L^∞-CODING OF IMAGES: A CONFIDENCE INTERVAL CRITERION

Lamia Karray, Olivier Rioul, *and Pierre Duhamel*

France Telecom–CNET/PAB/STC/SGV
38-40, rue du Général Leclerc,
92131 Issy-les-Moulineaux, France

ENST/SIG
46, rue Barrault
75013 Paris, France

ABSTRACT

A new image coding technique using statistical properties of quantization errors and L^∞-norm criterion is investigated. The original image is preprocessed, quantized, encoded and reconstructed within a given confidence interval. We focus on iterated filter banks as a preprocessing technique, and provide a comparison with linear prediction in the case of very good quality (almost lossless) image coding.

1. INTRODUCTION

The transmission and storage of digital signals require the original data to be compressed. In many applications, the image is transformed into a set of binary integers such that the original data can be recovered from the binary set within some level of degradation. These "lossy" image compression schemes are usually based on transform coding techniques. The original image is split into a set of coefficients using some invertible transformation which improves the signal statistics prior to encoding. Such transformations are: DCT, filter banks, wavelets [2], wavelet packets [7], etc. Linear prediction [6] can also be considered as a particular transformation. These transformations do not produce any bit-rate reduction, but prepare the following steps. The transform coefficients are then quantized (which produces some data compression, but usually generates distortion in the reconstructed signal) and possibly entropy-coded (further compressing the data, without additional distortion).

The compression scheme is optimized by choosing an appropriate set of parameters (transform, quantizer steps) so that the overall compression ratio is minimized for a given reconstruction error. Classical lossy compression applications minimize an m.s.e.-like criterion on the reconstruction error [7]. Thus, the overall distortion is evaluated and controlled using L^2-norm. However, this criterion is global (on the whole image) and does not exploit "local" knowledge on the signal, which is always available. At least, one knows the number of pixels on which the original image is encoded. Moreover, in many applications, one has indications on the precision with which the pixel values are obtained. The L^2-norm cannot take such information into account, since it "averages" the errors in the whole image. This paper proposes methods allowing such knowledge to be used as a criterion involving a "confidence interval". As a result, this paper sticks to the case of very good quality (almost lossless) image coding.

As a preprocessing, we considered two classical kinds of transformations: iterated "wavelet" filter banks and linear prediction. In each case, the original signal x is transformed, quantized, encoded and reconstructed to give $\hat{x}$, with a distortion $\Delta x = x - \hat{x}$. In our schemes, either prediction or filter banks achieve perfect reconstruction if no quantization is involved. If quantization occurs, the reconstruction error is only due to the quantization errors. Therefore, the problem is to determine the best quantizers in the compression scheme so as to achieve at least p% of errors inside the confidence interval. In other words, we minimize the overall bit rate under the constraint

$$prob\{|\Delta x| \leq t\} \geq p\% \tag{1}$$

As an example, the compression scheme can be tuned in such a way that only 2% the reconstruction error exceed half the initial quantization step: in other words, 98% the output pixels will be equal to the original ones.

2. FILTER BANKS CODING

The corresponding compression scheme (figure 2) uses a perfect reconstruction octave-band filter bank, scalar uniform quantization and global Huffmann coders. For a given set of perfect reconstruction filters, the problem is to determine the quantizers that minimize the global bit rate while ensuring that the constraint (1) is met. Solving this problem requires some statistical modelling of the reconstruction error.

2.1. Statistical properties of the reconstruction error

The transformation applied to the original signal increases the density of the original discrete data and allows a continuous modeling and a statistical study. It is shown in [1] that the signals in the various high-pass subbands are accurately modeled by Laplacian distributions.

After transformation, the resulting signals y are quantized to be encoded. Here, we consider only scalar uniform quantization with step q. The quantization error is thus an additive noise lying between $\frac{-q}{2}$ and $\frac{q}{2}$ [8]. It is shown in [8] that, if the ratio $\frac{\sigma}{q}$ is high (more than 0.7), the quantization noise can be modeled as a uniform distribution:

$$f_e(y) = \begin{cases} \frac{1}{q} & \text{if } \frac{-q}{2} \leq x < \frac{q}{2} \\ 0 & \text{otherwise} \end{cases}$$

with a zero-mean and a variance $\frac{q^2}{12}$. In our case (almost lossless coding), the above condition is satisfied and the quantization error is uniform.

0-8186-6950-0/94 $4.00

Now, since we use perfect reconstruction filters and lossless coders, the error is generated by the quantizers only. The system being linear, the reconstruction error is the result of the contribution of the various quantization errors to the actual output. In each subband, the quantization error is interpolated and filtered through the synthesis filters as depicted in figure 1 for a single iteration on a 1D-signal. For simplicity, we show in this 1-D context that the reconstruction error has a Gaussian distribution. The demonstration would be equivalent in the 2-D case.

Due to the oversampling, the odd and even samples of Δx depend on a different set of filter coefficients.

$$\begin{aligned}\Delta x_{2n} &= \sum_{k=0}^{\frac{N_g}{2}-1} \epsilon^0_{n-k} g_{2k} + \sum_{k=0}^{\frac{N_h}{2}-1} \epsilon^1_{n-k} h_{2k} \\ \Delta x_{2n+1} &= \sum_{k=0}^{\frac{N_g}{2}-1} \epsilon^0_{n-k} g_{2k+1} + \sum_{k=0}^{\frac{N_h}{2}-1} \epsilon^1_{n-k} h_{2k+1}\end{aligned} \quad (2)$$

where N_g (resp. N_h) is the filter g (resp. h) length. The quantization errors ϵ^i have, a uniform distribution in $[\frac{-q_i}{2}, \frac{q_i}{2}]$ with zero-mean and variance $\frac{q_i^2}{12}$.

The reconstruction error is thus a linear combination of uniform distributions, with different widths (due to the different values of the quantization steps in the subbands). It can be shown that the reconstruction error converges to a Gaussian distribution, provided that some conditions on its first three moments are satisfied (in order to use central limit ramifications developped in [5]). Due to lack of space, the demonstration is only outlined here.

Consider one subband of the synthesis phase, e.g. the low-pass one of figure 1, and denote by e its contribution to the total reconstruction error. We have $e_p = \sum_k \epsilon^0_{n-k} g_{2k+p}$ where $p = 0$ for the even samples and $p = 1$ for the odd samples. It is easily shown that e_p has zero-mean and variance $\sigma_p^2 = E[(\sum_k \epsilon^0_{n-k} g_{2k+p})^2]$.

Under appropriate conditions, which hold under the same hypothesis as the one recalled earlier for the quantization noise to be uniform, the reconstruction error can be shown to be a sum of such errors (which also holds for 2D-signals and many iterations of the filter bank). Thus, the generalized central limit theorem, as detailed in [5], can be used to show that an infinite sum of these errors has a Gaussian distribution. In practice, the filter banks are iterated several times. For $J = 5$, this corresponds to 16 subbands and the convergence to a Gaussian distribution is ensured.

Therefore, we end up with a Gaussian reconstruction error, with zero-mean and variance σ^2 depending on the quantization steps and the filter coefficients in all subbands. For example, when $J = 1$, in the 2-D case, we have

$$\begin{aligned}\sigma^2_{p,q=0,1} &= \frac{q_0^2}{12} \sum_{k,l} g^2_{2k+p} g^2_{2l+q} + \frac{q_1^2}{12} \sum_{k} h^2_{2k+p} g^2_{2l+q} \\ &+ \frac{q_2^2}{12} \sum_{k,l} g^2_{2k+p} h^2_{2l+q} + \frac{q_3^2}{12} \sum_{k,l} h^2_{2k+p} h^2_{2l+q}\end{aligned}$$

where p and q are associated to odd and even samples of the error. In general, for J iterations, the whole error is a contribution of the different samples $(\Delta x_{p,q})_{p,q=0,\ldots,2^J-1}$ due to the J interpolations. Furthermore, these samples have the same contribution in the whole signal, and are all Gaussian with variance $\sigma^2_{p,q}$.

2.2. Quantizer optimization

Since a statistical model of the reconstruction error is available, we can find the quantizers allowing a certain percentage of distortion as in (1). This requirement reads:

$$\text{prob}\{|\Delta x| \leq t\} = \frac{1}{2^{2J}} \sum_{p,q} \text{prob}\{\Delta x_{p,q}\} = \frac{1}{2^{2J}} \sum_{p,q} \text{erf}\left(\frac{t}{\sigma_{p,q}\sqrt{2}}\right) \quad (3)$$

where $\text{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-x^2} dx$ is the usual error function and $\sigma^2_{p,q}$ depends on the squared quantization steps as seen above.

Decimation in the analysis filters halves the size of the subimage in each subband and since we use separable dyadic filters, the contribution to the total bit-rate of a subimage after j iterations is weighted by $n_i = 2^{-2j}$. Let b denote the total bit rate, r_i and q_i the dynamic range and quantization step in subband $\#i$, respectively. The contribution of subband i to the total bitrate is b_i, which satisfies $q_i = r_i\, 2^{-b_i}$.

Quantizers have to be chosen such that the total bit rate, $b = \sum_i n_i b_i$, is minimum while (3) is satisfied. Thus, we end up with a constrained minimization problem

$$(P) \begin{cases} \min(\sum_i n_i \log_2(\frac{r_i}{q_i})) \\ \frac{1}{2^{2J}} \sum_{p,q} \text{erf}(\frac{t}{\sigma_{p,q}\sqrt{2}}) \geq p\% \end{cases} \quad (4)$$

In its present form, it is a nonlinear minimization problem with a nonlinear constraint. To avoid divergence of general optimization algorithms, we solve the problem in two steps: first, we determine the best quantizers which minimize the bit rate such that the maximum error is less than a given threshold. Then, using the relationship between the variance of the error and the quantization steps, we solve (1).

2.3. First step: Deterministic point of view

In a first step, a deterministic approach is used to find the quantizers such that the bit rate is minimum and the reconstruction error Δx does not exceed a given threshold t, i.e., $|\Delta x_n| \leq t$ for all n. This condition involves the L^∞ norm of the error, and leads to a constrained minimization problem:

$$\begin{cases} min(\sum_i n_i b_i) \\ \|\Delta x\|_\infty \leq t \end{cases} \quad (5)$$

2.3.1. Estimation of the distortion L^∞ norm

Since the exact value of the maximum depends on the original image, we can only give an upper bound as an estimate of the L^∞ norm of the reconstruction error. We first keep with the simplified context of a monodimensional signal, with a simple 2-band filterbank. The result is then extended to 2D-signals and J iterations.

According to (2), an upper bound is obtained if the quantization errors (ϵ^0 and ϵ^1) are extremum and have the same sign as the filter coefficients. Since ϵ^i satisfies $|\epsilon^i| \leq \frac{q_i}{2}$, where q_i is the quantization step, we obtain:

$$\|\Delta x\|_\infty \leq \max \begin{cases} \frac{q_0}{2} \sum_k |g_{2k}| + \frac{q_1}{2} \sum_k |h_{2k}| \\ \frac{q_0}{2} \sum_k |g_{2k+1}| + \frac{q_1}{2} \sum_k |h_{2k+1}| \end{cases}$$

Since separable filter banks are involved, the filtering and interpolation are separately done on lines then columns of the

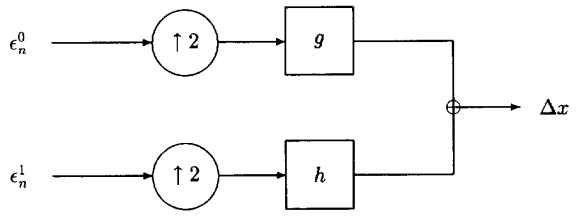

Figure 1: The reconstruction error. A single iteration of a one-dimensional signal is considered. ϵ_n^0 and ϵ_n^1 are quantization errors. Δx is the reconstruction error, g and h are the set of the filter coefficients.

2D-signals, which results in four types of samples, depending on the parity of their indexes, giving

$$\|\Delta x\|_\infty \leq \max \left\{ \begin{array}{l} (\Delta x_{2m,2n})_{\max}, (\Delta x_{2m,2n+1})_{\max}, \\ (\Delta x_{2m+1,2n})_{\max}, (\Delta x_{2m+1,2n+1})_{\max} \end{array} \right.$$

After J iterations of the filter bank we similarly obtain:

$$\|\Delta x\|_\infty \leq \max_{p,q} \left\{ \max_{m,n} (\Delta x_{2^J m+p, 2^J n+q}) \right\}$$

where p and q vary between 0 and $2^J - 1$ (i.e. $2^J \times 2^J = 2^{2J}$ possibilities) and come from the J successive interpolations in the synthesis; m and n describe the different signal samples.

Without any *a priori* knowledge on the signal, this estimator is optimal, but somewhat pessimistic for actual images. Note that it is a linear function of the quantization steps.

Since Δx is the contribution of the interpolated samples $\Delta x_{p,q}$, the relation $\|\Delta x\|_\infty \leq t$ has to be satisfied for every (p, q) combination, which yields to a linear system of constraints. Thus, the optimization problem (5) has a convex criterion: $\sum_i n_i \log_2(\frac{r_i}{q_i})$, using the notations of section 2.2, and linear constraints $\sum_i a_{ij} q_i \leq t$ where a_{ij} function of the set of filter coefficients involved in subband i and the combination j of p and q seen above.

2.3.2. Deterministic computation of the quantization steps

Finding quantization steps that minimize the bit-rate such that the reconstruction error is $\leq t$ is thus equivalent to solving the constrained optimization problem

$$(P_1) \left\{ \begin{array}{l} min(\sum_i -n_i log_2 q_i) \\ \sum_i a_{ij} q_i \leq t_j \quad j = 1, \ldots, 2^{2J} \end{array} \right. \tag{6}$$

with convex criterion and linear constraints (figure 3.a).

As illustrated in figure 3, the solution is unique, and lies on the boundary of the feasible region. Since, in actual problems, the number of constraints is huge (over 1,000 constraints when the low-pass filter is iterated five times in 2-D), general optimization procedures cannot be used. Moreover, these general optimization algorithms need a starting point close to the optimum so as to avoid divergence problems. Therefore, we have chosen to solve the problem in two steps: first, find an approximate solution, then use a general optimization algorithm using this approximation as a starting point to find the optimum solution.

Using equation $q_i = r_i \, 2^{-b_i}$, problem (6) can also be written

$$(P_2) : \left\{ \begin{array}{l} min(\sum_i n_i b_i) \\ \sum_i \alpha_{ij} 2^{-b_i} \leq t_j \end{array} \right. \tag{7}$$

with linear criterion and convex constraints (figure 3.b).

Problems (P1) and (P2) are equivalent and related through the function 2^{-x}. Since this function is convex and invertible, each vertex (Q_i) of (K_1) is associated to a "virtual" vertex (B_i) of (K_2) which saturates the same constraints. The feasible domain (D) determined by the constraints (B_i) is a convex polyhedron. Furthermore, the minimization of the linear criterion on (D) gives the best vertex of (6); the optimal solution is on the boundary. Since the criterion is linear and (D) a convex polyhedron, the best vertex is obtained using a variation of the simplex algorithm [9]. But, since (D) is not easily determined and (K_1) and (K_2) are equivalent, the best vertex in (D) is associated to the vertex in (K_1) where the criterion is minimum. The problem can thus be solved using version (6). However, we have to adapt the simplex algorithm to this particular case of a convex criterion, by evaluating the cost function at each vertex.

The obtained best vertex is an approximation of the optimum which is on the boundary of the feasible region. The obtained approximate solution is used in the second step as a starting point of some general optimization algorithm, to find the optimum solution. Its convergence is thus ensured and relatively fast.

So far, we have determined optimal quantization steps according to the deterministic approach in (6). As it is, the proposed solution already allows to perform lossless coding, by choosing $t = 0.5$, and requantizing $\hat{x}$ after reconstruction. However, some applications allow a "small" number of errors to exceed the given threshold. Hence we now return to the statistical problem (4).

2.4. Statistical computation of the quantization steps

Recall that the quantization steps depend linearly on the threshold t. Hence, if the threshold t is multiplied by some constant α, the steps q_i are also multiplied by α and the variance σ^2 of the quantization error by α^2. Therefore, it is natural to state the problem relatively to the one solved in section 2.3.2 : the new quantization steps are chosen as $q_i' = \alpha \, q_i$. This scale factor α is thus easily determined using the relation

$$erf(\frac{t}{\alpha \, \sigma \, \sqrt{2}}) \geq \frac{p}{100} \tag{8}$$

Taking into account the different samples $\Delta x_{p,q}$, we find α by solving the equation:

$$\frac{p}{100} = \frac{1}{2^{2J}} \sum_{p,q} \mathrm{erf}\left(\frac{t}{\alpha\ \sigma_{p,q}\ \sqrt{2}}\right) \qquad (9)$$

where $\sigma_{p,q}$ is the standard deviation associated to $\Delta x_{p,q}$ and is a function of the optimal quantization steps solution of the deterministic problem (6). Since (9) is nonlinear, it is solved using a general algorithm. However, since it depends on a single variable α, solutions are reliable and fast.

2.5. Results

A direct application of the above scale factor α to the quantization steps q_i computed as described in section 2.3.2, gives the results shown in table 1. These results are obtained in the context of "almost lossless coding". The first column provides true lossless coding ($t < 0.5$) using the optimal quantizers according to the L^∞-criterion. Since the estimate of the L^∞ norm is not achieved in the real images, the obtained maximum error is less than the imposed threshold $t = 0.5$. So, using an iterative algorithm and the linear dependency between the quantization steps and the threshold, we find the scale factor which brings the reconstruction error at $t = 0.5$. The other columns of table 1, show as expected that the bit rate reduction is more and more important when p decreases.

All this procedure may seem fairly complicated. However, note that all quantities are signal-independent: the whole optimization can be done off-line. It amounts to computing the quantization steps taking an L_∞ norm into account rather than an L_2 norm. Otherwise, the whole coding/decoding scheme is identical to the classical one: only the quantization steps differ. It should also be noted that both solutions using the different norms are *very* different: the repartition of the errors as well as their statistics is drastically changed.

3. COMPARISON BETWEEN LPC AND FBC

The linear prediction case has been investigated following the same lines as described above for filter banks. For the results presented in this paper, we used a third-order predictor (there is only a marginal gain beyond a third-order predictor):

$$\tilde{x_{n,m}} = h_1 x_{n-1,m-1} + h_2 x_{n,m-1} + h_3 x_{n-1;m}$$

where h_i are the predictor coefficients. This predictor is optimized according to the statistics of the image (see e.g. [6, 4]). Given this predictor, we choose the quantizer allowing a bit rate reduction and a reconstruction error within a given confidence interval.

An example of what can be done using this approach is shown in table 1 for two images. Again all results are provided in the context of "almost lossless coding", hence with reference to a threshold $t = 0.5$, with original images initially quantized on 8 bits per pixel (bpp). The first column provides the results obtained for $p = 100\%$ of errors in $[-0.5, 0.5]$, which corresponds to true lossless coding, since the image is perfectly obtained by requantizing the reconstructed signal on 8 bpp. The other columns show the variation of the bitrate by allowing more and more errors to exceed the chosen threshold, in a controlled manner, using the above methods.

Note that in both cases of linear prediction and filterbank coding, the observed percentages on the reconstructed image are almost equal to the required ones. Although the bit rate decrease is noticeable, it is not well improved when more errors are allowed, i.e. when the percentage $p\%$ decreases. A larger coding gain would be obtained only by working with a larger threshold t. However, this corresponds to an increase of the distortion, and it is well known that replacing the linear prediction by a filterbank may lead to better compression ratios in this case. This is clearly checked in table 1: LPC achieves better compression rate than FBC when lossless or near to lossless coding is required. When more loss is allowed ($p < 90\%$), the filter banks become more efficient.

4. CONCLUSION

The proposed approach has a very large flexibility:

- It allows to perform efficient *lossless* coding by simple means (see table 1).
- It allows a *local* control of the error. Varying α (the scale factor applied to the "optimal" quantization steps) according to the value of the signal allows *spatial masking* to be used.
- Whatever the application, any signal is measured with a given accuracy, hence defining a *confidence interval*. Our method is able to maintain the reconstruction errors in this confidence interval, thus resulting in practical lossless coding.

5. REFERENCES

[1] M. Antonini, "Transformée en Ondelettes et Compression Numerique des Images," phd Report, 1991.

[2] M. Antonini, M. Barlaud, P. Mathieu and I. Daubechies, Image Coding using Wavelet Transform, IEEE Trans. on Image Processing, Vol.1, NO. 2, pp. 205-220, April 1992.

[3] I. Daubechies, "Orthonormal Bases of Compactly Supported Wavelets," Comm. Pure Appl. Math, vol XLI, no. 7, pp. 909-294.

[4] A. Habibi, *Comparison of nth-Order DPCM Encoder with Linear Transformations and Block Quantization Techniques*, IEEE Trans. on Communication Technology, Vol. COM-19, NO. 6, Dec. 1971.

[5] M. Loève, *Probability Theory* , The university series, Princeton, NJ, 1960, chap. 6.

[6] M. Rabbani and P.W. Jones, " Digital Image Compression Techniques," SPIE Optical Engineering Press: Bellingham, Washington, USA, 1991.

[7] K. Ramchandran A. Ortega and M. Vetterli, "Best Wavelet Packet Bases in the Rate Distortion Sense," IEEE Trans. Image Processing, vol. 5, pp. 381-384, Apr. 1993.

[8] A.B. Sripad and D.L. Snyder, *A Necessary and Sufficient Condition for Quantization Errors to be Uniform and White*, IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. ASSP-25, NO. 5, Oct. 1977.

[9] D.A. Wismer and R. Chattergy, *Introduction to Nonlinear Optimization*, North Holland New York, 1978, (Appendix on linear programming).

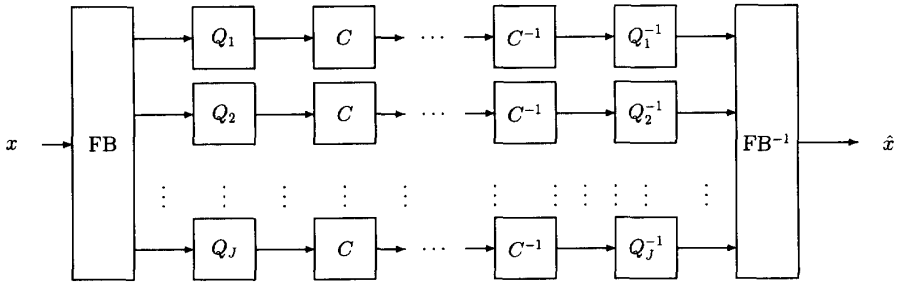

Figure 2: Filter bank compression scheme. The encoding part uses an octave-band separable filter bank iterated J times on the lowpass filter which splits the input image into $(3J+1)$ subimages. These subimages are then quantized and losslessly encoded (using Huffmann coding). The synthesis part reconstructs the approximate signal $\hat{x}$.

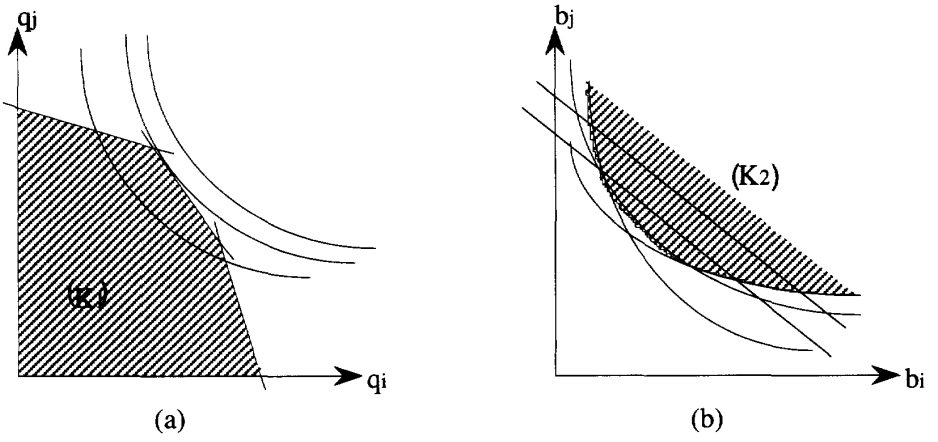

Figure 3: Equivalent constrained optimization problems. (a) Problem (P_1) has a convex cost function and linear constraints delimiting the feasible region (K_1). (b) Problem (P_2) has a linear cost function and convex constraints, the fesible region is (K_2). We easily see from $(P1)$ that there is an optimum solution which is on the boundary of (K_1) or (K_2).

LENA (512 × 512)							
percentage (%) of errors ≤ 0.5	required	100	99	95	90	85	80
	observed (LP)	100	98.99	94.98	90.06	84.79	79.95
	observed (FB)	100	99.30	95.30	89.90	84.63	79.40
GHC (bpp)	LP coding	4.64	4.62	4.56	4.48	4.39	4.31
GHC (bpp)	FB coding	5.61	5.01	4.61	4.37	4.18	4.03
Medical image (coronair, 256 × 256)							
percentage (%) of errors ≤ 0.5	required	100	99	95	90	85	80
	observed (LP)	100	99.03	95.01	90.35	85.51	80.09
	observed (FB)	100	99.01	95.32	89.89	84.56	79.36
GHC (bpp)	LP coding	2.89	2.88	2.83	2.76	2.69	2.60
GHC (bpp)	FB coding	3.76	3.13	2.82	2.61	2.43	2.31

Table 1: Almost lossless coding using filter banks (FB) or linear prediction (LP). We show the overall bitrates obtained after a Global Huffmann Code (GHC), for varying percentages but a constant threshold $t = 0.5$. The FB method here uses 5 iterations and the 12-tap Daubechies filter [3]. For $p = 100\%$, the deterministic criterion of 2.3.2 is used. The LP method uses three neighbor pixels predictor (see text).