

Image coding with an L∞ norm and confidence interval criteria

Lamia Karray, Pierre Duhamel, Olivier Rioul

▶ **To cite this version:**

Lamia Karray, Pierre Duhamel, Olivier Rioul. Image coding with an L∞ norm and confidence interval criteria. IEEE Transactions on Image Processing, 1998, 7 (5), pp.621-631. 10.1109/83.668016 . hal-03330179

HAL Id: hal-03330179

https://telecom-paris.hal.science/hal-03330179

Submitted on 10 Aug 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Image Coding with an L^∞ Norm and Confidence Interval Criteria

Lamia Karray, *Student Member, IEEE,* Pierre Duhamel, *Senior Member, IEEE,* and Olivier Rioul, *Member, IEEE*

***Abstract*— A new image coding technique based on an L^∞-norm criterion and exploiting statistical properties of the reconstruction error is investigated. The original image is preprocessed, quantized, encoded, and reconstructed within a given confidence interval. Two important classes of preprocessing, namely linear prediction and iterated filterbanks, are used. The approach is also shown to be compatible with previous techniques.**

The approach allows a great flexibility in that it can perform lossless coding as well as a controlled lossy one: specifications are typically that only *p%* the reconstructed pixels are different from the original ones.

***Index Terms*— Confidence interval criterion, filterbanks, lossless image compression.**

I. Introduction

TRANSMISSION and storage of digital signals both require some compression of the original data to reduce the overall system cost. In many applications, the image quality is reduced by the coding process: the image is converted into a set of binary digits in such a way that the original data can be recovered from the binary set within some distortion level. These "lossy" image compression schemes are usually based on some "transform coding" technique. The original image is split into a set of coefficients using some invertible transformation in order to improve the signal statistics prior to encoding. This transformation can be of several kinds: discrete cosine transform (DCT), filterbanks [1], wavelets [2], and so on. Linear prediction can also be considered as a particular transformation [3]. This first step (transformation) does not produce any data compression, but prepares the next steps. The transform coefficients are then quantized (which produces some data compression, but also 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, etc.) so that the overall compression ratio is minimized for a given reconstruction error measure. Classical lossy compression schemes basically minimize an MSE-like criterion on the reconstruction error [4]. Thus, the overall distortion is evaluated and controlled using the L^2-norm. However, this criterion is global (on the whole image) and does not exploit "local" knowledge, which is always available. One at least knows the number of bits on which each pixel of the original image is encoded. Moreover, in many applications (biomedical, for example), 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 on the whole image. In this paper, we propose a set of methods that are able to take such knowledge into account through a new criterion involving a "confidence interval." Therefore, our algorithms allow a pixel-by-pixel control of the error. We restrict ourselves in this paper to the case of very good quality (almost lossless) image coding, giving only a few indications on visual differences between L^2 and L^∞ coding.

More specifically, we determine the components in the compression scheme in order to ensure that a given percentage p of the reconstruction errors are smaller than some given threshold t. This is obtained without *a priori* knowledge of the image statistics, but some knowledge on the distortion statistics is required. Since the transform maps the initial discrete signal into an "almost continuous" one, a statistical model, depending on the transform, can be introduced and the statistical approach above becomes meaningful. As an example, we are able to optimize a compression scheme such that, e.g., say only 2% of errors exceed half the initial quantization step. That is, after requantizing (with the initial quantization step), only 2% the output pixels will be different from the original ones.

As a preprocessing, we consider two classical types of transformations: linear prediction and filterbanks. In each case, the original image x is transformed, quantized, encoded, and reconstructed to give $\hat{x}$. The distortion $\Delta x = x - \hat{x}$ is statistically modeled in order to be controlled accurately. In the chosen compression schemes, either prediction or filterbanks would achieve perfect reconstruction if no quantization was involved. Thus, the reconstruction error is due to quantization only.

The classical assumption made on quantization noise and transform coefficients are recalled in Section II. Since reconstruction errors depend only on quantization, the problem amounts to determining the quantizers in the compression scheme so as to achieve $p\%$ the values of the error $|\Delta x|$ inside the confidence interval which for simplicity is chosen of constant amplitude, depending on a given threshold t. Hence,

L. Karray is with the Department of Research in Speech Communication, CNET, Lannion, France.

P. Duhamel is with the Signal and Image Processing Department and URA CNRS 820, Ecole Nationale Supérieure des Télécommunications, 75634 Paris, France.

O. Rioul is with the Department Communications and URA CNRS 820, Ecole Nationale Supérieure des Télécommunications, Paris, France (e-mail: rioul@com.enst.fr).

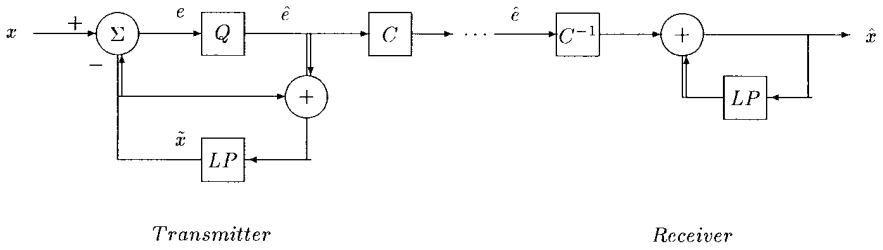


Fig. 1. DPCM block diagram. The transmitter includes the receiver to ensure identical prediction on both sides. LP denotes the predictor. The prediction error is quantized, encoded (global Huffmann coding) and transmitted. The receiver reconstructs the approximate signal $\hat{x}$.

we minimize the overall bit rate under the constraint

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

This problem is solved in the context of a predictive coding scheme in Section III, and in the context of wavelet transforms (iterated filterbanks) as a subband coding in Section IV. Section V-A provides a comparison between both techniques in the context of near lossless coding.

Compatibility with other lossless coding techniques is also examined in Section VI, showing that lossy plus lossless coding [3] can also be used with same performances if the lossy part is optimized using an L^∞ criterion. Finally, we give some indications on the visual differences between low bit rate systems optimized using L^2 and L^∞ norm criteria.

II. Preliminaries

The processing (transform) applied on the original signal in the compression scheme is introduced in order to improve the statistics of the coder input. Moreover, a fact that is seldom used, it also increases the density of the original discrete data and allows an almost continuous modeling of the statistical distribution of the reconstruction errors.

In the case of a predictive preprocessing, each pixel is predicted by a linear combination of its neighbors. The transmitted data is the prediction error.

For filterbanks, the original image is split into different subimages. Each subimage has a reduced bandwidth compared to the original full-band image. Furthermore [7], subband signals statistics can often be accurately modeled by Laplacian distributions.

After transformation, the resulting signals y are quantized prior coding. Here, we consider only scalar uniform quantization with stepsize q. The quantization error can be modeled as an additive noise lying between $-q/2$ and $q/2$. The input y is supposed to be Laplacian, with zero-mean and standard deviation σ. As shown in [6], if the ratio σ/q is high (more than 0.7), the quantization noise can be accurately modeled by a uniform distribution

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

with zero mean and variance $q^2/12$.

In our case (almost lossless coding), $q \leq 2$ and the standard deviation of the quantizer input signal is larger than one. So, the above condition is satisfied and the quantization error is always supposed to be uniformly distributed.

III. Coding with Linear Prediction

Linear predictive coding is particularly well adapted to lossless or high-quality coding, and easily implemented. Therefore, it is the first logical choice in our context. In a general predictive coding scheme, the correlation between the neighboring pixel values is used to compute the linear predictor which, when applied on the image, provides a prediction for each pixel. A well-known approach to predictive coding is differential pulse code modulation (DPCM). In DPCM, the prediction $\tilde{x}$ is subtracted from the actual pixel value x to form a differential image that is much less correlated than the original data and can be assumed to have a Laplacian distribution [3], [5].

The predictive error is then quantized and encoded. The quantized differential image is transmitted to reconstruct the pixel values. Since the quantization of the differential image introduces errors, the reconstructed values typically differ from the original ones. To ensure that identical predictions are formed at both the receiver and the transmitter, the transmitter also bases its prediction on the reconstructed values. This is accomplished by introducing the quantizer within the prediction loop at the analysis and the reconstruction side as shown in Fig. 1. In essence, the transmitter includes the receiver within its structure.

This compression scheme uses a linear predictor, scalar uniform quantization, and lossless coder. Since the differential image has a Laplacian distribution, the probability of occurrence of the various quantized levels is not uniform and the average number of binary digits required for their representation can be reduced by using some variable length coder. Here we use global Huffmann coding.

The predictor and the quantizer have to be chosen in such a way that the bit rate is minimum while the "confidence interval" constraint (1) is satisfied. Because the inclusion of the quantizer in the prediction loop results in a complex dependency between the prediction error and the quantization error, a joint optimization should ideally be performed. However, to avoid the complexity of modeling such interactions, both components are usually optimized separately.

A. Predictor Optimization

The classical criterion for computing the linear predictor is the minimization of the mean-squared prediction error [3], [5].

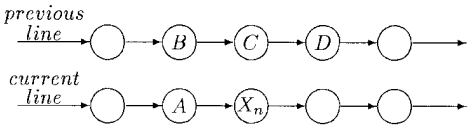


Fig. 2. DPCM predictor configuration. The pixel X_n is predicted using the neighboring pixels; e.g., for a third-order predictor, the prediction of X_n is a combination of pixels A, B, and C [3].

Under this criterion, the best linear estimate of a sample x_n is the value $\tilde{x}_n$ that minimizes

$$\sigma_e^2 = E[(x_n - \tilde{x}_n)^2]. \tag{2}$$

If $\tilde{x}_n = \sum_{i=1}^{N} h_i\, x_{n-i}$, where N is the predictor order and h_i the coefficients to be optimized, the minimization of criterion (2) yields to the following set of linear (Wiener–Hopf) equations:

$$\sum_{i=1}^{N} h_i\, R_{i,\,j} = R_{0,\,j} \qquad j = 1,\, 2,\, \cdots,\, N$$

where $R_{i,\,j} = E[x_i\, x_j]$. The optimal predictor coefficients are then obtained by solving the linear system above.

Regarding the predictor order, which means the number of samples involved in the weighting, experimental studies on various images show that there is only a marginal gain beyond a third-order predictor. In this paper, we use a third-order predictor of the form

$$\tilde{x}_n = h_1 A + h_2 B + h_3 C$$

where h_i are the predictor coefficients; A, B, and C are the neighboring points shown in Fig. 2.

Recall that the full scheme in Fig. 1 would achieve zero error if no quantizer was introduced. The reconstruction error $\hat{x} - x$ is thus fully determined by the quantizer, which has to be optimized as described below.

B. Quantizer Optimization

Given an optimal predictor, the next step is to choose the quantizer that minimizes the bit rate under the constraint that the reconstruction error lies within the confidence interval. To solve this problem, a statistical model of the reconstruction error is required. Using the compression scheme of Fig. 1, and assuming no channel errors, we easily find that $\Delta x = x - \hat{x} = e - \hat{e}$ which is the quantization error. Since the quantizer is uniform, with a step q, the quantization error is, under the conditions recalled in Section II, uniform in $[-q/2,\, q/2]$. The reconstruction error has consequently the same uniform distribution. Equation (1) can then be rewritten as

$$\int_{-t}^{t} \frac{1}{q}\, w_q(x)\, dx \geq p\% \tag{3}$$

where

$$w_q(x) = \begin{cases} 1, & \text{if } \dfrac{-q}{2} \leq x < \dfrac{q}{2} \\ 0, & \text{otherwise} \end{cases}.$$

Note that if $q < 2t$, the quantization error and consequently the reconstruction error are always less than the given threshold. Since condition (3) allows a certain percentage of errors to exceed t, the quantization step meeting (1) is larger than $2t$. Equation (3) is easily solved in t, yielding $\int_{-t}^{t} (1/q)\, dx = 2t/q \geq p\%$. The required value of the quantization step is thus $q = 2t/p$.

C. Results

An example of what can be done using this approach is shown in Table II for two images. 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 b/pixel. The first column provides the results obtained for $p = 100\%$ of errors in $[-0.5,\, 0.5]$. This corresponds to plain lossless coding using DPCM, since the image is exactly reconstructed by requantizing the reconstructed signal on 8 b/pixel. As is well known, this approach allows a noticeable reduction of the bit rate that is necessary for transmission (58% the initial bit rate is necessary for Lena, 36% for coronair). The other columns show the variation of the bit rate by allowing more and more errors to exceed the chosen threshold, in a controlled manner, by using the above approach.

We observe that the percentages computed on the reconstructed image are almost equal to the required ones, thus verifying the white noise assumption. However, although the bit rate decrease is noticeable, it does not decrease quickly when more errors are allowed, i.e., when the percentage $p\%$ decreases (allowing 20% pixels to be recovered with an error has allowed the bit rates to be further decreased of 8% and 11%, respectively). 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. In the following, filterbanks are introduced in the same framework.

IV. Filterbanks Coding

Since the efficiency of DPCM decreases when more distortion is allowed in the image compression scheme, filterbanks coding (FBC) is now considered.

The corresponding compression scheme is depicted on Fig. 3. It uses a perfect reconstruction octave-band filterbank, scalar uniform quantization, and global Huffmann coders. For a given set of perfect reconstruction filters, the problem is that of determining the best quantizers minimizing the global bit rate while a certain percentage of reconstruction errors lies within a given confidence interval, as in (1). However, the problem is much more difficult in this case, as will be clear below. An evaluation of the statistical properties of the reconstruction error is required in order to solve this problem.

A. Statistical Properties of the Reconstruction Error

Since we use perfect reconstruction filters and lossless coders, the error is generated by the quantizers only. Its distribution is thus a function of the quantization noise. Therefore, we first need to find the dependency between both signals.

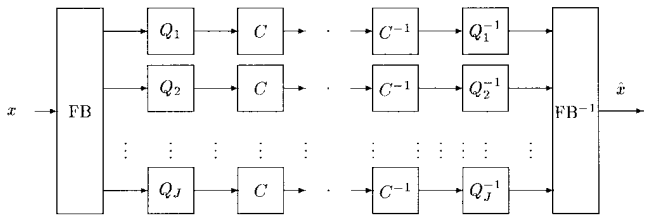


Fig. 3. Filterbank compression scheme. The encoding part uses an octave-band separable filterbank (wavelet transform) iterated J times on the lowpass filter which splits the input image into $(3J+1)$ subimages. These subimages are then quantized and losslessly encoded (using Huffmann coding). The synthesis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

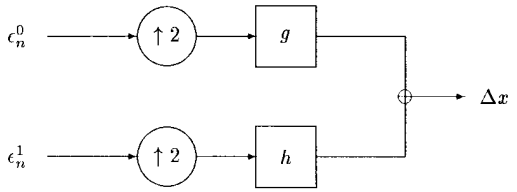


Fig. 4. Reconstruction error. Only one iteration of a 1-D signal is considered. ϵ_n^0 and ϵ_n^1 are quantization errors. Δx is the reconstruction error, g and h are the set of the filter coefficients.

Since the system is 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 structure depicted in Fig. 4 in the context of one iteration of a one-dimensional (1-D) signal. For simplicity, we show in this 1-D context that the reconstruction error has a Gaussian distribution. The demonstration would be similar in the two-dimensional (2-D) case.

Due to the interpolation in the synthesis phase, the odd and even samples of Δx depend on a different set of filter coefficients. So, two cases have to be considered:

$$\Delta x_{2n} = \sum_{k=0}^{(N_g/2)-1} \epsilon_{n-k}^0 g_{2k} + \sum_{k=0}^{(N_h/2)-1} \epsilon_{n-k}^1 h_{2k}$$
$$\Delta x_{2n+1} = \sum_{k=0}^{(N_g/2)-1} \epsilon_{n-k}^0 g_{2k+1} + \sum_{k=0}^{(N_h/2)-1} \epsilon_{n-k}^1 h_{2k+1} \quad (4)$$

where N_g (resp., N_h) is the filter g (resp., h) length.

The quantization errors ϵ^i have, as seen in Section II, a uniform distribution in $[-q_i/2, q_i/2]$ with zero-mean and variance $q_i^2/12$.

The reconstruction error is thus a linear combination of uniform distributions. Because of the filter coefficients and the different values of the quantization steps in the various subbands, these uniform distributions have different bandwidths and different energies. Thus, to prove that the reconstruction error converges to a Gaussian distribution, some conditions on its first three moments have to be satisfied in order to use central limit ramifications developed in [8]. These conditions are that the mean and the third moment should be zero, while the variance should take finite value. This is checked below.

Consider one of the subbands of the synthesis phase, e.g., the lowpass one of Fig. 4, denoted e. We have $e_{2n+u} = \sum_k \epsilon_{n-k}^0 g_{2k+u}$ where $u = 0$ for even samples and $u = 1$ for odd samples. We easily find that e_{2n+u} has a zero-mean and a variance $\sigma_u^2 = E[(\sum_k \epsilon_{n-k}^0 g_{2k+u})^2]$.

Since quantization is scalar, we can assume that quantization errors are independent. We further assume that the quantization errors in different subbands are also independent. All these conditions can be shown to hold under the same hypothesis as the one recalled in Section II for the quantization noise to be uniform. Under these conditions, we have

$$E[\epsilon_{n-k}^i \epsilon_{n-l}^i] = \begin{cases} \dfrac{q_i^2}{12}, & \text{if } k = l \\ 0, & \text{otherwise} \end{cases}$$

and

$$\sigma_u^2 = \frac{q_i^2}{12} \sum_k g_{2k+u}^2.$$

Note that the variance has a finite value (since we use finite length filters). Under the independence hypothesis, it can be shown that the third moment is null: $\gamma_u = E[e^3] = 0$.

The results above are easily extended to 2-D signals and many iterations. We can thus use generalized central limit theorems detailed in [8] and [9] to show that an infinite sum of those errors has a Gaussian distribution.

In practice, the filterbanks are iterated several times, which increases the number of subbands, since for J iterations of a 2-D signal we have $3J+1$ subbands. Thus, e.g., for $J = 5$ iterations we have 16 subbands and the convergence to a Gaussian distribution is ensured.

Finally, we end up with a Gaussian distribution reconstruction error, with a zero-mean, and a variance σ^2 depending on quantization step values and on the set of filter coefficients involved in each subband. For example, for $J = 1$ we have

$$\sigma_{u,v=0,1}^2 = \frac{q_0^2}{12} \sum_{k,l} g_{2k+u}^2 g_{2l+v}^2 + \frac{q_1^2}{12} \sum_{k,l} h_{2k+u}^2 g_{2l+v}^2 + \frac{q_2^2}{12} \sum_{k,l} g_{2k+u}^2 h_{2l+v}^2 + \frac{q_3^2}{12} \sum_{k,l} h_{2k+u}^2 h_{2l+v}^2$$

where u and v are associated to odd and even samples of the error.

For J iterations, the whole error is a contribution of the different samples $(\Delta x_{u,v})_{u,v=0,\cdots,2^J-1}$ due to the J interpolations. Furthermore, these samples have the same contribution in the whole signal, and a probability density $f_{u,v}(x) = 1/(\sigma_{u,v}\sqrt{2\pi}) e^{-(x^2/2\sigma_{u,v}^2)}$.

B. Quantizers Optimization

Since a statistical model of the reconstruction error is available, a control of this error is feasible. So, we can find the quantizers allowing a certain percentage of distortion such as in (1).

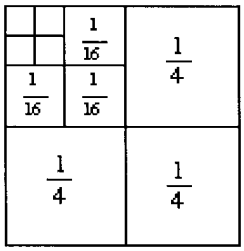


Fig. 5. Contribution of the various subbands. The lowpass subimage is transformed to give four other subimages four times smaller due to the decimation in the analysis phase. The contribution of each subimage in the global bit rate is thus divided by four in each iteration.

Taking into account the uniform contribution of the various $(\Delta x)_{u,v}$, we have

$$\begin{aligned}\text{prob}\{|\Delta x| \leq t\} &= \frac{1}{2^{2J}} \sum_{u,v} \text{prob}\{\Delta x_{u,v} \leq t\} \\ &= \frac{1}{2^{2J}} \sum_{u,v} \text{erf}\left(\frac{t}{\sigma_{u,v}\sqrt{2}}\right) \end{aligned} \tag{5}$$

where $\text{erf}(a) = 2/(\sqrt{\pi}) \int_0^a e^{-x^2}\, dx$.

Recall that $\sigma_{u,v}^2$ is a function of the squared quantization steps. Besides, the optimal quantization steps satisfying $\text{prob}\{|\Delta x| \leq t\} > p\%$ must also minimize the global bit rate, which is the sum of the bit rates in the various subbands. Note also that the decimation in the analysis phase divides by two the size of the subimage in each subband and since we use separable dyadic filters, the contribution of a subimage resulting from j iterations is $n_i = 1/2^{2j}$. This is illustrated in Fig. 5.

Let b denote the total bit rate, r_i the dynamic range in subband number i, and b_i the bit rate in subband i. For simplicity, we evaluate the contribution b_i of subband i as a function of quantization step q_i by

$$q_i = r_i 2^{-b_i}. \tag{6}$$

Although this is only a rough evaluation of the actual bit rates, which are much lower after Huffmann coding, this formula has been chosen in order to be able to perform the optimization. Despite this approximation, the results are quite accurate, as will be seen in Section V. This is certainly due to the fact that, for such high-precision coders, the bit rate evaluated by formula (6) has been observed to be almost linear in terms of the Huffman bit rate. Quantization steps have to be chosen such that the total bit rate, $b = \sum_i n_i b_i$, is minimum while (5) is satisfied. Thus, we end up with a constrained minimization problem

$$(P): \begin{cases} \min \left[\sum_i n_i \log_2\left(\frac{r_i}{q_i}\right)\right] \\ \frac{1}{2^{2J}} \sum_{u,v} \text{erf}\left(\frac{t}{\sigma_{u,v}\sqrt{2}}\right) \geq p\% \end{cases}. \tag{7}$$

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 (deterministic problem). Then, using the relationship between the variance of the error and the quantization steps, we solve (1) (statistical problem).

C. 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., $|\Delta x_n| \leq t\ \forall n$. This condition can be written using the maximum of the reconstruction error, also called L^∞-norm: $||\Delta x||_\infty \leq t$. We finally have to solve the constrained minimization problem

$$\begin{cases} \min\left(\sum_i n_i b_i\right) \\ ||\Delta x||_\infty \leq t \end{cases}. \tag{8}$$

An evaluation of the maximum reconstruction error is thus required.

1) Estimation of the Max Norm of the Distortion: Since the exact value of the maximum depends on the filter coefficients and 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 two-band filterbank. The result is then extended to 2-D signals and J iterations.

According to (4), an upper bound is found for each kind of distortion samples [$(\Delta x_{2n})_{\max}$ for even samples and $(\Delta x_{2n+1})_{\max}$ for odd ones]. The L^∞-norm estimate is then the maximum of both:

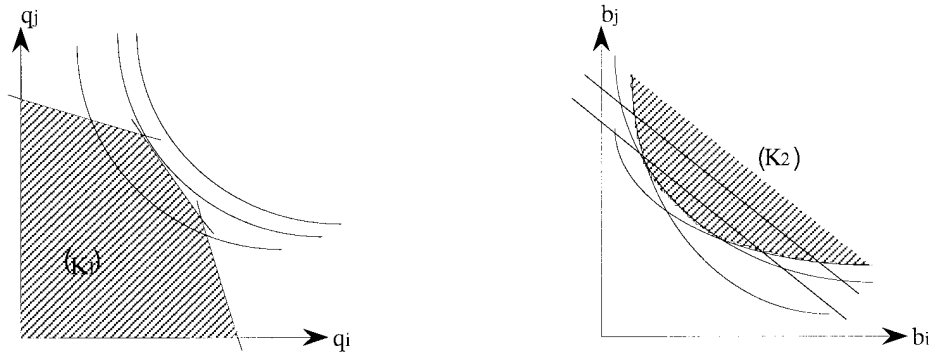
$$||\Delta x||_\infty \leq \max \begin{cases} (\Delta x_{2n})_{\max} \\ (\Delta x_{2n+1})_{\max} \end{cases}.$$

An upper bound is obtained if the quantization errors (ϵ^0 and ϵ^1) have maximum amplitude and have the same sign as the filter coefficients. Since ϵ^i satisfies $|\epsilon^i| \leq q^i/2$, 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 filterbanks are involved, the filtering and interpolation are separately done on lines then columns of the 2-D signals, which results in four types of samples, depending on the parity of their indexes, giving

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



(a) (b)

Fig. 6. Equivalent constrained optimization problems. (a) Problem (P_1) has a convex cost function and linear constraints delimiting the feasible domain (K_1). (b) Problem (P_2) has a linear cost function and convex constraints, the feasible domain is (K_2). It is easily seen from $(P1)$ the optimum solution is on the boundary of (K_1) or (K_2).

After J iterations of the filterbank we similarly obtain

$$\|\Delta x\|_\infty \leq \max_{u,\,v}\{\max_{m,\,n}(\Delta x_{2^J m+u,\,2^J n+v})\}$$

where u and v vary between 0 and 2^J-1 (i.e., $2^J \times 2^J = 2^{2J}$ possibilities, coming from the J successive interpolations in the synthesis phase) and m and n describe the different signal samples.

Without any *a priori* knowledge on the signal, this estimator is optimal, but overly pessimistic for many real images. Furthermore, it is a linear function of the quantization steps.

Since Δx is the contribution of the interpolated samples $\Delta x_{u,\,v}$, the relation $\|\Delta x\|_\infty \leq t$ has to be satisfied for every $(u,\,v)$ combination, yielding a linear system of constraints. Thus, the optimization problem (8) has a convex criterion: $\sum_i n_i \log_2(r_i/q_i)$, using the notations of (6), and linear constraints $\sum_i a_{ij} q_i \leq t_j$ where coefficients a_{ij} depend on the set of filter coefficients involved in subband i and the combination j of u and v seen above.

2) Deterministic Computation of the Quantization Steps: Finding the optimal quantization steps minimizing the bit rate such that the reconstruction error is at most equal to the threshold t is thus equivalent to solving the constrained optimization problem

$$(P_1)\colon \begin{cases} \min\left(\sum_i -n_i \log_2 q_i\right) \\ \sum_i a_{ij} q_i \leq t_j \qquad j=1,\cdots,2^{2J} \end{cases} \tag{9}$$

with convex criterion and linear constraints [see Fig. 6(a)].

As illustrated in Fig. 6, we can prove that there is a unique solution which is on the boundary of the feasible domain. In order to prove it, consider for instance Fig. 6(a). The feasible domain $(K1)$ is a convex polyhedron. It is also closed, nonempty, and bounded. So, using general theorems, we conclude that the problem has an optimal solution. Then, we can easily show, *ab absurdo*, that the optimum is unique and on the boundary of the feasible domain (more detailed proof is available upon request).

Now, in actual problems, the number of constraints is huge (over 1000 constraints when the lowpass filter is iterated five times in two dimensions), which makes the use of general optimization procedures very difficult. 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 (6), the problem (9) can also be written as follows:

$$(P_2)\colon \begin{cases} \min\left(\sum_i n_i b_i\right) \\ \sum_i \alpha_{ij} 2^{-b_i} \leq t_j \end{cases} \tag{10}$$

with linear criterion and convex constraints [see Fig. 6(b)].

These two problems are equivalent and linked by the function 2^{-x}. Since this function is convex and bijective, each vertex (Q_i) of (K_1) is associated to a "virtual" vertex (B_i) of (K_2) which saturates the same constraints, see Fig. 7. The feasible domain (D) determined by the constraints (B_i) is a convex polyhedron. Further, the minimization of the linear criterion on (D) gives the best vertex of (9); the optimal solution is on the boundary. Since the criterion is linear and (D) a convex polyhedron, the best vertex is obtained using the simplex algorithm [10], [11]. 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. We can thus solve the problem using the (9) version. 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 domain. Note that the further the vertices are from the axis, the better the approximation is. The obtained approximate solution is used in the second step as a starting point of the general optimization algorithm, to find the optimum solution. Its convergence is thus ensured and relatively fast.

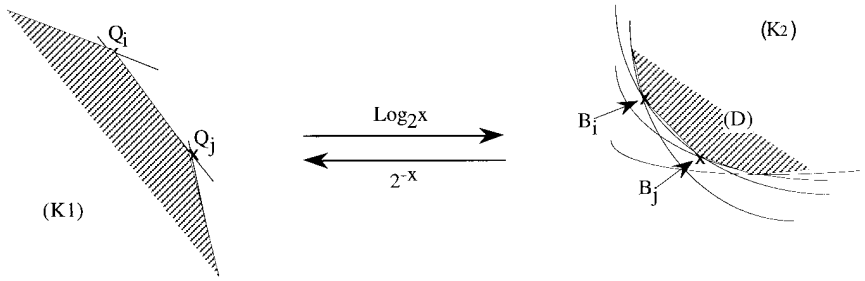


Fig. 7. Vertices in the equivalent constrained optimization problems. Each vertex (Q_i) of the feasible domain $(K1)$ is linked by the function $x \longmapsto \log_2 x$ to a virtual vertex (B_i) in $(K2)$. The virtual vertices of $(K2)$ are the vertices of the convex polyhedron (D).

TABLE I
BIT RATES (HUFFMANN CODING) OBTAINED BY FBC, AND REQUESTING THE MAX ERROR BE LOWER THAN $t = 0.5$. THE NUMBERS BETWEEN PARENTHESES SHOW THE ACTUAL MAXIMUM ERROR. THE LAST ROW HAS BEEN OBTAINED BY FORCING THE MAX ERROR TO BE EQUAL TO $t = 0.5$ BY A SCALE FACTOR ON THE QUANTIZATION STEPS

GHC (bpp)	LENA (512 × 512)	Medical (256 × 256) (Coronair)	Aerial (512 × 512) (Aerial view of Corse)
Original	7.47	6.27	3.95
Preliminary Results	7.17 (0.20)	4.96 (0.20)	3.44 (0.22)
Final Results	5.61 (0.5)	3.73 (0.5)	2.42 (0.5)

TABLE II
ALMOST LOSSLESS CODING USING DPCM. WE SHOW THE OVERALL BIT RATES OBTAINED AFTER A GLOBAL HUFFMANN CODING (GHC) FOR VARIOUS PERCENTAGES OF ERRORS EXCEEDING $t = 0.5$ (SEE FIG. 2)

LENA (512 × 512)							
percentage (%)	required	100	99	95	90	85	80
of errors ≤ 0.5	observed	100	98.99	94.98	90.06	84.79	79.95
GHC (bpp)		4.64	4.62	4.56	4.48	4.39	4.31
Medical image (coronair, 256 × 256)							
percentage (%)	required	100	99	95	90	85	80
of errors ≤ 0.5	observed	100	99.03	95.01	90.35	85.51	80.09
GHC (bpp)		2.89	2.88	2.83	2.76	2.69	2.60

TABLE III
ALMOST LOSSLESS CODING USING FBC. WE SHOW THE OVERALL BIT RATES OBTAINED AFTER A GLOBAL HUFFMANN CODING (GHC) FOR VARIOUS PERCENTAGES OF ERRORS EXCEEDING $t = 0.5$, THE 12-TAP DAUBECHIES FILTER [12]) IS ITERATED FIVE TIMES. FOR $p = 100\%$, THE DETERMINISTIC CRITERION OF IV-C2 IS USED

LENA (512 × 512)							
percentage (%)	required	100	99	95	90	85	80
of errors ≤ 0.5	observed	100	99.30	95.30	89.90	84.63	79.40
GHC (bpp)		5.61	5.01	4.61	4.37	4.18	4.03
Medical image (coronair, 256 × 256)							
percentage (%)	required	100	99	95	90	85	80
of errors ≤ 0.5	observed	100	99.01	95.32	89.89	84.56	79.36
GHC (bpp)		3.76	3.13	2.82	2.61	2.43	2.31

However, note that the constraints have been taken into account through an upper bound of the max norm of the reconstruction error. As already pointed out, this upper bound of the L^∞-norm is not achieved in actual images, hence the obtained maximum error is usually less than the required threshold t (an order of magnitude of the observed threshold is $t/2$). Based on the linear dependency between the quantization steps and the upper bound, we apply some scale factor to all quantization steps in order to bring the reconstruction error up to the required threshold t. Note that this is only an approximation, since the dependency between the actual threshold and the quantization steps is not linear. The scale factor is tuned using an iterative optimization procedure. In Table I we show the difference between Huffmann coding applied to the original image, applied to the image obtained by using the quantization steps as provided by the constrained optimization procedure (the observed maximum error is seen to be about 0.2, while 0.5 was requested), and after applying the scale factor to the quantization steps.

So far, we have determined optimal quantization steps according to the deterministic approach in (9). As it is, the proposed solution 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. We recover the statistical problem (8).

D. Statistical Computation of the Quantization Steps

Recall that the quantization steps depend linearly on the upper bound of the error, which is close to t. Hence, if the threshold t is multiplied by some constant α, the steps q_i should also (almost) be multiplied by α and the variance σ^2 of the error by α^2. Therefore, it is natural to state the statistical problem relatively to the one solved in Section IV-C2: the new quantization steps are chosen as $q_i' = \alpha q_i$. Since the reconstruction error has been shown to be nearly Gaussian, α is easily evaluated by computing the enlargement factor of the Gaussian such that $p\%$ of the error is larger than the threshold t. This corresponds to the relation

$$\operatorname{erf}\left(\frac{t}{\alpha\sigma\sqrt{2}}\right) \geq \frac{p}{100}. \tag{11}$$

Taking into account the different samples $\Delta x_{u,v}$, we find α by solving the equation

$$\frac{p}{100} = \frac{1}{2^{2J}} \sum_{u,v} \operatorname{erf}\left(\frac{t}{\alpha\sigma_{u,v}\sqrt{2}}\right) \tag{12}$$

where $\sigma_{u,v}$ is the standard deviation associated to $\Delta x_{u,v}$ and function of the optimal quantization steps solution of the deterministic problem (9). Since (12) is nonlinear, it is solved using a general algorithm. However, since it depends on a single variable α, solutions are reliable and fast.

E. FBC Used as an "Almost Lossless Coder"

A direct application of the above scale factor α to the quantization steps q_i computed as described in Section IV-C2, gives the results shown in Table III. These results are

obtained with the same context of "almost lossless coding" as the linear prediction approach. For simulations, we use a 12-tap Daubechies filter [12] iterated five times. This choice is experimentally proved to give the best results in terms of compression gain among all Daubechies filters, and any possible number of iterations.

The first column is obtained in lossless context ($t < 0.5$) using the optimal quantizers according to the L^∞-criterion. For the other columns of Table III, we note that the bit rate reduction is more and more important when p decreases. The compressed image quality is, however, high and sufficient for some applications where some details can be ignored. The important point, however, is that the distance to lossless is controlled through the percentage of pixels outside the confidence interval.

Furthermore, it is observed that there is a good agreement between the required and the observed percentages, which proves the efficiency of the statistical Gaussian model of the reconstruction error.

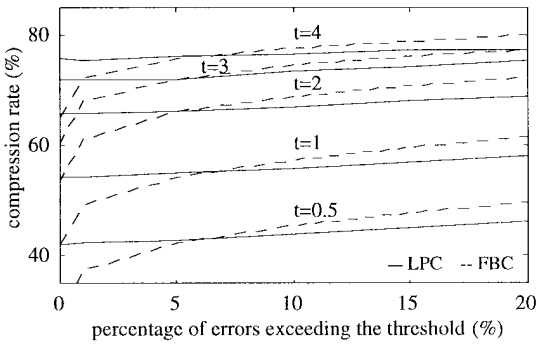


Fig. 8. Comparison between DPCM (solid lines) and FBC (dashed lines). We give the results obtained after a global Huffmann code (GHC) after linear prediction, or after a wavelet transform for various thresholds. This is done for different percentages of errors exceeding a given threshold. We show the compression ratio (defined as 100 times the initial bit rate divided by the one, after compression). DPCM achieves higher compression ratios for small percentages of errors, while FBC performs better when more errors are allowed.

V. Comparison Between Various Lossless Techniques

A. Comparison Between DPCM and FBC

As shown in Tables II and III, DPCM achieves better compression rate than FBC when lossless or near to lossless coding is required. When more loss is allowed ($p < 90\%$), the filterbanks become more efficient.

This observation has been confirmed by using other threshold values. For a given threshold t', we evaluate the compression rates obtained with various percentages of errors exceeding t'. The results obtained with different values of the threshold are summarized in Fig. 8. Note that the property mentioned above is maintained, i.e., FBC gives better results for larger confidence intervals (when the percentage p, of errors less than a given threshold, decreases). However, note that the threshold value must be rather small so that the assumption required for the statistical modeling of the distortions [see (2)] are valid. This means that the statistical confidence interval criterion only holds in the context of near lossless coding.

As for the deterministic approach, which does not need any modeling, the threshold value is also limited according to the relationship (6) between the bit rate, the dynamic range and the quantization steps. The quantization steps can not exceed the dynamic range values. So, in the context of very high compression rates, we have to consider additional constraints ($q_i \leq r_i$) in the optimization problem of Section IV-C2. These constraints depend on the original image. Recall that the whole optimization procedure is otherwise signal independent.

VI. Comparison to Other Classical Lossless Techniques

A number of lossless coding techniques for continuous-tone images have appeared in the literature [3], [13]. These techniques are based on such concepts as runlength coding, bit plane processing, predictive coding, etc.

The simplest techniques consist in applying lossless coders on the original image. These methods are not usually very efficient in terms of compression ratio because the 2-D pixel intercorrelations are not easily taken into account. Hence, several improvements of the statistics of the coders' inputs are made by preprocessing the original signal. Linear prediction is an example of those *hybrid techniques*, as explained above.

An other example, useful when short transmission time is required, is the *lossy plus lossless residual coding* [3]. This technique is based on the transmission of a lossy but high-quality version of the image, followed by a lossless encoding of the remaining difference (residual) image to achieve perfect reconstruction of the original image. In the lossy part of the scheme, lossy transform coding techniques, such as filterbanks, are used. The lossy part of the coder is usually optimized by an L^2-criterion. Performances in terms of lossless compression depend on the bit rate repartition between the lossy and the lossless part. In the simulations given below, the tradeoff between the bit allocation for the lossy part and the residual has been chosen in order to obtain the best total entropy.

Based on the results of this paper, we developed a transform-based technique able to achieve lossless coding using an L^∞-criterion. We described how to choose the quantizers in the FBC scheme to have a lossless compression. Using the same procedure, we can also perform lossy plus lossless coding, with a lossy coder optimized using an L^∞-norm. The threshold in the lossy L^∞ part of the coder has been chosen in such a way that the total entropy of the lossy plus lossless coder in minimized.

In Table IV we show the entropy bitrates obtained on the Lena image by these various schemes. The lossy plus lossless coders have been tuned such that the amount of loss (either in an L^2 or L^∞ sense) results in a minimum for the global entropy. Note that, if no preprocessing is applied to the original image, lossless coders are not very efficient, since only a compression gain of (1.07) is achieved. Plain lossless FBC (iterated filterbank) follows.

This gain improves with DPCM, but it is clearly seen that lossy plus lossless compression schemes still have better performances. When the lossy part is optimized with an L^∞-

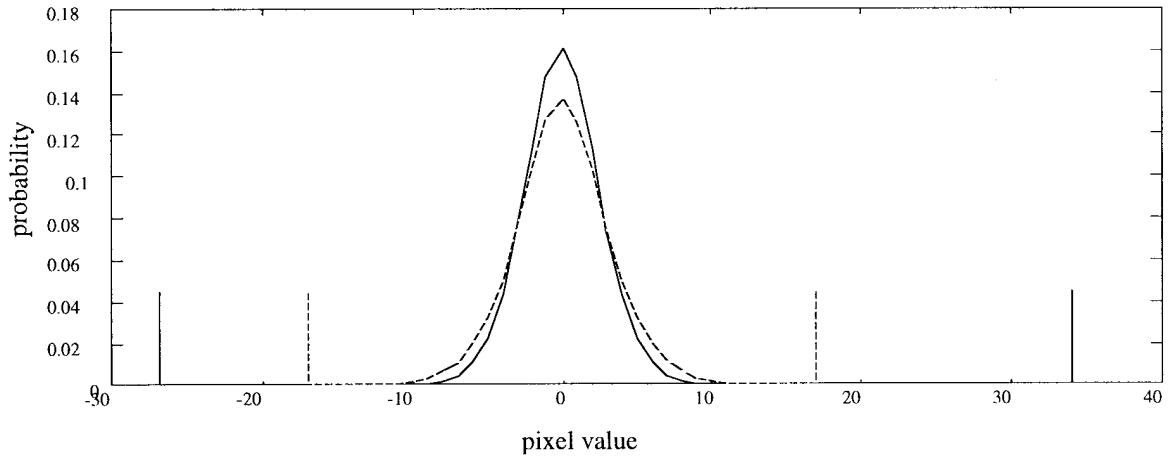


Fig. 9. Histogram of the errors with L^2 and L^∞ criteria in the reconstructed Lena images at the bit rate of 1.4 b/pixel (given by a global Huffmann coding) using the MSE-criterion (solid curve) and the L^∞-criterion (dashed curve). The vertical lines (dashed lines for L^∞ and solid lines for L^2) show the minimum and maximum error values in both cases. As expected, the maximum error is higher with L^2-criterion than with L^∞-criterion.

TABLE IV
COMPARISON BETWEEN VARIOUS LOSSLESS CODING TECHNIQUES IN TERMS OF ZERO-ORDER ENTROPIES. THE FIRST THREE LINES CORRESPOND TO CLASSICAL TECHNIQUES, WHILE THE LAST TWO ONES ARE EITHER THE PLAIN TECHNIQUES EXPLAINED IN THIS PAPER (LINE 4) OR MIXED TECHNIQUES (LAST LINE)

Lossless Coding Technique	bit rate (bpp)
zero-order entropy	7.45
DPCM	4.64
lossy (DWT) plus lossless residual coding (L^2-criterion)	4.58
lossless DWT (L^∞-criterion)	5.61
lossy (DWT) plus lossless residual coding (L^∞-criterion)	4.56

criterion, the best performances are obtained (4.56 b/pixel for the DWT lossy plus lossless coder versus 4.64 for the DPCM scheme).

It should, although, be noted that FBC performance is greatly improved in the context of lossy plus lossless residual coding. As seen in Section V, FBC is efficient when some loss is allowed. Moreover, the obtained residual is well decorrelated. This could explain this improvement.

Nevertheless, a conclusion about the best transform is not easy [13], since the choice of particular technique is not determined strictly by the achievable bit rate. Each strategy offers certain features that may be useful to meet side requirements that might exist in a particular environment.

VII. THE L^∞-CRITERION IN THE CONTEXT OF LOW BIT RATE CODING

As seen above, the L^∞-norm criterion allows a local control of the distortion in such a way that the difference between the reconstructed and the original image does not exceed a given threshold. Typically, although transform and quantization are used, we can perform lossless compression when the threshold is half the initial quantization step.

However, these schemes also have the potential for performing low bit rate coding with a different criterion than the classical L^2-norm. The question is how do these two criteria compare in this context.

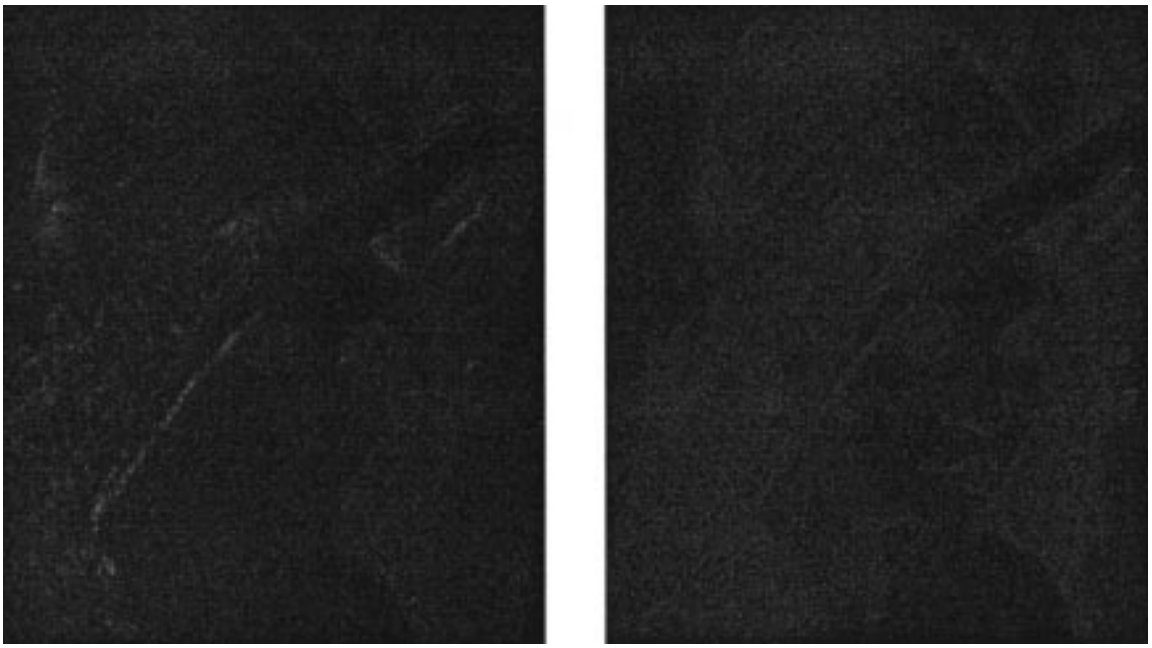
In order to compare both L^∞ and L^2 (MSE) criteria, a full coding and decoding of images was implemented. In the MSE approach, the choice of the quantizers is based on the minimization of the L^2-norm of the reconstruction error, as explained in [4]. However, the structure of the filterbank is kept fixed. The procedure is as follows: For a given image, we first apply the L^∞-criterion technique to obtain a certain bit rate. Then, using L^2-criterion, we choose the quantizers minimizing the MSE for the same bit rate. The reconstructed images as well as the reconstruction errors are compared using objective and visual means, with the same bit rate for the two criteria.

The repartition of the distortions generated in the reconstructed Lena images at 1.4 b/pixel (global Huffmann coding) in the two approaches is plotted in Fig. 9. We notice that the errors have different repartitions. We should notice especially that, with the MSE-criterion, errors have higher intensity (maximum error is 34 with MSE-criterion and 17 with L^∞-criterion). This means that visual distortions may appear earlier in the reconstructed image using the L^2-norm.

This phenomenon is also seen on the error images (obtained from the difference between the original image and the reconstructed ones at the same bit rate 1.4 b/pixel) where the errors are higher in L^2 approach, see Fig. 10. Note that the errors are concentrated on the edges [Fig. 10(a)].

This becomes very noticeable when the compression ratio increases, and the L^∞ coded images often look sharper than L^2 coded images. This must, however, be confirmed by intensive simulations.

We although should notice that in the MSE context, we minimize the distortion for a given entropy bit rate computed on the image, whereas in the L^∞ context we constrain the error to be less than a given threshold and minimize the bit rate given by (6). Particularly, the quantizers are optimized to give minimum entropy bit rate in the first case but only a theoretical bit rate in the second one. This simplified the optimization procedure in L^∞-criterion case and made it signal independent. But since visual quality comparison between the two criteria is required, we should minimize the same cost. Therefore, by means of some statistical assumptions on the coders' inputs, we plan to evaluate the entropy bit rate to be minimized. This procedure may improve the visual quality in



(a) (b)

Fig. 10. Error images between the original image (Lena, 512 × 512) and the reconstructed ones after a compression to 1.4 b/pixel (global Huffmann coding) with L^2 and L^∞ criteria. (a) Error image obtained in the L^2-norm context. (b) Error image obtained with L^∞-norm. The errors in (a) are higher and more concentrated on the edges than in (b).

L^∞ context and give a more realistic comparison between both approaches.

VIII. Conclusion

In this paper, we derived a technique of high-quality image coding based on a confidence interval criterion. The proposed approach has been explained in the context of predictive as well as filterbank coding. It has a very large flexibility.

- It allows efficient *lossless* coding to be performed by simple means (see Tables II and III). The DPCM is then more efficient than FBC in this context of lossless compression (see Fig. 8).
- It is based on a *local* control of the distortions, which allows one to go progressively from lossless compression to a lossy one, since the reconstruction error is controlled sample-by-sample so as to achieve a given quality of the reconstructed image. As shown in Fig. 8, the use of filterbanks is recommended if more than 5% of original image pixel values exceed the given threshold.
- The use of an L^∞-norm is compatible with lossy plus lossless coding. It happens that the most efficient scheme we could find makes use of a lossy FBC optimized with an L^∞-norm, plus a lossless coding of the residual error.
- Whatever the application, any signal is measured within a given accuracy, hence defining a *confidence interval*. Our method is able to maintain the reconstruction errors within this confidence interval, thus resulting in practical lossless coding in this case.
- We hope that since we can control the reconstruction error pixel by pixel, we may envision in the future to vary this error inside an image so that spatial masking properties can be used. This would already be useful is some medical applications where regions of interest can be defined, in which no loss is allowed, while some distortion can be allowed elsewhere. This can be done in our scheme simply by varying the quantization step values.
- Finally, it has been outlined that L^∞ coding could also be useful in the context of lossy coding (more likely in a high-quality context) due to the better coding of edges in images.

References

[1] M. Vetterli, "A theory of multirate filter banks," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-35, pp. 356–372, Mar. 1987.

[2] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing,* vol. 1, pp. 205–220, Apr. 1992.

[3] M. Rabbani and P. W. Jones, *Digital Image Compression Techniques*. Bellingham, WA: SPIE, 1991.

[4] 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. 1996.

[5] A. Habibi, "Comparison of nth-order DPCM encoder with linear transformations and block quantization techniques," *IEEE Trans. Commun.,* vol. COMM-19, pp. 948–956, Dec. 1971.

[6] A. B. Sripad and D. L. Snyder, "A necessary and sufficient condition for quantization errors to be uniform and white," *IEEE Trans. Acoust., Speech, Signal Processing,* vol. ASSP-25, pp. 442–448, Oct. 1977.

[7] M. Antonini, "Transformée en ondelettes et compression numérique des images," Ph.D. dissertation, Nice Univ., Nice, France, 1991.

[8] M. Loéve, in *Probability Theory*. Princeton, NJ: Princeton Univ. Press, 1960, ch. 6.

[9] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: Wiley, 1967, vol. I, ch. 7.

[10] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 1988, ch. 10.

[11] D. A. Wismer and R. Chattergy, *Introduction to Nonlinear Optimization*. Amsterdam, The Netherlands: North Holland, 1978.

[12] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math,* vol. XLI, pp. 909–294, 1988.

[13] A. K. Jain, "Image data compression: A review," *Proc. IEEE,* vol. 69, pp. 349–389, Mar. 1981.



Lamia Karray (S'95) was born in Sfax, Tunisia, in 1967. She graduated as an engineer from the École Nationale Supérieure de l'Electronique et de ses Applications, France, in 1992. She received the Diplôme d'Études Approfondies and the Ph.D. degree in automatics and signal processing from the University of Paris XI, France, in 1992 and 1995, respectively. From 1992 to 1995, she completed her Ph.D. and worked on high-quality image coding in the Centre National d'Études des Télécommunications (CNET) in Paris.

Since 1995, she has been with the Department of Research in Speech Communication,CNET, Lannion, France, where she is currently working on robust speech recognition.



Pierre Duhamel (M'87–SM'87) was born in France in 1953. He received the Ingénieur degree in electrical engineering from the National Institute for Applied Sciences, Rennes, France, in 1975, and the Dr. Ing. and Doctorat es Sciences degrees, both from Orsay University, France, in 1978 and 1986, respectively.

From 1975 to 1980, he was with Thomson-CSF, Paris, France, where his research activities were in circuit theory and signal processing, including digital filtering and analog fault diagnosis. From 1980 to 1993, he was with the National Research Center in Telecommunications, Issy-Les-Moulineaux, France, where his research interests were in the design of recursive CCD filters, fast Fourier transform and convolution algorithms, with application to adaptive filtering, spectral analysis, and wavelet transforms. He is now with the École Nationale Supérieure des Télécommunications, Paris, France, where his research activities include fast transform algorithms, image coding and compression, channel equalization, and array processing.

Dr. Duhamel served as an Associate Editor for IEEE TRANSACTIONS ON SIGNAL PROCESSING from 1990 to 1992, and was Associate Editor for IEEE SIGNAL PROCESSING LETTERS. He is chairman of the DSP Committee of the IEEE Signal Processing Society.



Olivier Rioul (M'94) was born in Strasbourg, France, in 1964. He received the Dipl. Ing. degree from the École Polytechnique, Palaiseau, France, in 1987, and the Dipl. Ing. Télécommunications and Doctorat d'Ingénieur des Télécommunications (Ph.D.) degrees from the École Nationale Supérieure des Télécommunications, Paris, in 1989 and 1993, respectively.

In 1988, he worked for AT&T Bell Laboratories, Murray Hill, NJ, on acoustics and speech representation using neural networks. From 1989 to 1994, he was with the Centre National d'Études des Télécommunications (CNET-France Télécom), Issy-Les-Moulineaux, France, where he worked on image compression, wavelets, multirate signal processing, computational complexity, and signal analysis. In 1994, he joined the École Nationale Supérieure des Télécommunications, where he is currently Assistant Professor of Electrical Engineering and Digital Communications. His research interests include joint source-channel coding, error control codes, information theory, multicarrier modulation, wavelets, and multirate signal processing.