专 题 汇 编

ICASSP' 图象编码

9

北京邮电学院图书馆

12770

ADAPTIVE DCT IMAGE CODING BASED ON A THREE-COMPONENT IMAGE MODEL[†]

Xiaonong Ran and Nariman Farvardin
Electrical Engineering Department
and Systems Research Center
University of Maryland
College Park, Maryland 20742

ABSTRACT

In this paper, we describe a three-component image model developed based on psychovisual studies of the human visual system. This image model consists of the primary component which contains the strong edge information of the image, the smooth component which represents the background slow-intensity variations and the texture component which contains the textures. Using this image model, we develop an adaptive DCT image coder in order to achieve high subjective performance at low bit rates (≤ 0.5 bits/pixel). The simulation results show that this adaptive DCT coder performs better than JPEG both objectively (in PSNR) and subjectively especially at low bit rates.

I. INTRODUCTION

Adaptive Discrete Cosine Transform (DCT) coding is a well established technique for image data compression applications [1] - [3]. In this scheme, the image is segmented into blocks of equal size and each block is operated upon by the 2D DCT. The transform blocks are then classified into several classes; each block is subsequently encoded with an encoder matched to its class. While these DCT coders offer good perceptual performance at bit rates about 1.0 bit/pixel and above, they usually produce specific types of visible distortion, e.g., blockiness, blurred edges, etc., at lower bit rates.

We observe that the fidelity criterion used in most of the existing DCT-based image coding systems is the Mean-Squared-Error (MSE), which owes its popularity to its mathematical tractability and to the fact that generally small values of MSE correspond to perceptually high quality reconstructed images. This fact is important because the human eye is usually the final judge of the quality of the reconstructed image. There is evidence, however, that the human eye is not an MSE detector [4], and that among all image coders of the same subjective performance, those based on the MSE do not necessarily have the lowest bit rate. These observations and the limitations experienced by the MSE-based image coding systems suggest that using a subjective-based fidelity criterion other than MSE may yield image coding systems of superior subjective performance at lower bit rates.

To this end, based on some observations of psychovisual aspects of the human perception [8], [9], we have introduced a three-component image model (3CIM) consisting of (1) the *primary*, (2) the *smooth* and (3) the *texture components*; the primary component contains the strong edge information, the smooth component provides the background slow-intensity variations and the texture component simply contains the textures of the image. These three components have different importance to the human perception [5], and thus should be treated accordingly in an image coding system.

In this paper, in addition to describing the 3CIM, we report an adaptive DCT coder based on the 3CIM, in which the primary component is encoded using the chain code representation [7] and the smooth and texture components are encoded by an entropy-coded adaptive DCT image coder. In this scheme, the blocks are classified according to the ac energies of the corresponding smooth and texture component blocks. The bits are allocated efficiently using the steepest descent from zero (SDFZ) algorithm [6]. The DCT coefficients are quantized with uniform-threshold quantizers (UTQs) and encoded using Huffman codes (HCs). The simulation results show that this system has an encouraging subjective performance especially at lower bit rates.

This paper is organized as follows. The 3CIM is described in Section II, followed by descriptions of the coding techniques for the three components in Section III and simulation results in Section IV. Section V contains a summary and conclusions.

II. A THREE-COMPONENT IMAGE MODEL

Based on the work in [8] [9], a study of psychovisual aspects of the human perception is conducted. This study indicates that the process of formation of the human visual perception can be modeled by the mechanism of appropriately formulated energy minimization problems [5]. For example, for an image consisting of broad black and white figures $X^o \equiv \{x^o_{i,j}\}$, $i, j = 0, \ldots, M-1$, the visually perceived image, denoted as $X^p \equiv \{x^p_{i,j}\}$, is modeled as the solution of the following problem,

$$\min_{\{x_{i,j}\}} \sum_{i=0}^{M-2} \sum_{j=0}^{M-2} [(x_{i,j} - x_{i,j+1})^2 + (x_{i,j} - x_{i+1,j})^2], \tag{1}$$

subject to $x_{i,j} = x^o_{i,j}$ for $(i, j) \in B$, where B contains the pixels at the locations of edges [5]. Intuitively, this model, named as a *minimum information principle* (MIP), suggests that the human visual system (HVS) estimates the brightness of an image only from the brightness variations in the image. Other situations can be modeled as similar minimization problems and interprated intuitively as follows. In the area close to a "strong edge" of an image, the HVS detects less textures as compared to areas without strong edges; for the area of smooth brightness changes, the HVS tends to detect less brightness changes.

A. Stressed Image

Motivated by the above psychovisual study of the HVS, we attempt to extract the strong edge information - the most important information for the formation of the human perception - from the image, in order to give it a special treatment in the image coding system to be designed. To do so,

[†] This work was supported in part by National Science Foundation grants NSFD MIP-86-57311 and NSFD CDR-85-00108.

0-7803-0532-9/92 $3.00 © 1992 IEEE

we introduce in the following, the concept of a *stressed image* associated with the original image.

Let the original image be denoted by $X = \{x_{i,j}\}$, $i, j = 0, \ldots, M-1$, where $x_{i,j}$ is the intensity value of the pixel (i, j), and M is the size of the image; similarly, the corresponding stressed image is denoted by $X^s = \{x^s_{i,j}\}$. The stressed image X^s has the following properties: (i) at the strong edges of X, X^s closely approximates X, i.e., the squared-errors, $(x_{i,j} - x^s_{i,j})^2$, are small at these locations; (ii) in other areas such as smooth and texture areas, X^s is smooth and its squared-errors from X are only loosely constrained. In other words, X^s contains only the strong edge information and the smooth intensity variations of X. The smoothness at pixel (i, j) is measured by the so-called *pixel row-curvature energy* $C^r_{i,j}$ and *pixel column-curvature energy* $C^c_{i,j}$ defined as follows, $C^r_{i,j} = (x_{i,j-1} - 2x_{i,j} + x_{i,j+1})^2$, $C^c_{j,i} = (x_{j-1,i} - 2x_{j,i} + x_{j+1,i})^2$, for $i = 0, \ldots, M-1$ and $j = 1, \ldots, M-2$, and $C^r_{i,j} = 0$, $C^c_{j,i} = 0$, otherwise. In order to generate X^s, we consider a linear combination of the squared-errors between X and X^s and the curvature energies, $C^r_{i,j}$ and $C^c_{i,j}$, at pixel (i, j),

$$E_{i,j}(X, X^s, \Lambda_{i,j}) \equiv \lambda^1_{i,j}(x_{i,j} - x^s_{i,j})^2 + \lambda^2_{i,j} C^r_{i,j} + \lambda^3_{i,j} C^c_{i,j}, \quad (2)$$

where the parameters in $\Lambda_{i,j} \equiv (\lambda^1_{i,j}, \lambda^2_{i,j}, \lambda^3_{i,j})$ are three non-negative real numbers. Then, X^s can be defined as the solution of the following minimization problem with a proper parameter set $\Lambda \equiv \{\Lambda_{i,j}, i, j = 0, 1, \ldots, M-1\}$,

$$\min_{\{x^0_{i,j}\}} \sum_{i=0}^{M-1} \sum_{j=0}^{M-1} E_{i,j}(X, X^0, \Lambda_{i,j}), \quad (3)$$

where $X^0 \equiv \{(i, j, x^0_{i,j})\}$[1]. The relation of X and X^s governed by (3) can be modeled as a 2D low-pass filter with input X and output X^s; the cutoff frequencies at (i, j) of the filter are controlled by $\lambda^1_{i,j}$, $\lambda^2_{i,j}$ and $\lambda^3_{i,j}$. Larger values of $\lambda^1_{i,j}$ give higher cutoff frequencies in both directions, while larger values of $\lambda^2_{i,j}$ and $\lambda^3_{i,j}$ lead to lower cutoff frequencies in the row-direction and column-direction, respectively [5]. Therefore, a proper parameter set Λ should consists of small values of $\lambda^2_{i,j}/\lambda^1_{i,j}$ and $\lambda^3_{i,j}/\lambda^1_{i,j}$ at strong edges and large values at other locations. Since the locations of strong edges are not known a priori, the stressed image X^s is generated iteratively. Starting with a uniform parameter set, the problem is solved to get X^1. Then the parameter set is updated by changing $\lambda^2_{i,j}/\lambda^1_{i,j}$ and $\lambda^3_{i,j}/\lambda^1_{i,j}$ inversely proportional to the curvature energies $C^r_{i,j}$ and $C^c_{i,j}$, respectively, of X^1. The above procedure is repeated until the relative variation of the objective function in (3) for two consecutive iterations is less than a given threshold (see [5] for details).

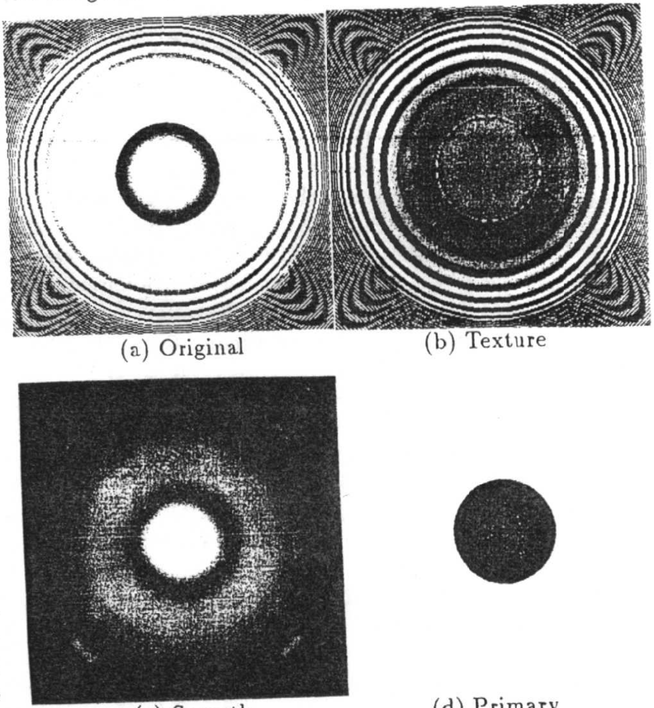
B. The Three Components

The strong edge information of X can be easily determined by identifying pixels of large curvature energies in X^s. These pixels characterize the brim contours of strong edges [5]. A contour is defined as a sequence of triples: $\overline{b} \equiv \{(i^k, j^k, x^s_{i^k,j^k})\}$, such that $|i^{k-1} - i^k| \leq 1$, $|j^{k-1} - j^k| \leq 1$, and $(i^{k-1}, j^{k-1}) \neq (i^k, j^k)$, for $1 \leq k < m$, and $\sigma_{\overline{b}} \equiv \max_{1 \leq k < m} |x^s_{i^k,j^k} - \overline{x}| \leq T_c$, where $T_c \geq 0$, $\overline{x}$ is the average intensity on $\overline{b}$, m is called the *length* of the contour, and $\sigma_{\overline{b}}$ is referred to as the *maximum-variation* of the intensities. This definition of a contour is similar to the previous ones [7] except for the additional condition on the intensities. The contours are generated using a local search algorithm as described in [5].

The primary component of X, denoted by $P \equiv \{p_{i,j}\}$, is generated from the extracted strong edge information by solving a minimization problem similar to (1). Since X^s contains the strong edge information and the smooth intensity variations of X, the difference image $X \ominus X^s \equiv \{(x_{i,j} - x^s_{i,j})\}$ consists of the textures of X. We define $T = \{t_{i,j}\} = X \ominus X^s$ and $S = \{s_{i,j}\} = X^s \ominus P$ corresponding to the texture and the smooth components, respectively. Therefore, we have

$$X \equiv T \oplus S \oplus P \equiv \{(t_{i,j} + s_{i,j} + p_{i,j})\}. \quad (4)$$

The above decomposition is referred to as the 3-component image model (3CIM). An example of the decomposition of a test image into its three components is provided in Fig. 1.


(a) Original (b) Texture

(c) Smooth (d) Primary

Fig. 1: The original image and its 3 components.

III. CODING OF THE THREE-COMPONENTS

The primary component is coded by encoding the information in the strong edge brims. The geometrical information of the brim contours are coded using *N-ring chain codes* [7], [10]. The outputs of the N-ring chain coder are then encoded using an HC. The intensities of the contour are represented by the average $\overline{x}$ (see Section II.B), which is then quantized with a UTQ and coded using an HC. From the received information, the receiver can obtain a replica of the the primary component $\hat{P}$, based on the MIP.

[1] This problem is named as an *Energy Minimization Model* (EMM) problem due to its analogy to a mechanical system [5].

The smooth and texture components are encoded with an entropy-coded adaptive DCT coder. More precisely, T and S are blocked into $(L \times L)$ blocks $\mathbf{t}_{m,n}$ and $\mathbf{s}_{m,n}$, respectively. The 2D DCT coefficients of $\mathbf{t}_{m,n}$ and $\mathbf{s}_{m,n}$ are denoted by $\bar{\mathbf{t}}_{m,n}$ and $\bar{\mathbf{s}}_{m,n}$, respectively. Similar to [1], the blocks are classified into four classes, except that the classification is made in two stages. In the first stage, the blocks are classified into one of two classes by comparing the ac energies of $\{\bar{\mathbf{s}}_{m,n}\}$ against a threshold T_1; this threshold is chosen such that the resulting two classes contain the same number of blocks. In the second stage, each resulting class of the first stage is further divided into two classes by comparing the ac energies of $\{\bar{\mathbf{t}}_{m,n}\}$ against another threshold. The two thresholds T_2^1 and T_2^2 in the second stage (one for each class of the first stage) are also chosen such that the resulting classes have the same number of blocks in them. With this two-stage classification, the frequency distribution of the ac energies is utilized since S (T) contains only low (high) frequency energy.

We denote the variance of the (i,j)th DCT coefficient of $T \oplus S$ in the kth class by $\sigma_k^2(u,v)$, where $k = 0,\ldots,3$, and $u, v = 0,\ldots,L-1$, and assume that the $(0,0)$th, $(0,1)$th and $(1,0)$th coefficients have Gaussian distributions (GD), while all other coefficients have Laplacian distributions (LD). This assumption is based on a study of the DCT coefficients' distributions. The DCT coefficients are quantized with UTCs whose outputs are then encoded using appropriately designed HCs. The two UTC-HC coders for the GD and LD used here are identical to those used in [11]. The variance-normalized MSEs associated with the UTC-HC coder operating at r bits/sample are denoted by $d_G(r)$ and $d_L(r)$ for the GD and LD, respectively. The overall MSE is

$$D \equiv \sigma^2(0,0)d_G(r_{0,0}) + \frac{1}{4}\sum_{k=0}^{3}\{\sigma_k^2(0,1)d_G(r_{0,1}^k) + \\ \sigma_k^2(1,0)d_G(r_{1,0}^k) + \sum_{(u,v)\neq(0,0),(0,1),(1,0)} \sigma_k^2(u,v)d_L(r_{u,v}^k)\}, \quad (5)$$

where $\sigma^2(0,0)$ and $r_{0,0}$ are the variance and the coding rate, respectively, for the dc coefficient which is encoded in the same manner regardless of which class it belongs to, and $r_{u,v}^k$ is the coding rate for the (u,v)th coefficient in the kth class. The overall bit rate is given by

$$R \equiv r_o + r_c + \frac{1}{L^2}\{r_{0,0} + \frac{1}{4}\sum_{k=0}^{3}\sum_{u,v\neq(0,0)} r_{u,v}^k,\}\ \text{bits/pixel}, \quad (6)$$

where r_o is the bit rate for coding the overhead information, namely, the class information (2 bits/block), the mean and variance of the dc coefficients (64 bits for each frame), the bitmaps[2] and the normalization factor c (32 bits for each frame) defined below and r_c is the bit rate for coding of the primary component. The bit rates $r_{0,0}$ and $r_{u,v}^k$'s are determined efficiently with the SDFZ algorithm [6].

To reconstruct the DCT coefficients at the receiver side, $\sigma_k^2(u,v)$'s ($(u,v) \neq (0,0)$) are needed. Notice that if the quantization errors of DCT coefficients are known, $\sigma_k^2(u,v)$'s can be computed easily from the rates. To investigate the quantization errors of the DCT coefficients, we use the Shannon lower bound (SLB) to approximate $d_G(r)$ and $d_L(r)$ [12]

$$d_G(r) = exp(-2r), \qquad d_L(r) = exp(-2r)/\pi. \qquad (7)$$

Substituting (7) into (5) and using the Lagrange multiplier, one can solve this constrained minimization problem (with the equality constraint (6)) and find that the optimal quantization errors for different coefficients are the same. Therefore, in the ideal case, only one number is required for the receiver to calculate the variances $\sigma_k^2(u,v)$'s ($(u,v) \neq (0,0)$) from the bitmaps; this number is called the normalization factor c and equals the quantization error of any DCT coefficient (except the $(0,0)$th). In the system actually implemented, the average of the quantization errors is used as the value of c.

IV. SIMULATION RESULTS

The test image, referred to as LENA, is a 512 × 512 monochrome image shown in Fig. 3 (a). The primary component is encoded using a bit rate equal to 0.04 bits/pixel; the HCs for the output of the 2-ring chain code and the UTC-HC coder of contour intensity values are designed based on the statistics of the contours of eight images excluding LENA. The block size is 16 × 16 for the simulations of the DCT schemes developed here.

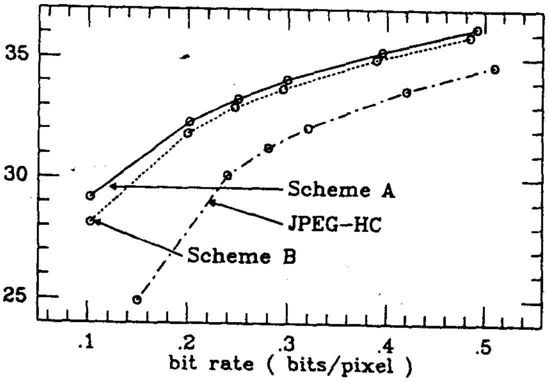


Fig. 2: PSNR performance (in dB).

The PSNR performance of the above mentioned adaptive DCT coder based on 3CIM (Scheme A) is depicted in Fig. 2 for bit rates between 0.1 and 0.5 bits/pixel. Also included in Fig. 2 are the PSNR performance of JPEG with HC (JPEG-HC) as well as that of an entropy-coded adaptive DCT coder (Scheme B) similar to Scheme A but with the exception that the 3CIM is not used. In Scheme B, the classification algorithm is identical to that of [1]. Scheme B can be thought of as an entropy-coded modification of the scheme in [1]. It can be seen from these PSNR results that Scheme A offers noticeable improvements over JPEG-HC especially at lower bit rates and also performs better than Scheme B. A few reconstructed images are shown in Fig. 3. In comparing the results

[2] The bitmap $\{b_{u,v}^k\}$ $(k = 0,\ldots,3)$ is encoded using two integers $(2 \times \log_2 L$ bits) to specify the largest indices u_k and v_k such that $b_{u_k,v}^k \neq 0$ and $b_{u,v_k}^k \neq 0$ for some v and u and using 6 bits to represent every $b_{u,v}^k$, for $u \leq u_k$ and $v \leq v_k$. Thus the overall bit rate for the bitmaps will be $4(2\log_2 L + 6u_k v_k)$ bits for each frame.

of Scheme A and Scheme B, we note that the reconstructed images obtained from Scheme A have better perceptual quality especially at the locations of strong edges. Our experimental studies indicate that while the coding of the primary component in Scheme A is responsible for the better subjective quality of Scheme A (as compared to Scheme B), the PSNR improvements are due to the two stage classification algorithm in Scheme A.

V. SUMMARY AND CONCLUSIONS

In this paper, an adaptive DCT image coder with a subjective-based fidelity criterion is introduced in order to achieve good subjective performance at bit rates equal to and below 0.5 bits/pixel. A three-component image model, which is developed based on a psychovisual study of the HVS, is utilized for the special treatment of the primary component and for the block classification to capture the frequency distribution of ac energies. Close to optimal UTC-HC coders are used to quantize the DCT coefficients. The variances of the DCT coefficients are estimated from the bitmaps based on an argument using the SLB. The simulation results show that this new DCT coder performs better than JPEG-HC in PSNR, and offers high subjective quality at low bit rates.

REFERENCES

[1] W. H. Chen and C. H. Smith, "Adaptive coding of monochrome and color images," *IEEE* COM-25, pp. 1285-1292, Nov. 1977.

[2] N. S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall, Englewood Cliffs, NJ. 1984.

[3] W. A. Pearlman, "Adaptive Cosine Transform Image Coding with Constant Block Distortion," *IEEE* COM-38, pp. 698-703, May 1990.

[4] A. N. Netravali and J. O. Limb, "Picture coding: a review," *Proc. of IEEE*, vol. 68, No. 3, pp. 366-406, Mar. 1980.

[5] X. Ran, "A Three-Component Image Model for Human Visual Perception and Its Applications in Image Coding and Processing," Ph.D. dissertation, Univ. of Maryland, in preparation.

[6] X. Ran and N. Farvardin, "Combined VQ-DCT coding of images using interblock noiseless coding," in *Proc. IEEE ICASSP*, Apr. 1990, pp. 2281-2284.

[7] H. Freeman, "On the Encoding of Arbitrary Geometric Configuration," *IRE Trans. Elec. Com.*, EC-10, pp. 260-268, Jun. 1961.

[8] H. Helmholtz, *Treatise on Physiological Optics*, edited by J. Southall, vol. III, *The Perceptions of Vision*, O. S. A., George Bonta Pub. Co., Menasha, WI, 1925.

[9] T. Cornsweet, *Visual Perception*, Academic Press, 1970.

[10] D. L. Neuhoff and K. G. Castor, "A Rate and Distortion Analysis of Chain Codes for Line Drawings," *IEEE* IT-31, pp. 53-67, Jan. 1985.

[11] N. Tanabe and N. Farvardin, "Subband Image Coding Using Entropy-Coded Quantization Over Noisy Channels," in *Proc. IEEE ICASSP*, Apr. 1990, pp. 2105-2108.

[12] T. Berger, *Rate Distortion Theory*, Prentice-Hall, 1971.

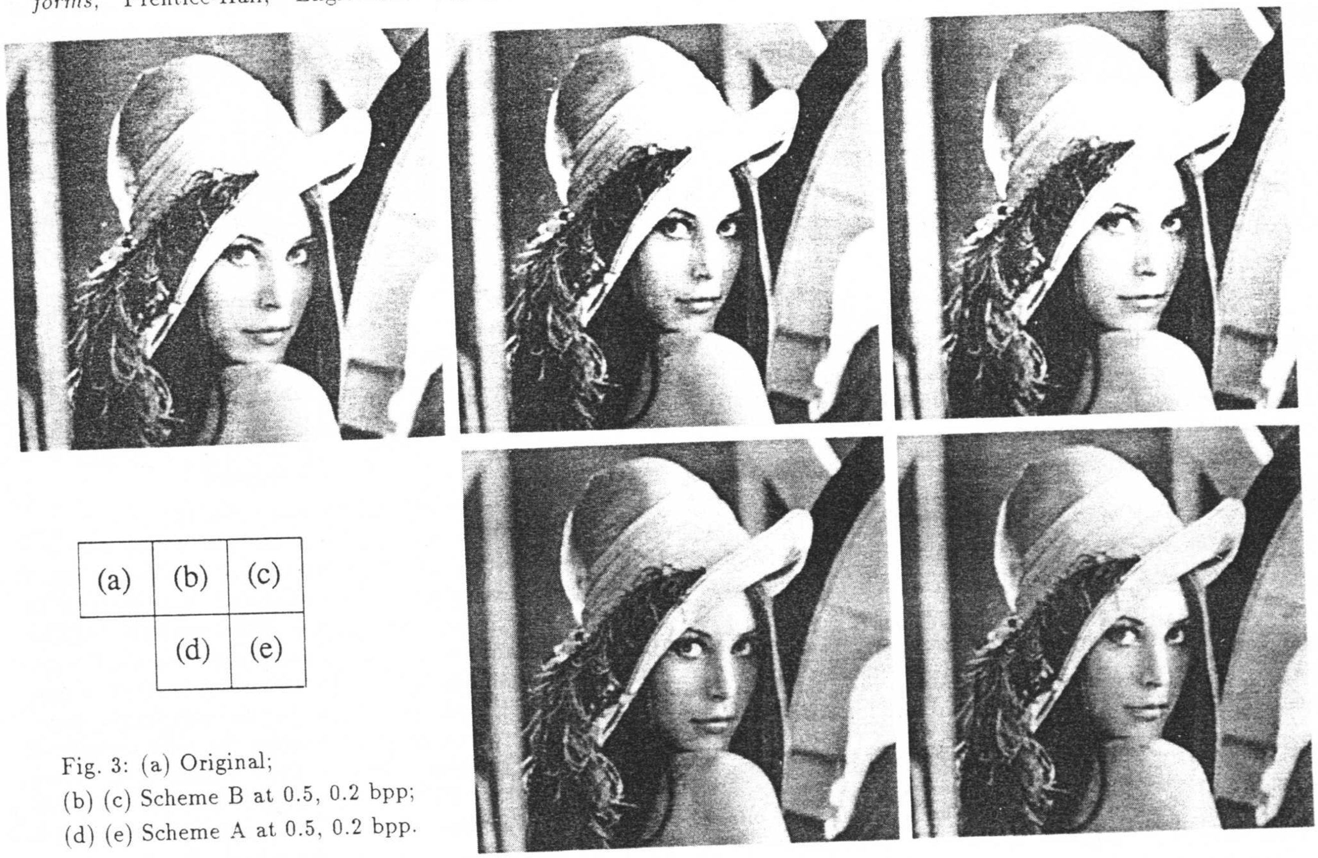


Fig. 3: (a) Original;
(b) (c) Scheme B at 0.5, 0.2 bpp;
(d) (e) Scheme A at 0.5, 0.2 bpp.

REAL-TIME RECURSIVE TWO-DIMENSIONAL DCT FOR HDTV SYSTEMS [†]

C.T. Chiu and K.J. Ray Liu
Electrical Engineering Department and Systems Research Center,
University of Maryland, College Park, MD 20742 USA

ABSTRACT

The two-dimensional discrete cosine transform (2-D DCT) has been widely recognized as the most effective technique in image data compression. In this paper, we propose a new algorithm to compute the 2-D DCT from a frame-recursive point of view. Based on this approach, a real-time parallel lattice structure for the 2-D DCT is developed. The system is fully-pipelined with throughput rate N clock cycles for an $N \times N$ successive input data frame. This is the fastest pipelined structure known so far. Moreover, the 2-D DCT architecture is modular, regular, and requires only two 1-D DCT blocks which can be extended directly from the 1-D DCT array. We also propose a parallel 2-D DCT architecture and a new scanning pattern for the HDTV system to achieve higher performance.

1. INTRODUCTION

In recent years, much research has been focus on image data compression, especially for the application of the next generation TV, "HDTV". To make HDTV systems practical, bit rate reduction and data compression are indispensable [5]. The DCT coding approach has obtained most attention due to its superior energy compaction property and much simpler computations than the optimal Karhunen-Loeve transform (KLT). To satisfy the high speed video transmission system, fast and efficient algorithm to implement the 2-D DCT with simple hardware is strongly demanded [1, 2, 4]. The irregularity, global communication, and transposition delay of the existing 2-D DCT architectures have severe impacts on high speed video signal processing systems. Here we propose a new real-time recursive 2-D DCT architecture which requires only two 1-D DCT arrays and no transposition is required.

2. FRAME RECURSIVE 2-D DCT ARCHITECTURE

A new algorithm for the 2-D DCT by employing the frame-recursive concept [3] on successive input frames is presented. We adopt the frame-recursive approach since in digital signal transmission data arrive seriesly. Such approach can obtain the 2-D DCT in real-time recursively. Based on this method, a parallel and fully-pipelined 2-D DCT lattice structure which can dually generate the 2-D DCT and discrete sine-cosine transform (DSCT) is developed. The 2-D DCT $\{X_c(k,l,t) : k,l = 0,1,...,N-1.\}$ and 2-D discrete sine-cosine transform (DSCT) $\{X_{sc}(k,l,t) : k = 1,2,...,N; l = 0,1,...,N-1.\}$ of an $N \times N$ 2-D data sequence $\{x(m,n) : m = 0,1,2,...; n = 0,1,...,N-1.\}$ is defined as

$$X_c(k,l,t) = \frac{4}{N^2}C(k)C(l)\sum_{m=t}^{t+N-1}\sum_{n=0}^{N-1}x(m,n) \cdot \cos\left[\frac{\pi[2(m-t)+1]k}{2N}\right]\cos\left[\frac{\pi(2n+1)l}{2N}\right] \quad (1)$$

and

$$X_{sc}(k,l,t) = \frac{4}{N^2}C(k)C(l)\sum_{m=t}^{N+t-1}\sum_{n=0}^{N-1}x(m,n) \cdot \sin\left[\frac{\pi[2(m-t)+1]k}{2N}\right]\cos\left[\frac{\pi(2n+1)l}{2N}\right] \quad (2)$$

where

$$C(k) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } k = 0 \text{ and } k = N, \\ 1 & \text{otherwise.} \end{cases}$$

In the following, we call $X_c(k,l,t)$ and $X_{sc}(k,l,t)$ the the t'th frame 2-D DCT and 2-D DSCT of an $N \times N$ 2-D data frame $x(m,n)$. The recursive relations for the $(t+1)$'th frame transformed data

[†] The work is partially supported by ECD-8803012.

$X_c(k,l,t+1)$ and $X_{sc}(k,l,t+1)$ as well as the t'th frame transformed data $X_c(k,l,t)$ and $X_{sc}(k,l,t)$ are given by

$$X_c(k,l,t+1) = \overline{X}_c \cos\left(\frac{\pi k}{N}\right) + \overline{X}_{sc} \sin\left(\frac{\pi k}{N}\right), \quad (3)$$

and

$$X_{sc}(k,l,t+1) = \overline{X}_{sc} \cos\left(\frac{\pi k}{N}\right) - \overline{X}_c \sin\left(\frac{\pi k}{N}\right), \quad (4)$$

where

$$\overline{X}_c = \frac{4}{N^2} C(k) C(l) \sum_{m=t+1}^{t+N} \sum_{n=0}^{N-1} x(m,n) \cdot \cos\left[\frac{\pi[2(m-t)+1]k}{2N}\right] \cos\left[\frac{\pi(2n+1)l}{2N}\right] \quad (5)$$

and

$$\overline{X}_{sc} = \frac{4}{N^2} C(k) C(l) \sum_{m=t+1}^{t+N} \sum_{n=0}^{N-1} x(m,n) \cdot \sin\left[\frac{\pi[2(m-t)+1]k}{2N}\right] \cos\left[\frac{\pi(2n+1)l}{2N}\right] \quad (6)$$

The relations between $\overline{X}_c$ and $\overline{X}_{sc}$ and the previous transformed data $X_c(k,l,t)$ and $X_{sc}(k,l,t)$ are

$$\overline{X}_c = X_c(k,l,t) + \delta_c(k,l,t) \frac{2}{N} C(k) \cos\left(\frac{\pi k}{2N}\right), \quad (7)$$

and

$$\overline{X}_{sc} = X_{sc}(k,l,t) + \delta_c(k,l,t) \frac{2}{N} C(k) \sin\left(\frac{\pi k}{2N}\right). \quad (8)$$

And the intermediate values $\delta_c(k,l)$ are

$$\delta_c(k,l) = \frac{2}{N} C(l) \sum_{n=0}^{N-1} \left[(-1)^k x(N,n) - x(0,n)\right] \cdot \cos\left[\frac{\pi(2n+1)l}{2N}\right]. \quad (9)$$

The relation between $X_c(k,l,t+1)$ and $X_c(k,l,t)$ is realized by lattice array II with lattice module shown in Fig. 1. It is noted that $\delta_c(k,l)$ in (9) is the 1-D DCT of the data vector which is the difference between the parity of the $(t+N)$'th row and t'th row of the 2-D input sequence. It can be shown that $\delta_c(k,l)$ can be generated time recursively by the lattice array I whose lattice module is plotted in Fig. 2. The fully-pipelined lattice structure for the 2-D DCT and DSCT is shown in Fig. 3 which includes one LAI, one $LAII$, and two circular shift arrays and shift register arrays.

The circular shift array in the middle of the system is an $N \times 1$ shift register array. This special shift register array loads an $N \times 1$ data vector from the LAI every N clock cycles, then it shifts the data circularly and sends the data to the $LAII$ every clock cycle. There are three inputs in $LAII$, δ_c, $X_c(k,l,t)$ and $X_s(k,l,t)$, where the δ_c comes from the circular shift array, and $X_c(k,l,t)$ and $X_s(k,l,t)$ from the shift register arrays located behind the $LAII$. We divide the $LAII$ into two groups: the $LAII_{even}$ and $LAII_{odd}$. Each includes $N/2$ lattice modules as shown in Fig. 3. The $LAII_{even}$ contains only those lattice modules for even transformed components k, while $LAII_{odd}$ contains only the odd lattice modules. The shift register array contains $2N \times N$ registers which are used to delay data for N clock cycles.

We will show how to apply the frame-recursive concept to obtain the 2-D DCT. Our approach is to send the input sequence $x(m,n)$ row by row directly into the LAI. It takes N clock cycles for the LAI to complete the 1-D DCT of one row input vector, then the array sends the 1-D DCT data in parallel to the $CSMII$ as shown in Fig. 3. The circular shift matrix II ($CSMII$) is an $(N+1) \times N$ sequential shift register. At the output of the $CSMII$, the 1-D transformed data of the $(t+N)$'th row and t'th row are added together according to (9) depending on the sign of the k components (see Fig.3). Then the results are sent to CSAs. The upper CSA translates the intermediate value $\delta_c(k,l)$ to the lattice array II_{even}, as do the lower CSA except that the signs of the output of the CSA are changed before being sent to the lattice array II_{odd}. Since $LAII_{even}$ and $LAII_{odd}$ have only $N/2$ modules, every δ_c is floating for $N/2$ clock cycles. It is noted that a specific 2-D transform data $X_c(k,l,t+1)$ and $X_{sc}(k,l,t+1)$ are updated recursively every N clock cycles from $X_c(k,l,t)$ and $X_{sc}(k,l,t)$. Therefore the outputs of the $LAII$ are sent into the shift register array (SRA) where data are delayed by N clock cycles. Each SRA contains $N/2$ shift registers each with length N. The data in the rightest registers are sent back as the $X_c(k,l,t)$ and $X_{sc}(k,l,t)$ of $LAII$. At the N^2 clock cycle, the 2-D DCT and DSCT of the 0'th frame are available. After this, the 2-D transformed data of successive frames can be obtained every N clock cycles.

There are many interesting results in this structure. First, the lattice array can be viewed as a filter bank. It is because every lattice module itself is an independent digital filter with different frequency components $l = 0, 1, ..., N-1$. Moreover, all the lattice modules in this architecture have the same structure which is regular, modular, and without global communication. Second, the system requires only 2 1-D DCT arrays and is fully pipelined with throughput rate N clock cycle for frame-recursive approach. A comparison with existing algorithms is given in Table 1.

3. APPLICATION TO HDTV SYSTEMS

Most of the 2-D DCT implementations in HDTV systems are based on the row-column decompositions methods [5, 6]. Although fast algorithms exist for the 1-D DCT, the second 1-D DCT cannot start until all the first 1-D DCTs are completed. To speed up the operations, one method is to execute the first 1-D DCT in parallel. For the 8 × 8 case, there are 8 1-D DCT blocks to perform the first transform simultaneously. Assuming that each signal is 10-bit long, in order to to satisfy the precision, then the total number of bits required in the input is 640 bits, which is not practical in the circuit realizations. From this point of view, our serial input 2-D DCT system is more practical in hardware implementations. Moreover, if the speed of the circuit components, such as the ROM and adder, is high enough, our 2-D DCT system can be executed as fast as the sample clocking rate.

Although our 2-D DCT implementations are effective, transforming a video frame of 1080 x 1920 still requires intensive computations. Therefore, we designed a 2-D DCT architecture suitable for the HDTV system to achieve higher performance. The block diagram of the 2-D DCT encoder is shown in Fig. 4, where five 2-D DCT chips are included. Five chips were used because the ratio of pixel numbers per line for luminance signal Y and color difference signals U and V is 4:2:2. As the sampling frequency of HDTV is very high, the pixels of Y are divided into four groups, in order to carry out DCT in parallel. Additionally, the color difference signal Y and U are switched alternatively to another DCT coder. The scanning processor shown in Fig. 4 is used to divide the signal into four luminance components and one color difference component. The outputs of the 2-D DCT transformed data are sent to the entropy encoder in parallel or through multiplexers. Since the transform block size is 8 × 8, we divided the frame into 135 × 240 blocks and 240 channels as shown in Fig. 5. The 2-D DCT are executed on each channel whose scanning pattern is shown in Fig. 5. This scanning pattern reflects the fact that our system is based on row by row scanning order and is fully pipelined. Thus, such a scanning method would maximize the system throughput.

4. CONCLUSIONS

In this paper, we propose a new 2-D DCT algorithm based on a frame-recursive approach. The resulting 2-D DCT architecture can be obtained by using only two 1-D DCT arrays, at the same time, the transposition procedure is eliminated. It, therefore, does not have the drawback of the row-column decomposition method in which a transposition is needed between the first and the second 1-D DCT. The parallel 2-D DCT architecture and the scanning pattern proposed in Section 3 can process the video data in real time and eliminate the waiting time in the DCT codings so that the system performance can be maximized. Consequencely, our real-time parallel and fully-pipelined 2D-DCT structure is very attractive in high speed transmission systems.

References

[1] W. Ma, " 2-D DCT systolic array implementation," Electronics Letters, Vol. 27, No. 3, pp. 201-202, 31st Jan. 1991.

[2] P. Duhamel and C. Guillemot, "Polynomial Transform computation of the 2-D DCT," IEEE ICASSP Proc., pp. 1515-1518, March. 1990.

[3] K. J. R. Liu, and C. T. Chiu, "Unified Parallel Lattice Structures for Time-Recursive Discrete Cosine/Sine/Hartley Transforms," submitted to IEEE Trans. Acoust., Speech, Signal processing.

[4] B. Silkstrom, *et al.*, "A high speed 2-D Discrete Cosine-Transform," Integration, VLSI journal 5, pp. 159-163, 1987.

[5] K. Kinoshita, T. Nakahashi, and . Eto, "130 M bit/s (H4 rate) HDTV Codec based on the DCT algorithms," Electronics Letters, Vol. 26, No. 16, pp. 1245-1246, 2nd Aug. 1990.

[6] S. Cucchi, and F. Molo, "DCT-based Television Codec for DS3 digital Transmission," SMPTE Journal, pp. 640-646, Sep. 1989.

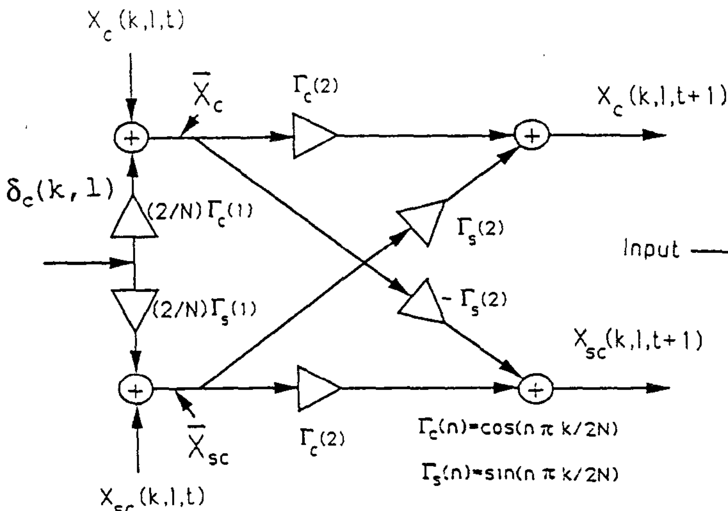


Figure 1: The lattice module of lattice array II.

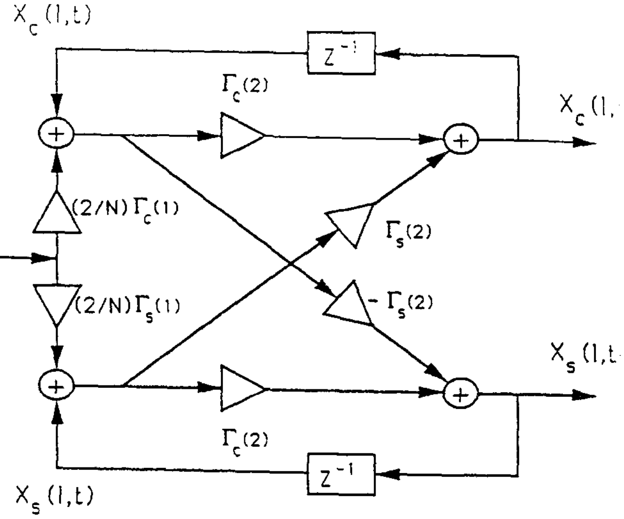


Figure 2: The lattice module of lattice array I.

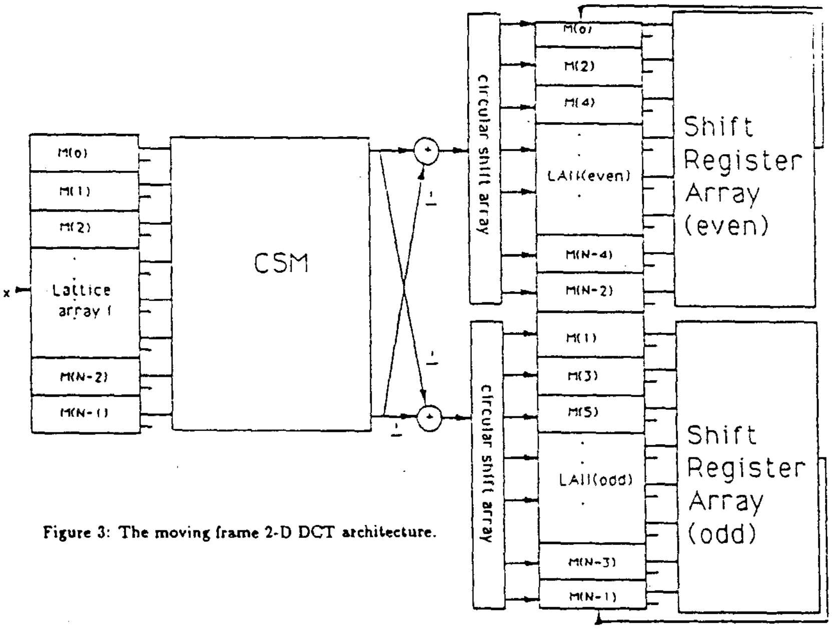


Figure 3: The moving frame 2-D DCT architecture.

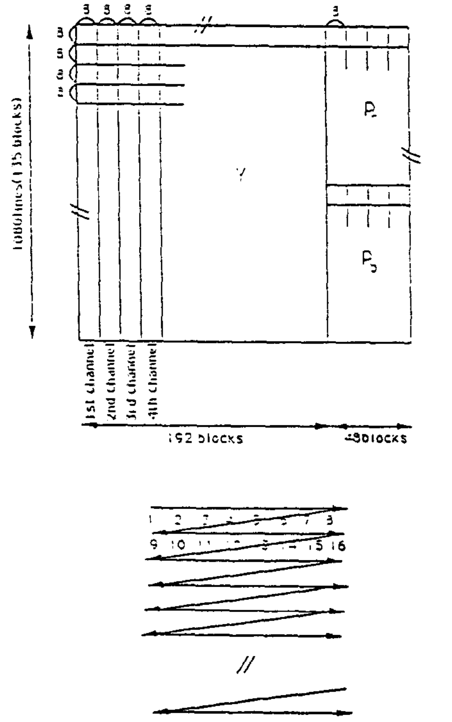


Figure 5: Block construction of a video frame and proposed scanning patter

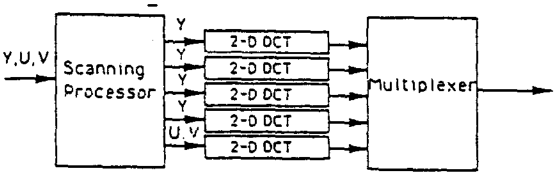


Figure 4: The block diagram of the DCT encoder.

	row-column method based on Chen in[2]	Duhamel[2] *et. al.*	Ma [1]	Liu-Chiu2D
No. of multipliers	$2N^2ln(N)$ $-6N^2/2+8N$	N^2 $+2N+2$	$4N(N+1)$	$8N$
Throughput	$N+$ transposition	$2N$	$2N+1$	N
Limitation on transform size N	power of 2	power of 2	no	no
Communication	global	global	local	local
I/O operation	*PIPO*	*PIPO*	*SIPO*	*SIPO*
Approach of of algorithm	indirect	direct	indirect	direct

Table 1: Comparisons of different 2-D DCT algorithms.

SELECTIVE DECOMPRESSION ON A HIERARCHICALLY CODED IMAGE

H. Torbey*, D. Freidlander*, J. Barda**

* Center for Telecommunications Research
and Department of Electrical Engineering
S.W. Mudd Bldg. Room 1220
Columbia University, NY, NY 10027

** AVELEM
"La BILLARDIERE", 36190 GARGILESSE - FRANCE

ABSTRACT

A novel application, selective decompression, is introduced. It enables to access a hierarchically coded image and retrieve that information necessary to decode a selected image portion excluding any other data. This techniques allows the handling of large images and provide a solution to the problems associated with them, namely: storage, excessive transmission delays, congestion of network resources, and mismatch between image and screen dimensions.

Technological advances have made available large communication bandwidth, large storage and powerful processing capacities. However these advances have also spurred the emergence of very demanding new services involving the transfer of such large data volumes that heavy loads are likely to be imposed on network and users resources alike [1]. The strains on network resources will translate into congestion of resources and transmission delays even if transmission occurs over broadbank network such as ATM based B-ISDN [2]. The strains on user resources will translate into a shortage of storage capacity. These new services we are alluding to center mainly around imaging applications. In the Visual Arts System for Archiving and Retrieval of Images (VASARI) project, for example [3], an image database accessible mostly to researchers is being created. At a planned scanning density of upt to 250 pixels/inch and with 7 color components per pixel (Red, Green, Blue, 2 infra-red frequencies, and 2 ultraviolet frequencies) and 8 bits per component, the system has the potential of generating extremely large image files. A two square meter painting for example, will represent more than 12 Gbits of data. Similarly, the "Bibliotheque de France" is an electronic archival and retrieval system involving images from the collections of major libraries and museums across France. The Request For Quotes [4] on the system specifies that the images, scanned in full color, will have sizes up to 6000 x 8000 pixels leading to a data volume of more than 1 Gbits when scanned in 24-bit color. In another example, large engineering companies and the nuclear industry [3], are setting-up image distribution systems to handle their maintenance applications. In a typical configuration, a central site will hold all the engineering drawings and documents, and will be accessed remotely on-demand. The drawings can be very large. A common size drawing of 32" by 24", if scanned at 200 pixels per inch and coded with 256 gray levels, will lead to a 6400 x 4800 pixels image having a data volume of more than 234 Mbits.

Note that in the applications just mentioned, a user will be faced with an important additonal problem, that of a mismatch between the image and screen dimensions for most of the "high resolution" displays readily available have dimensions of about

0-7803-0532-9/92 $3.00 © 1992 IEEE

1000 x 1000 pixels, a fraction of the dimensions of the large images likely to be handled.

A pragmatic solution to the above problems may reside in the concept of selective access : A user, faced with local display/ storage limitations, relatively limited bandwidth availability and wishing to display a portion of a high resolution image, is provided with the means to access exclusively the information relevant to decoding the image portion of interest. Selective access on an uncompressed image is trivial but does not constitute an attrative solution. Without compression, the source (and the destination) where the images are stored will be faced with storage problems and transmission of the selected data will be less efficient. Selective access on a compressed image, referred to as selective decompression, solve the above problems. It is however a complex, coder dependent algorithm. In this paper we introduce a system in which an image is coded and compressed using a multi-resolution hierarchical technique. The compressed image is transmitted, and at the destination, it is uncompressed and decoded up to a given intermediate level in the hierarchy. This intermediate level is selected such as to have a resolution suitable to the display limitation of the receiver. The user is given the means to select part of this intermediate level as an area to be decompressed and decoded further. Using a simple mathemaical relation, only the coded and compressed information relevant to the selected area is retrieved from the compressed data stream at the source and transmitted to the destination, which will combine it with the pixels of the selected area to achieve the requested expansion. This process can be repeated as needed until the resolution of the original image is reached.

Section 2 will formalize the selective decompression procedure while the equations for selecting the compressed values relevant to the area of interest to the user are detailed in section 3 as applicable to the JPEG proposed standard in its lossless hierarachical mode [5]. Section 4 will conclude by reviewing the advantages of the selective decompression procedure.

2. Selective Decompression Using a Hierarchical Image Coder

The selective decompression technique is particularly attractive in association withmulti-resolution hierarchical coding of images. This type of coding affords the creation of a good quality intermediate image at the decoder, with dimensions adaptable to those of the receiving display, while transmitting a relatively small fractionof the total image compressed data volume. Further transmission, decompression and decoding will be done on the basis of an area of interest defined by the user on the intermediate image.

The basic sequence of operation is as follows:
1- An image is scanned at an appropriately high resolution.
2- The image is coded using a multi-resolution hierrchical technique to allow multi-resolution progressive or hierarchical transmission/decompression/decoding and it is compressed to reduce its data volume.
3- Multi-resolution progressive transmission and decompression allow the display of the decoded image at a user-selectable lower resolution version of the original image. This lower resolution is in most cases dependent upon the relative dimensions of the display screen (or window) at the receiver and those of the original image.
4- A size and position adjustable window is scrolled acroos the lower resolution image under user control. Once these parameters are fixed to indicate user designation of an image portion, they define what is referred to as a "tile of interest" or a "tile" (i.e. - a user selected image portion).
5- The parameters of the tile, i.e. size of the designated image portion and its positioning relative to a reference point in the lower resolution image, are transferred back to the transmission source.
6-Using a mathematical relation, only the compressed data relevant to the decompression and decoding of the tile of interest is accessed at the source and then trans-

mitted. At the destination, the decompressed and decoded tile is displayed on the user's display up to a user-selectable resolution, or up to the resolution capability of the user's display.
7- Steps 4, 5, and 6 can be repeated if the previous step did not yield the final resolution for the tile of interest.

This process is illustrated in figure 1 where selective decompression is repeated twice before the final image has reached its final resolution.

3. Mathematical Relation for JPEG's Hierarchical Lossless Coder

In this section we present the mathematical relation enabling selective decompression on an implementation of the JPEG lossless multi-resolution hierarchical coder as described in annex J and annex K of the draft strandard [5]. Its operation can be summarized as follows :

"An image is encoded in a DPCM-coded frame followed by a sequence of differential frames which are two complements differences between source input components 4:1 downsampled, and the reference components 1:4 upsampled. The reference components are reconstruct components created by previous frames in the hierarchical process. for either the DPCM-coded frame or the differential frames, reconstructions fo the components are generated as a reference components for a subsequent frame in the hierarchical process." The DPCM-coded frame and the differentialframes are sent to the Huffman coder for compression. The compressed image is then constituted by the Huffman coded DPCM differential frames (C(i), i=1,..., F).

The mathematical relation in this case is straightforward. Assume the intermediate level obtained from the decoding of C(i) to have vertical dimensions Vi and horizontal dimensions Hi. The user will define a tile identified by the coordinates of its upper-left hand corner and lower right hand corner (Xt1, Yt1) and (Xt2, Yt2) respectively, given the following format (horizontal coordinate, vertical coordinate). The values relevant to the selective decompression process inlevel C(k), k>i are identified as follows :

```
let a = $\overline{(Vk/V)}$ = $\overline{(Hk/Hi)}$ and x lm ε C(k)

for (l=0; l<Vk; l++) {
        for (m=0; m<Hk; m++) {
                if(a.Yt1 <= l <= a.Yt2) &&
(a.Xt1 <=m <= a.Xt2) select $x_{lm}$ ε C(k);
                else reject $x_{lm}$ ε C(k)
}
```

An overbar refers to rounding up to the nearest integer value.

4. Conclusion

We have introduced in this chapter an algorithm allowing decompression of a user selectable portion of a compressed image. This technique was demonstrated in association with the lossless hierarchical coder described in chapter two. The user is first sent the necessary information to display a subsampled version of the image. The resolution of the subsampled image is generally determined by the resolution of the display. On the subsampled image, the user selects a portion of the image to be decompressed further. Given the coordinates of the selected portion relative to the upper left hand corner of the subsampled image, the algorithm is able to select from the compressed data of the whole image, only what is necessary to decode the selected portion.

Aside from the fact that the proposed selective access scheme provides a practical solution to the problem of the mismatch between original image resolution and screen resolution, it also allows for a substantial bandwidth and storage savings byavoiding the transmission of unwanted data. Consequently, decoding time and user resources have also been saved. The savings can be seen from the selective decompression example of an image, "church", 1855 x 2350 pixels 8bits/pixel. Should all the compressed image been transmitted, it would have represented a data volume of 12.39 Mbits, requiring more than 22 minutes of transmission over a 9600 bits/second modem. The receiver would also have had to solve the problem of creating storage space for the uncompressed image representing 33.26 Mbits of data. Using selective decompression, a user can first decompresse the image up to dimensions 232 x 294 pixels then zoom on a 499 x 400 pixels area at the original resolution in 2 steps as in figure 1. The total volume transmitted is 699 kbits in 1 minute and 15 seconds over a 9600 bits/second modem. All the images used can very well fit on a VGA screen.

[1] H. Torbey, "Lossless Multi-Resolution still Image Coding and Transmission", Ph. D. Dissertation, Columbia University 1992.

[2] L. Kleinrock, "ISDN-The path to broadland networks", proceedings of the IEEE, February 1991, pp. 112-117.

[3] L. Mac Donald, "Europe's group growing support for imaging in art", Advanced Imaging, September 1990.

[4] Bibliothèque de France, "cahier des charges pour les tests de transfert sur support électronique des images fixes', March 1991.

[5] JPEG Committe Draft, CD10918-1, "Digital compression and coding of continuous-tone still images", Part I : Requirements and guidlines, January 1991.

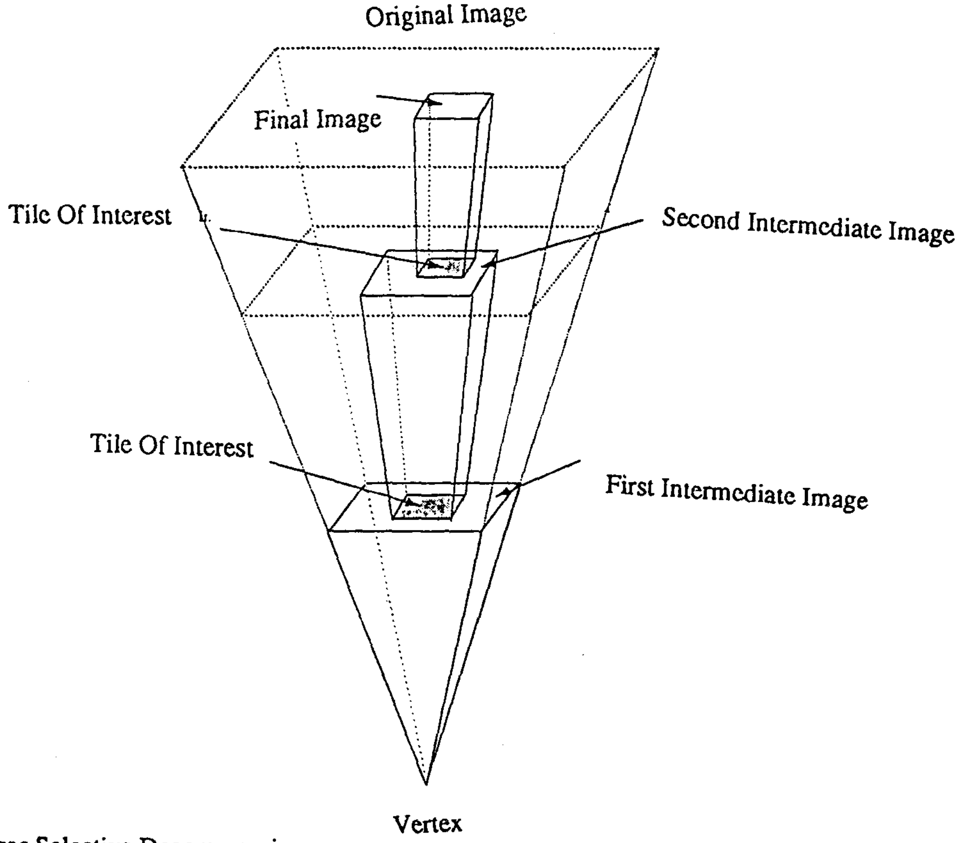


Figure 1 : 2- Stages Selective Decompression

A MULTI-START ALGORITHM FOR SIGNAL ADAPTIVE SUBBAND SYSTEMS

David Taubman and Avideh Zakhor[*]
Department of Electrical Engineering and Computer Sciences
University of California
Berkeley, CA 94720

ABSTRACT

There appears to be considerable motivation to investigate signal adaptive subband systems in which the polyphase transfer matrix is paraunitary, for application to still and motion image coding. This paper proposes a method of optimizing the filter coefficients of such systems using a mean-squared-error criterion. We have developed an efficient algorithm for locating numerous local optima in the coefficient space, permitting a degree of confidence in the location of globally optimal or near optimal solutions. We investigate both separable systems and a particular class of non-separable filter systems. The application of our algorithm to a number of images is described.

1. INTRODUCTION

Subband coding schemes have been proposed for image and video coding applications [1]. In addition, filter-decimate-interpolate schemes, which may be regarded as subband coding systems in which only one subband is kept, are the key element in pyramid coding schemes which have been proposed for ATV [2]. To date, however, comparatively little effort has been directed to signal adaptive systems of this form. In this paper our objective is to design subband filters which concentrate as much of the signal energy as possible in a single subband of a perfect reconstruction subband system with paraunitary polyphase transfer matrix[3]. The motivation for this objective is outlined in section 3. We concentrate on adapting relatively small filters and finding global optima. Particularly in this regard our approach differs from the work of Delsarte, et al. [4], who seek only a single locally optimal filter. The adaptation depends only on the second order statistics of the input signal and so may just as easily be applied to a class of signals as to a single signal. We consider both separable and non-separable two-dimensional systems, however computational constraints prevent us from applying our optimization algorithm to very general non-separable systems.

In outline, after demonstrating the motivation for our investigations in section 3, we summarize the equations and algorithmic features of our adaptation strategy in section 4. Some illustrative results are presented in section 5.

2. NOTATION

We denote sequences by $\langle\cdot\rangle$ and sets by $\{\cdot\}$. $l^2(\mathcal{Z})$ is the Hilbert space of square-summable sequences with the 2-norm denoted by $\|\cdot\|$. We write $span\{v_i\}$ for the smallest Hilbert space containing the vectors v_i. We adopt the usual notation $F(z)$ for the z-transform of a filter, F,whose impulse response is $\langle f(n)\rangle$. We define a filter with impulse response $\langle f(n)\rangle$ to be N-orthogonal if $\{\langle f(n-Nk)\rangle\}_k$ is an orthogonal set of vectors in $l^2(\mathcal{Z})$. By an invertible filter, we mean a filter with stable impulse response, $\langle f(n)\rangle$, where $F(z)$ has no zeroes on the unit circle – note $\langle f(n)\rangle$ and the impulse response of its inverse, $\langle f^{-1}(n)\rangle$, must satisfy $f * f^{-1} = \langle\delta(n)\rangle$ and may be two-sided in general. We signify the average over the sequence elements of $f = \langle f(n)\rangle$ as $\overline{f}$. Finally, a matrix of transfer functions, $\mathbf{E}(z_1,\ldots,z_n)$, is said to be paraunitary if $\mathbf{E}(z_1,\ldots,z_n)\mathbf{E}^T(z_1^{-1},\ldots,z_n^{-1}) = I$ [1, p 91].

3. MOTIVATION

In this paper we describe a signal adaptive technique for subband systems which have paraunitary polyphase transfer matrices, in which we optimize the analysis filters so as to maximize the energy in one of the subbands. In this section we make two observations, providing the motivation for this choice of system and optimization objective.

To begin with, consider the scheme outlined in Figure 1, which arises as a component of pyramid coding schemes as proposed for ATV coding [2], for example.

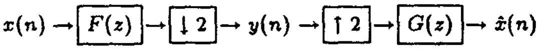


Figure 1: Filter-Decimate-Interpolate Scheme

If we are given filter G, it is natural to enquire how F must be chosen so that $\|x-\hat{x}\|$ is minimized. However, noting that the linear operator $T : l^2(\mathcal{Z}) \to l^2(\mathcal{Z})$, defined by $T(x) = \hat{x}$, depends only on G, having range space $span\{\langle g(n-2k)\rangle\}_k$, our question may be recast as: "How must F be chosen so that T is a projection operator onto its range space?" It has been shown that T is a projection whenever the system of Figure 1 is one of the two branches of a two-channel perfect reconstruction system with paraunitary polyphase transfer matrix [5]. Our first observation is, essentially, that this is necessarily the case. The proof is included in the Appendix.

Observation I *If F and G are FIR in the system of Figure 1 then T is a projection operator $\Longleftrightarrow$ F and G belong, to within a scale factor, to a two-channel FIR perfect reconstruction system with paraunitary polyphase transfer matrix. In this case G is 2-orthogonal.*

The above observation implies that every pair of FIR filters F and G for which T is a projection arise, to within a scale factor, as one of the branches of some two-channel perfect reconstruction system with FIR paraunitary polyphase transfer matrix.

Our second observation concerns subband coding. We consider the perfect reconstruction system of Figure 2 with FIR paraunitary polyphase transfer matrix. In the light of Observation I, then, T and $\tilde{T}$, given by $T(x) = \hat{x}_1$ and $\tilde{T}(x) = \hat{x}_2$, must be projection operators onto orthogonal subspaces of $l^2(\mathcal{Z})$. For simplicity we quantize y_1 and y_2 with the same uniform quantizer with quantization interval Δ and obtain $y_{1_q} = y_1 + q_1$

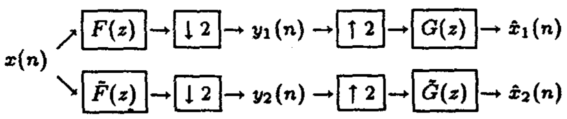


Figure 2: Two Channel Perfect Reconstruction Scheme

[*] This work was supported by National Science Foundation PYI award MIP–9057466. David Taubman was partially supported by a traveling scholarship from the University of Sydney, Australia.

0-7803-0532-9/92 $3.00 © 1992 IEEE

and $y_{2q} = y_2 + q_2$ where y_{1q}, y_{2q} are the quantized subband sequences and q_1, q_2 are the quantization errors. We will assume that y_1 and y_2 are wide sense stationary ergodic random processes, uncorrelated with q_1 and q_2 respectively, the latter having uniform distributions. Then, noting that $\{(g(n-2k))\}_k$ and $\{(\tilde{g}(n-2k))\}_k$ are orthonormal sets of vectors spanning orthogonal subspaces of $l^2(\mathcal{Z})$, we have

$$\overline{q^2} = \frac{1}{2}(E[q_1^2] + E[q_2^2]) = \frac{\Delta^2}{12},$$

$$\overline{x^2} = \frac{1}{2}(\overline{y_1^2} + \overline{y_2^2}) \text{ and } \overline{xq} = 0$$

If the sample rate of x is normalized to 1 so that y_1 and y_2 have normalized sample rates of $\frac{1}{2}$ then, applying the naive $4-\sigma$ rule[1] to the coding of y_{1q} and y_{2q} we obtain an overall bit rate of

$$\frac{1}{2}\left\{\left\lceil \log_2 \left\lceil \frac{8\sqrt{E[y_1^2]}}{\Delta} \right\rceil \right\rceil + \left\lceil \log_2 \left\lceil \frac{8\sqrt{E[y_2^2]}}{\Delta} \right\rceil \right\rceil \right\}$$

$$\approx \frac{1}{4}\log_2 \frac{2^{12}\overline{y_1^2}\,\overline{y_2^2}}{\Delta^4}$$

Since $\overline{y_1^2} + \overline{y_2^2} = \overline{x^2}$ is constant for a given signal, x, minimizing $\overline{y_1^2}\,\overline{y_2^2}$ corresponds to maximizing, or equivalently minimizing, $\overline{y_1^2}$. We have:

Observation II *In the subband system of Figure 2 with paraunitary polyphase transfer matrix, the bit rate is optimized by maximizing, or equivalently minimizing, $\overline{y_1^2}$ for a given input x.*

$\overline{y_1^2}$, then, becomes the objective function for our signal adaptive algorithm of section 4.

Observations I and II provide a common motivation to optimize perfect reconstruction filter systems with paraunitary polyphase transfer matrices so as to maximize the energy in a single subband, y_1.

4. ADAPTATION OF TWO-DIMENSIONAL SYSTEMS

Although we discussed only one-dimensional systems in section 3, the motivation derived therein applies also to two-dimensional systems of the form shown in Figure 3 in which $\mathbf{D}$ is a decimation matrix with determinant of 4 and the sequences are two-dimensional.

$$x(n_1,n_2) \rightarrow \boxed{F_1} \rightarrow \boxed{\downarrow \mathbf{D}} \rightarrow y_1 \rightarrow \boxed{\uparrow \mathbf{D}} \rightarrow \boxed{G_1} \rightarrow \hat{x}_1(n_1,n_2)$$

Figure 3: 2-D Filter-Decimate-Interpolate Scheme

In polyphase form the paraunitary analysis system equations may be written as

$$\vec{Y}(z_1,z_2) = \mathbf{E}(z_1,z_2)\vec{X}(z_1,z_2) \qquad (1)$$

where $\vec{y} = (y_1, y_2, y_3, y_4)^T$ is the vector of subband sequences, $\mathbf{E}(z_1,z_2)$ is the 4x4 paraunitary polyphase transfer matrix and $\vec{x} = (x_{00}, x_{01}, x_{10}, x_{11})^T$ is a vector of subsequences of x where each subsequence, $x_{ij}(n_1,n_2)$, is given by

$$x_{ij}(n_1,n_2) = x\left(\mathbf{D}\begin{pmatrix} n_1 \\ n_2 \end{pmatrix} - \begin{pmatrix} i \\ j \end{pmatrix}\right) \text{ for } i,j \in \{0,1\}$$

When $\mathbf{E}$ and $\mathbf{D}$ are separable F_1 is separable and so the applicability of the one-dimensional results is clear. When $\mathbf{E}$ is a separable matrix, the one dimensional results are also directly applicable, even when $\mathbf{D}$ and hence F_1 are not separable, because it is always possible to reorganize the sequence $x(n_1,n_2)$ so that we need only consider $\mathbf{D} = 2\mathbf{I}$.

We now describe our optimization equations and algorithm.

4.1. The Equations

We wish to adapt the analysis system of equation (1) so as to maximize $\overline{y_1^2}$. Clearly this involves optimization of both the decimation matrix, $\mathbf{D}$, and the polyphase transfer matrix, $\mathbf{E}$. As noted above, however, we can always reorganize the sequence x so that the decimation matrix can be taken as $\mathbf{D} = 2\mathbf{I}$ for the purpose of optimizing $\mathbf{E}$. In this way we are able to concentrate on the optimization of $\mathbf{E}$, leaving that of $\mathbf{D}$ for future investigation.

We would like to consider cascade structures of the form

$$\mathbf{E}(z_1,z_2) = \mathbf{H}_N\Lambda(z_1,z_2)\mathbf{H}_{N-1}\cdots\Lambda(z_1,z_2)\mathbf{H}_1. \qquad (2)$$

$\Lambda(z_1,z_2) = \mathrm{diag}(1, z_1^{-1}, z_2^{-1}, z_1^{-1}z_2^{-1})$ and $\mathbf{H}_1,\ldots,\mathbf{H}_N$ are 4x4 orthogonal coefficient matrices, leading to a paraunitary system with $6N$ degrees of freedom. It is straightforward to see that $\mathbf{E}$ is a *separable* paraunitary FIR polyphase transfer matrix if and only if it satisfies equation (2) with each $\mathbf{H}_n$ of the form

$$\mathbf{H}_n = \begin{pmatrix} C_{\alpha_n}C_{\beta_n} & -C_{\alpha_n}S_{\beta_n} & -S_{\alpha_n}C_{\beta_n} & S_{\alpha_n}S_{\beta_n} \\ C_{\alpha_n}S_{\beta_n} & C_{\alpha_n}C_{\beta_n} & -S_{\alpha_n}S_{\beta_n} & -S_{\alpha_n}C_{\beta_n} \\ S_{\alpha_n}C_{\beta_n} & -S_{\alpha_n}S_{\beta_n} & C_{\alpha_n}C_{\beta_n} & -C_{\alpha_n}S_{\beta_n} \\ S_{\alpha_n}S_{\beta_n} & S_{\alpha_n}C_{\beta_n} & C_{\alpha_n}S_{\beta_n} & C_{\alpha_n}C_{\beta_n} \end{pmatrix}$$

with C_{α_n} and S_{α_n} standing for $\cos\alpha_n$ and $\sin\alpha_n$ etc. The separable case, then, is characterized by $2N$ degrees of freedom.

From equation (1) we immediately write down $Y_1(z_1,z_2) = E_{1,1}(z_1,z_2)X_{00}(z_1,z_2) + \cdots + E_{1,4}(z_1,z_2)X_{11}(z_1,z_2)$ and so $\overline{y_1^2}$ depends only on the first row of $\mathbf{E}$ and the second order statistics of x. In fact, from equation (2), $\overline{y_1^2}$ is a trigonometric polynomial in the $6N$ rotation angles which characterize the orthogonal matrices $\mathbf{H}_N,\ldots,\mathbf{H}_1$, the coefficients of this polynomial being determined by the second order statistics of the two-dimensional signal, x. These statistics may be evaluated for a particular signal or for a class of signals. In numerical work the trigonometric polynomial itself is generated symbolically for a given size filter.

4.2. The Algorithm

It is well-known that the optimization of filter systems is quite dependent on an "initial guess" for numerical convergence. It is common to start with an unconstrained filter designed to satisfy desirable frequency response specifications and then perform a constrained numerical optimization. Such a technique is not well suited to the signal adapted optimization problem at hand. Delsarte, et al. [4] find a single local optimum using a fixed starting point for a one-dimensional signal adaptive subband system with paraunitary polyphase transfer matrix. Their objective is identical to ours – to maximize $\overline{y_1^2}$. Our approach, however, is to pursue a globally optimal filter system.

We have already noted that the objective function is a trigonometric polynomial in, say, R degrees of freedom (angles). The local optima, then, are a subset of the set of zeroes of the gradient of this function – i.e. the common zeroes of the set of R trigonometric polynomials corresponding to each of the partial derivatives of this objective function. As such, homotopy methods [6] could in theory be used to find all local optima and hence the global optimum. The astronomical number of such common zeros, mostly complex-valued, however, render such an approach impractical.

Our optimization technique involves collecting local optima from the $R-1$ dimensional problem derived by holding the R'th angle fixed at each of, say, M values from some fixed grid. A numerical tracking algorithm is then applied to each of these solutions in order to find a set of R dimensional local optima.

The tracking algorithm allows us to move from a local optimum in the $R-1$ dimensional problem, derived by holding the R'th angle fixed, to a local optimum in the full R angles. It involves tracking the $R-1$ dimensional local optimum as the R'th angle is stepped so as to increase the objective function, $\overline{y_1^2}$. Tracking involves prediction and step size adaptation to maximize efficiency and incorporates many measures to avoid losing the path of local optima in $R-1$ dimensions. A Newton-Raphson technique is used to stay on this path and also to converge to the

[1] In the $4-\sigma$ rule we hard-limit the signal to within ± 4 standard deviations before quantizing

full R dimensional solution once the gradient vector becomes sufficiently small.

The recursive approach outlined has two noteworthy advantages: First it is easy to guarantee that we converge no more than twice to any given local optimum. The tracking of a solution in $R-1$ variables as the R'th variable is moved, is what allows us to avoid converging to the same local optimum again and again – a serious problem with naive multi-start algorithms. It is this same feature (tracking) which allows homotopy methods to arrive at all local optima exactly once. The other key advantage of our algorithm over naive multi-start approaches is the fact that most of the work is done in solving lower dimensional problems, where numerical complexity is much lower. There is also an efficient way to derive the $R-1$ dimensional objective function from the R dimensional objective function. In this way we converge to numerous distinct local optima. The confidence with which we find a globally optimal solution depends on the size of M and hence on the time we are prepared to spend in optimization.

5. RESULTS

In experimental work we have optimized separable filters for $N \leq 7$. The object in our experiments was to compare the performance of the adapted filters with a non-adaptive approach. The experimental context is shown in Figure 4, which may be viewed as a pyramidal subband decomposition structure with L levels, in which all but one of the 4^L branches have been discarded. Although this is not in itself a useful scheme for image compression, it allows a qualitative evaluation of the benefits of filter adaptation. The separable fixed filters with which we compared our adapted filters were taken from [7]. As discussed in section 4, we always use $\mathbf{D} = 2\mathbf{I}$ for this work.

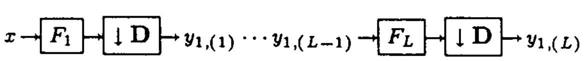


Figure 4: Experimental Context

The numerical results tabulated indicate the energy captured in sequence $y_{1,(l)}$ with the energy in the original sequence, x, normalized to 1. The "fixed" values were generated by applying the appropriate filters from [7] and keeping the subband with the most energy at each stage. The "adapted" values were generated by using individually adapted filters at each stage.

The result of applying our algorithm to the carpet texture of Figure 5 is shown in table 1. As seen, the adaptive algorithm improves the energy compaction at the third level by 60 percent. Figure 6 shows the magnitude response of the row filter used to generate $y_{1,(3)}$, which is clearly distinct from either a high- or low-pass filter – the only choices in a non-adaptive system. We observed similar performance in other highly regular textures. For example, when applied to the 128x128 image of a brick wall in perspective, energy compaction at the third level improves by 70 percent - see table 2. As expected, the same effect was observed on an artificially generated sinusoidal grating with 45° orientation and frequency of 0.5 rad/pixel – see table 3. On complex images, however, the signal-adapted filters only slightly outperformed the fixed filters.

The general trend with images in which the adaptation did have a significant effect was that the adaptive filters would perform similarly to the fixed filters at each stage, l, until some point at which we would notice a considerable difference. This transition point is easily understood as the point at which regular features in $y_{1,(l-1)}$ may be captured by the relatively small region of support of the filters. The transition was seen to occur for lower values of l when the filter sizes were increased, exactly as one would expect, leading to the obvious conclusion that we should adapt filters with as large a support as possible.

Table 1: Separable Filter Results for "Carpet" Texture

l	N	Size of $y_{1,(l)}$	$\overline{y^2_{1,(l)}}$ (fixed)	$\overline{y^2_{1,(l)}}$ (adapted)
1	6	262x262	0.8771	0.8879
2	6	137x137	0.5270	0.6095
3	6	74x74	0.2879	0.4582

Table 2: Separable Filter Results for Brick Wall in Perspective

l	N	Size of $y_{1,(l)}$	$\overline{y^2_{1,(l)}}$ (fixed)	$\overline{y^2_{1,(l)}}$ (adapted)
1	6	70x70	0.4496	0.5031
2	6	41x41	0.1780	0.2715
3	7	27x27	0.0877	0.1498

In non-separable work we considered only the case in which $\mathbf{H}_1, \ldots, \mathbf{H}_{N-1}$ in equation (2) are separable matrices and $\mathbf{H}_N$ is non-separable. Only three of the six angles which characterize $\mathbf{H}_N$ need affect its first row and hence the first row of E, and hence y_1, from equations (1) and (2). In this way we have $2N+1$ angles in the objective function for $\overline{y_1^2}$, only one more than the $2N$ of the separable case. Note that the fact that three of the angles which characterize $\mathbf{H}_N$ do not affect y_1 means that these can be independently adjusted to optimize y_2, y_3, y_4 in a full subband system. Experimental results were obtained for these non-separable filters with $N = 4$. We found optimization of larger filters, with any confidence of being close to a global optimum, to be prohibitive due to computational complexity.

It was observed that non-separable filters barely outperformed their separable counterparts, which we attribute to the fact that we were only able to optimize $2N+1$ degrees of freedom, just one more than in the separable case, rather than the full $6N$ of equation (2). Even on images with distinctly non-separable spectra, non-separable features were observed only outside the passband of the adapted filters.

6. CONCLUSIONS

There is strong motivation in image coding applications to adapt the parameters of perfect reconstruction subband systems with paraunitary polyphase transfer matrices to a given signal or class of signals. Our experimental results indicate that adaptable separable filters result in significant energy compaction for images of a highly regular nature. In order to allow orientational tuning with a separable polyphase transfer matrix it should also be important to adapt the decimation matrix to optimize the second order statistics of the vector $\tilde{x}$ in equation (1) prior to adapting the $2N$ angles which characterize a separable matrix, $\mathbf{E}$. This phase of our research into adaptive filter-decimate-interpolate schemes remains to be investigated.

REFERENCES

[1] *Subband Image Coding, John W. Woods Editor*, Kluwer Academic Publishers, 1990.

[2] K. Metin Uz, Martin Vetterli, and Didier J. LeGall. "Interpolative Multiresolution Coding of Advanced Television with Compatible Subchannels," *IEEE Transactions on Circuits and Systems for Video Technology*. Vol 1, No 1, March 1991. pp 86-99.

[3] R. Rinaldo, D. Taubman and A. Zakhor. "Applications of Multi-Resolution Analysis to Images," *Seventh Workshop on Multidimensional Signal Processing*. September 1991, Lake Placid, New York.

[4] Philippe Delsarte, Benoit Macq, and Dirk T.M. Slock. "Efficient Multiresolution Signal Coding via a Signal-Adapted Perfect Reconstruction Filter Pyramid," *Proceedings of the International Conference on ASSP, Toronto 1991*. pp 2633-2636.

Table 3: Separable Filter Results for Sinusoidal Grating

l	N	Size of $y_{1,(l)}$	$\overline{y^2_{1,(l)}}$ (fixed)	$\overline{y^2_{1,(l)}}$ (adapted)
1	4	260x260	0.9985	0.9994
2	4	134x134	0.8883	0.9950
3	4	71x71	0.6840	0.9808
4	4	39x39	0.6478	0.9696
5	6	25x25	0.2969	0.7405

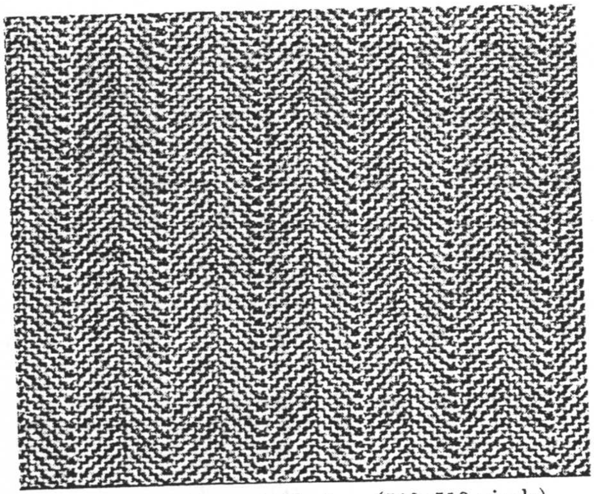


Figure 5: "Carpet" Texture (512x512 pixels)

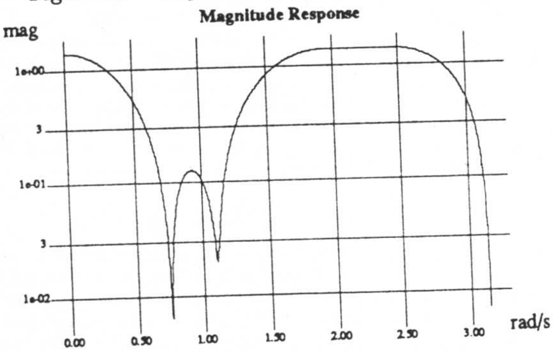


Figure 6: Adapted Row Filter for "Carpet" Texture

[5] Olivier Rioul and Martin Vetterli. "Wavelets and Signal Processing," *IEEE Signal Processing Magazine*. Vol 8, No. 4, October 1991. pp 14-38.

[6] L T Watson and R T Haftka. "Modern Homotopy Methods in Optimization," *Computer Methods in Applied Mechanics and Engineering*. Vol 74. Sept. 1989.

[7] Ingrid Daubechies. "Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*. Vol 41. 1988. pp 909-996.

A APPENDIX

We prove the observation stated in section 3. To see this, consider the following theorem, which is important in itself as it is not restricted to FIR filters:

Theorem *Given the invertible filter, G, there exists a unique filter, F, such that T is a projection operator onto its range space. Moreover, if $G(z)G(z^{-1}) + G(-z)G(-z^{-1}) \neq 0$ on the unit circle then F is given by*

$$F(z) = \frac{G(z^{-1})}{\frac{1}{2}[G(z)G(z^{-1}) + G(-z)G(-z^{-1})]} \tag{3}$$

Proof:

Observe that for input $\langle\delta(i)\rangle$ we get $F(\langle\delta(i)\rangle) = \langle f(n)\rangle$ and so we must have

$$\begin{aligned} f(2n) &= G^{-1}(T(\langle\delta(i)\rangle))(2n) \\ \text{and } f(2n+1) &= G^{-1}(T(\langle\delta(i+1)\rangle))(2n) \end{aligned} \tag{4}$$

demonstrating the existence and uniqueness of F.

All we need now is an orthonormal basis for $T(l^2(\mathcal{Z}))$ and the determination of F from equation (4) is then straightforward. We will demonstrate such a basis, $\{u_k\}_k$ with $u_k = \langle u_k(i)\rangle = \langle u_0(i-2k)\rangle$ or, equivalently,

$$U_0(z)U_0(z^{-1}) + U_0(-z)U_0(-z^{-1}) = 2$$

Noting that $u_0 \in T(l^2(\mathcal{Z})) = span\{\langle g(n-2k)\rangle\}_k$, we need only find a stable sequence $\langle a(n)\rangle$ satisfying

$$A(z^2)A(z^{-2}) = \frac{2}{G(z)G(z^{-1}) + G(-z)G(-z^{-1})} \tag{5}$$

and then put

$$u_0 = \sum_l \langle g(n-2l)\rangle a(l) \implies U_0(z) = A(z^2)G(z)$$

If $G(z)$ is a rational polynomial in z then $G(z)G(z^{-1}) + G(-z)G(-z^{-1})$ is a rational polynomial in z^2 and so $A(z)$ may be found by spectral factorization. Using this orthonormal basis, $u_k = \sum_l \langle g(n-2k-2l)\rangle a(l)$, we obtain

$$\begin{aligned} T(\langle\delta(i)\rangle)(n) &= \sum_{k,l,m} g(-2k-2l)g(n-2k-2m)a(l)a(m) \\ &= G(\langle y(i)\rangle) \end{aligned}$$

where

$$y(2i+1) = 0 \text{ and } y(2i) = \sum_{l,m} a(l)a(m)g(-2i+2m-2l)$$

$$\implies f(2n) = \sum_{l,m} a(l)a(m)g(-2n+2m-2l)$$

Similarly we find the odd subsequence of $\langle f(n)\rangle$ by considering $T(\langle\delta(i+1)\rangle)$ and, combining, we obtain

$$\begin{aligned} f(n) &= \sum_{l,m} a(l)a(m)g(-n+2m-2l) \\ \implies F(z) &= A(z^2)A(z^{-2})G(z^{-1}) \end{aligned} \tag{6}$$

Equations (5) and (6) yield the result – equation (3). □

In the FIR case, simple manipulation of equation (3) yields:

Corollary *If G is an invertible FIR filter and T is a projection operator then F is an FIR filter $\iff$ $G(z)G(z^{-1}) + G(-z)G(-z^{-1})$ is a constant – i.e. G is 2-orthogonal.*

We now demonstrate that G 2-orthogonal is equivalent, to within a scale factor, to requiring G to be one of the synthesis filters of a two-channel perfect reconstruction system with paraunitary polyphase transfer matrix – see Figure 2. For, if $G_0(z)$ and $G_1(z)$ are the polyphase components of $G(z)$ – i.e. $g_0(n) = g(2n)$ and $g_1(n) = g(2n-1)$ – then G_0 and G_1 are power complementary filters because

$$\begin{aligned} A\delta(k) &= \sum_i g(i)g(i-2k) \\ &= \sum_i [g_0(i)g_0(i-k) + g_1(i)g_1(i-k)] \end{aligned}$$

for some constant A. Then, setting

$$\tilde{G}_0(z) = -G_1(z^{-1}) \text{ and } \tilde{G}_1(z) = G_0(z^{-1}) \tag{7}$$

it is clear that the matrix

$$\frac{1}{\sqrt{A}}\begin{pmatrix} G_0(z) & \tilde{G}_0(z) \\ G_1(z) & \tilde{G}_1(z) \end{pmatrix}$$

is an FIR paraunitary polyphase matrix. Let $\tilde{G}_0$ and $\tilde{G}_1$, defined in equation (7), be the polyphase components of filter $\tilde{G}$ in Figure 2 and scale all filters by $\frac{1}{\sqrt{A}}$. Then, with $\tilde{F}(z) = \tilde{G}(z^{-1})$ it is easily verified that the two branches of Figure 2 project x onto orthogonal subspaces and hence $\hat{x}_1 + \hat{x}_2 = x$. So we have a perfect reconstruction subband system with FIR paraunitary polyphase transfer matrix.