions [7].

For efficient quantisation, it has been shown in the discussion regarding the quantisers, that knowledge of the coefficient density functions is necessary. The next paragraph investigates the different functions that has been proposed in the literature.

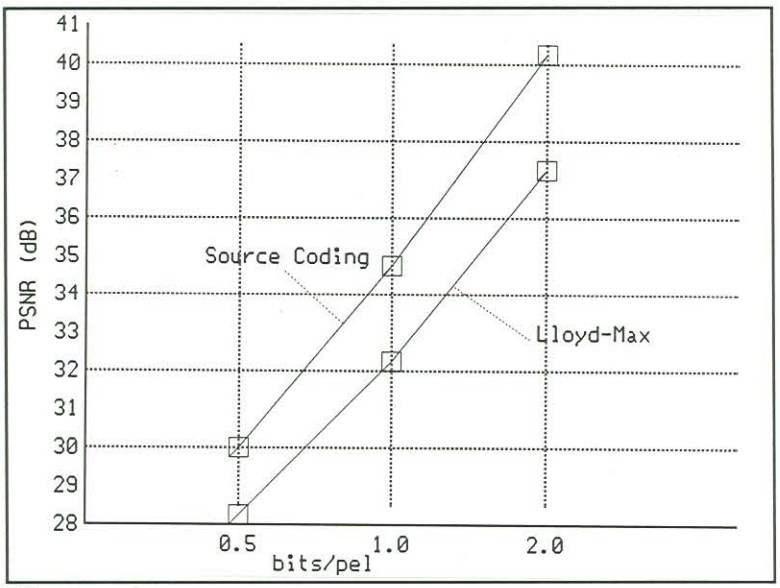

Figure 18 Comparison of the performance of the Source Encoder versus the Lloyd-Max quantiser. Transform=DCT (8x8)

Figure 19 The image GIRL DCT coded to 2.0 bits/pel, using uniform quantisation with source encoding.

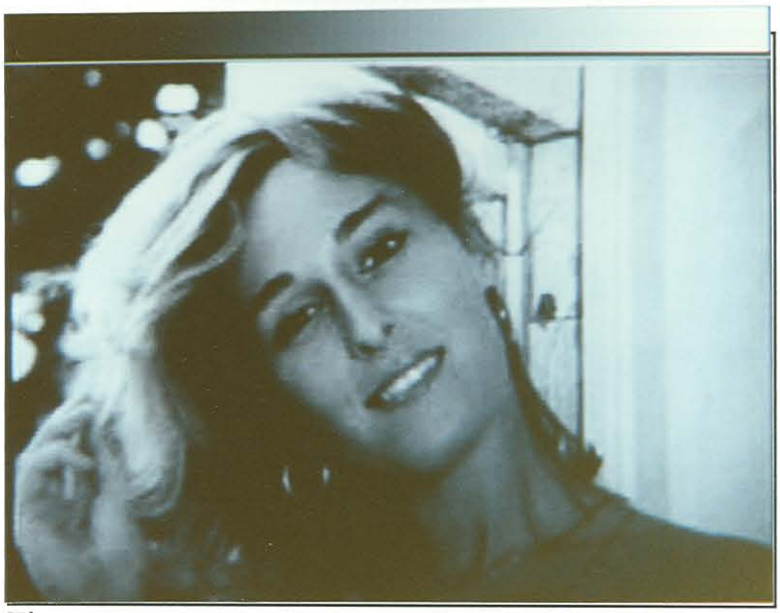

Figure 20 The image GIRL DCT coded to 1.0 bits/pel, using uniform quantisation with source encoding.

Figure 21 The image GIRL DCT coded to 0.5 bits/pel, using uniform quantisation with source encoding.

D. Coefficient Statistics

After the image has been subdivided into a number of smaller blocks, to take advantage of the spatial variant nature of images, these blocks are transformed using the DCT or a similar transform to decorrelate the images. The different coefficients, of similar index, are then grouped together and quantised using one of the quantisation methods just described. Since all of these methods need to know the distribution of the coefficients a priori, in order to determine decision levels or to achieve a certain rate, the determination of a representative distribution is important.

Several distributions have been suggested by various authors. Pratt [1] suggested that the DC coefficient should have a Rayleigh distribution since it was the sum of positive values, and that, based on the central limit theorem, the other coefficients should be Gaussian. Netravali and Limb [7] agreed with the above assumption and also stated that the histogram of non DC coefficients were roughly bell-shaped. On the other hand, some authors [30] thought that the non-DC coefficients were not Gaussian, but Laplacian. A few authors agreed that the DC coefficient was Gaussian. These different assumptions have led Reininger and Gibson [30] to perform goodness-of-fit tests on the transform coefficients in order to identify the distribution that best approximates the statistics of the coefficients. In the tests they considered the Gaussian, Laplacian, Gamma, and Rayleigh distributions. The test that they used was the well-known

Kolmogorov-Smirnov test. The results of their paper indicate that for a large class of images, the DC coefficient is best modeled by a Gaussian distribution, and the non-DC coefficients are best modeled by the Laplacian distribution.

The histograms of the coefficients of the two test images were computed, see figure 22. The results shown in figure 22 were for the coefficient [1,1] as well as the DC coefficient. The images were transformed with an 8x8 DCT transform. The results show that the non-DC coefficients were Laplacian distributed as experimentally determined. The DC coefficient could however not be described as anywhere near Gaussian, but it is recognised that a larger database was used in [30]. It was decided to code the DC-coefficient using a uniform distribution and the other coefficients using a Laplacian distribution. The next paragraph will investigate the question as to which coefficients should be coded and how many bits should be used for every coefficient.

E. Bit Assignment

Bit assignment is normally discussed in the literature under Block quantisation. Block quantisation is an efficient scheme to quantise a block of independently distributed random variables. In block quantisation each element of a vector is quantised on an element by element basis. These elements are normally not quantised equally well as will become clear in the following paragraph.

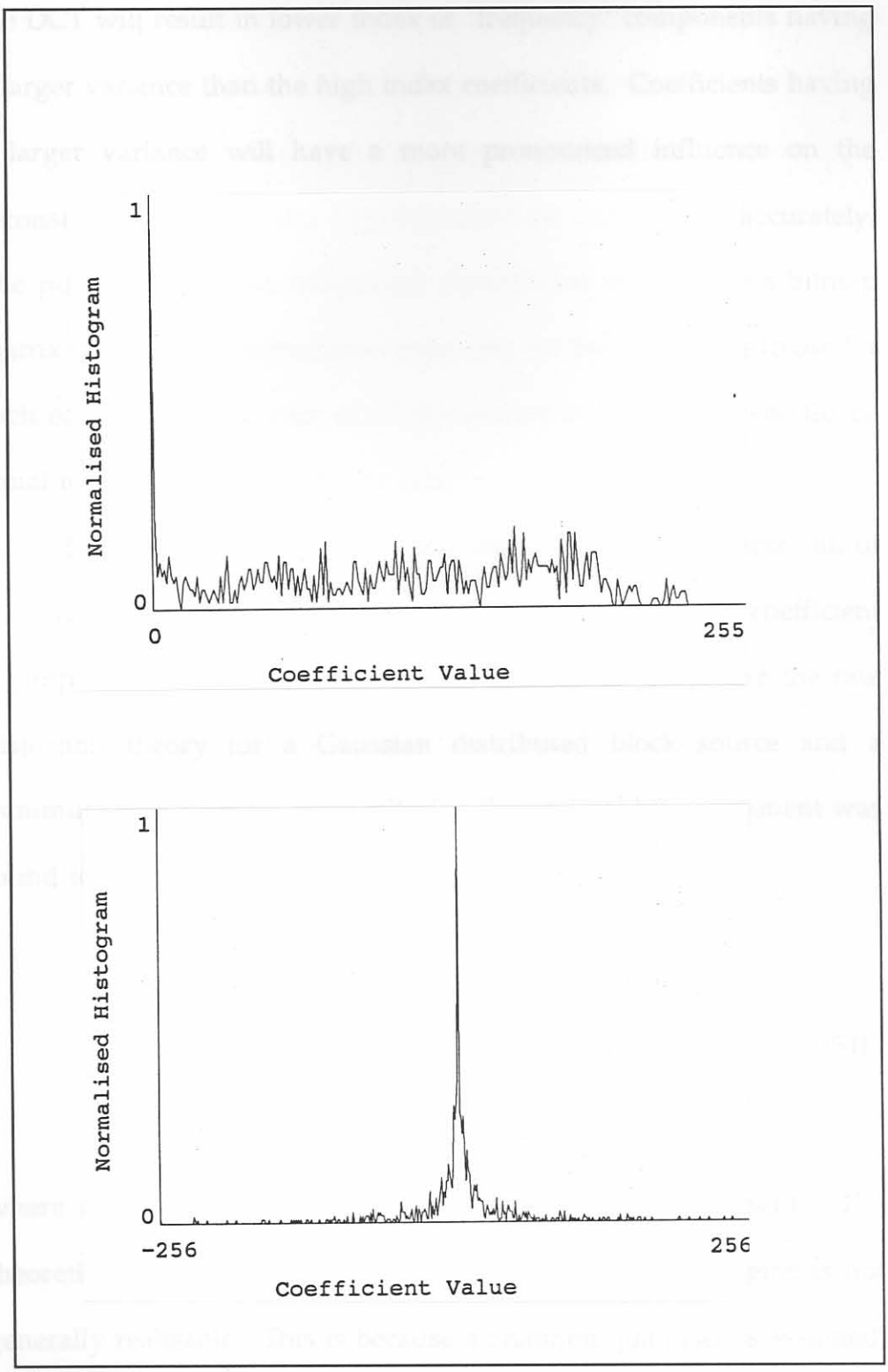

Figure 22 Coefficient distributions for the images GIRL using an 8x8 DCT.

The output of a two dimensional DCT is considered to be a vector under block quantisation. The energy compactation property of

the DCT will result in lower index or "frequency" components having a larger variance than the high index coefficients. Coefficients having a larger variance will have a more pronounced influence on the reconstructed image, and therefore need be coded more accurately. The purpose of the bit assignment algorithm is to generate a bitmap matrix that contains information regarding the number of bits to use for each coefficient. The sum of all the entries in this matrix should be equal to the desired average bit rate.

Several methods exist for computing this bitmap matrix, all of them have in common that the number of bits assigned to a coefficient is proportional to the variance of that coefficient. Applying the rate distortion theory for a Gaussian distributed block source and a minimum mean square error criterion the optimal bit assignment was found to be [1,7],

$$b_i = \theta + \frac{1}{2}\log_2 \frac{\lambda_i}{(\lambda_1 \lambda_2 \cdots \lambda_N)^{1/N}} \qquad (51)$$

where λ_i *is the variance of the i th element,*

θ = *desired average number of bits*

where b_i is the number of bits assigned to the ith element. The theoretical performance of the rate-distortion theoretic scheme is not generally realisable. This is because a Shannon quantiser is assumed, which is difficult to achieve practically [7], and the optimally assigned number of bits is not necessarily integer, and may even be negative.

Rounding to zero or the nearest integer is required, which may offset the optimality of the bit assignment.

Besides the rate distortion theoretic method, there is a computational approach to bit allocation [58]. This approach allocates the total number of available bits to the vector on a bit-by-bit basis using a marginal analysis technique.

The overall average quantisation error for an N-dimensional source is

$$D = E\left[\sum_{i-1}^{N} (x_i - \hat{x}_i)^2\right] \tag{52}$$

where x_i is original element, $\hat{x}_i$ is the reconstructed element

Let *f(n)* be the mean-square quantisation error (msqe) of an n-bit quantiser for a source with unity variance. Then, (52) can be written as

$$D = \sum_{i-1}^{N} \lambda_i f(n_i) \tag{53}$$

where n_i is the number of bits assigned to the ith element of the vector. The bit assignment is initialised by setting all the n_i equal to zero, i.e. a zero bit map matrix. Using the variances of each coefficient a marginal return can be computed for the coefficients. The marginal returns basically determine which coefficient would gain most by assigning another bit to that coefficient's bitmap.

The marginal returns are defined by

$$\Delta_i = \lambda_i [f(n_i) - f(n_i+1)]$$

$$\textit{where } \lambda_i \textit{ is the variance of the } i \textit{ th element,} \qquad (54)$$

$$f(n) \textit{ is the msqe}$$

The bit is allocated to the coefficient that has the largest marginal return. This method is repeated until all the bits has been assigned. The computational approach is superior to the rate-distortion theoretic approach since there is no round-off error in the bit assignment. However, its computational load is much heavier than the other approach. The rate distortion approach further assumes that each element of the vector is Gaussian distributed and that a Shannon quantiser is employed. This is not the case for the coefficients of the DCT transform which is predominantly Laplacian distributed [30]. When using a pdf optimised nonuniform quantiser the error for this quantiser should be used in the bit assignment. A list of the msqe of pdf-optimised nonuniform quantisers for Gaussian, Laplacian, and Gamma distributions is given in Table II [4,58]. It is interesting to note that the differential decrease in error for all distributions tends toward that of the Shannon quantiser, i.e. four, for large n. This means that it might not be necessary to compute the exact error values for large n.

n	f(n)			
	Shannon	Gaussian	Laplacian	Gamma
0	1.0	1.0	1.0	1.0
1	0.25	0.3634	0.5	0.668
2	0.0625	0.1188	0.1963	0.32
3	0.015625	0.03744	0.07175	0.1323
4	0.0039062	0.01154	0.02535	0.0501
5	0.00097656	0.003495	0.008713	0.01784
6	0.00024414	0.001041	0.002913	0.006073
7	0.00006104	0.0003035	0.0009486	0.001996
8	0.00001526	0.00008714	0.0003014	0.0006379

Table II Quantisation Errors of Shannon Quantiser and Max Quantisers for Gaussian, Laplacian, and Gamma Distributions.

F. Adaptive Quantisation

It has been noted in previous discussions that the statistics of real images are spatially variant. The performance of the quantiser could be increased if we could make the quantiser adaptive to local statistics in the image. The main advantage associated with adaptive quantisation is the improvement in the ability of the codec to code detail in the image.

Several methods to achieve this has been presented in the literature [5,7,8,17]. Most of the techniques use some activity measure defined for each block. This activity measure is mostly related to the ac-energy or the variance of the block's coefficients. In the method by Chen [5], the image is classified into four different classes according to

the variance of the blocks. A bitmap matrix is generated for every class. This increases the overhead associated with the quantisation information. It has been estimated by Chen that the overhead is approximately 0.034 bits per class. This places a limit on the number of classes that are practical for adaptive coding.

The purpose of this section is to investigate the relationship between the improvement in image quality and the number of *classes*. In other words the optimal number of classes will be determined using empirical techniques. The classification will consist of computing the ac-energy of each block and dividing the blocks equally between the different classes. To keep the bit assignment optimal, the algorithm assigns bits, using the marginal technique, concurrently to all bitmap matrixes. In other words the bits are spread out between the classes on a largest "error reduction" basis.

The results of the simulation is shown graphically in figure 23. Figures 24 to 27 show the test images coded using one class and eight classes respectively. It is clear from the results achieved that adaptive quantisation based on an activity index improves the quality of the image. Other more advanced techniques that exist but have not been used in simulations are the use of masks in the computation of the block energy. This basically involves generating masks that group certain features of the image, i.e. horizontal, vertical, and diagonal lines [17]. Some techniques use the sensitivity of the HVS to different DCT coefficients to generate the masks [12]. Most of these techniques

require a large number of adaptation of parameters to specific images, making them less useful.

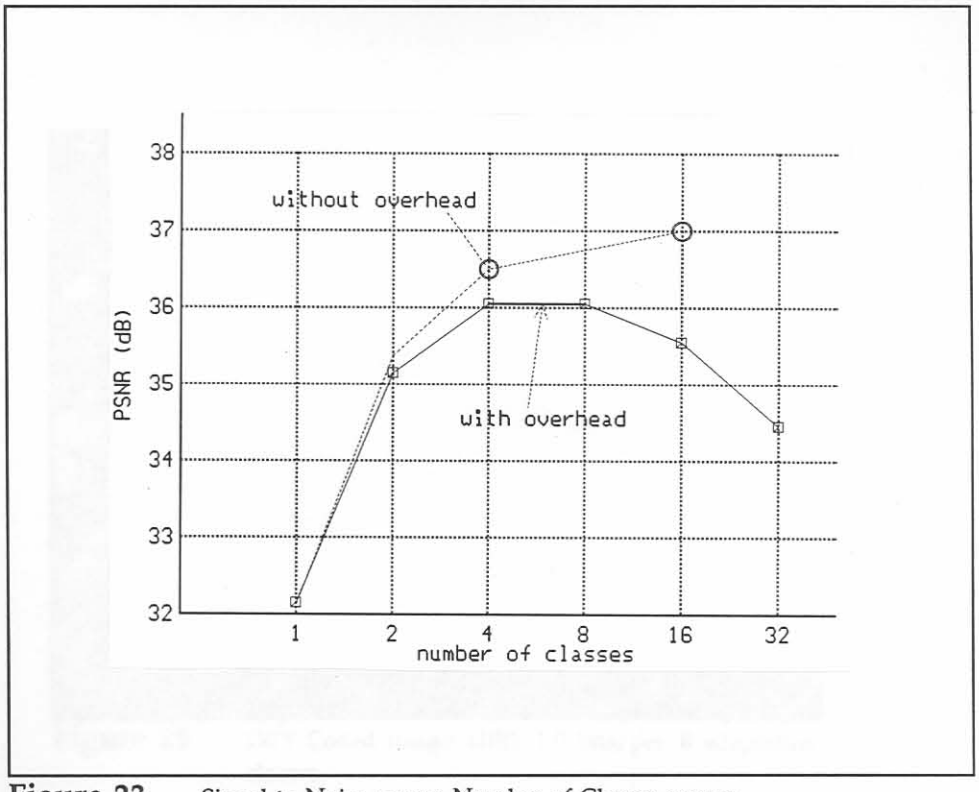

Figure 23 Signal to Noise versus Number of Classes curves

Figure 24 DCT (8x8) Coded Image: GIRL 1.0bits/pel, one adaptation class.

Figure 25 DCT Coded Image: GIRL 1.0 bits/pel. 8 adaptation classes.

Figure 26 DCT Coded Image: ROAD 1.0 bits/pel. 1 Adaptation Class.

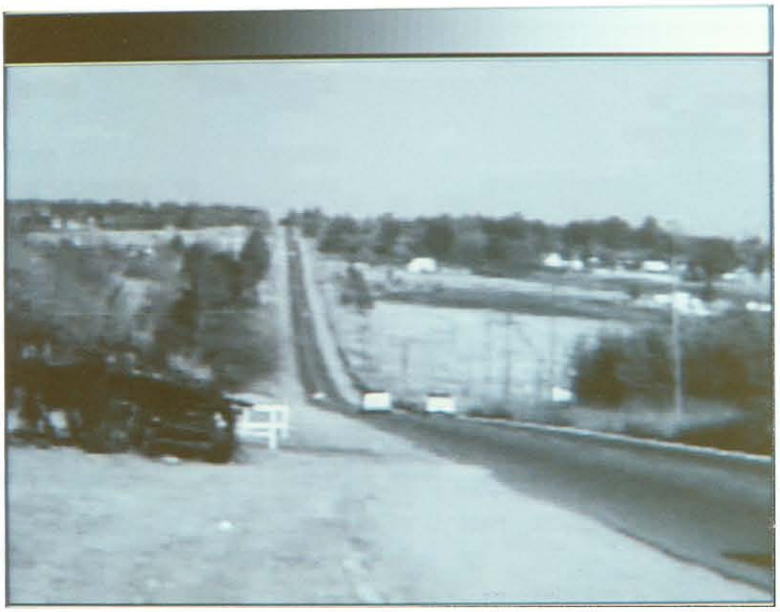

Figure 27 DCT Coded Image: ROAD 1.0 bits/pel. 4 Adaptation Classes.