Correspondence

An Error Analysis for 2-D Block Implemented Digital Filters

MAHMOOD R. AZIMI-SADJADI AND ROBERT A. KING

Abstract—**This correspondence considers error analysis of block-implemented 2-D digital filters. Expressions for error bound and mean-square error for roundoff error accumulation are derived using fixed-point arithmetic, and compared with the results obtained using ordinary 2-D difference equations.**

I. INTRODUCTION

Block implementation technique has attracted considerable attention in recent years, applied almost exclusively to 1-D digital filters [1]. The primary motivation behind this study was to develop a parallel processing model for recursive digital filters, with increased data-throughput rate and reduced computational effort. Subsequent work [2], [3], however, reveals several other fruitful features of this method which stem from its unique structure.

In the 1-D case, Barnes and Shinnaka [3] have shown that the roundoff noise variance for a 1-D fixed point digital filter, realized by a block-state structure, is reduced by a factor equal to the block length when compared with the noise variance in the simple state-space realization model [4]. Moreover, the dynamic ranges of the state variables are preserved under the transformation from a simple state-space structure to a block-state structure.

The block implementation technique for 2-D recursive digital filters was extended by Azimi-Sadjadi [5], [6]. The method is particularly useful when the linear filtering of an image is performed by a 2-D recursive digital filter.

In addition to the advantages of 2-D block processing in the implementation, several other benefits relating to its structure can also be exploited, as in the 1-D case.

In this paper, a bound on the norm of the error produced due to roundoff of multiplications is derived for a 2-D block-implemented digital filter. This bound is shown to be considerably smaller than that obtained by Ni and Aggarwal [7] when the filter is implemented using ordinary 2-D difference equations.

The method of analysis adapted here is analogous to that developed in [7], [8] but using a matrix model. The technique of Barnes *et al.* cannot easily be extended to the 2-D case, because of the complexity of the state matrices in the 2-D block-state realization model [9].

II. 2-D BLOCK RECURSIVE EQUATION

A two-dimensional recursive digital filter is described by the linear difference equation

Manuscript received June 26, 1981; revised April 1, 1982, June 7, 1983, and July 6, 1983.

M. R. Azimi-Sadjadi is with the Department of Electrical Engineering, State University of New York College of Technology, Utica, NY 13502.

R. A. King is with the Department of Electrical Engineering, Imperial College of Science and Technology, London SW7 2BT, England.

$$y_{m,n} = \sum_{i=0}^{M_b} \sum_{j=0}^{N_b} b_{i,j} x_{m-i,n-j} - \sum_{\substack{i=0 \\ i+j \neq 0}}^{M_a} \sum_{j=0}^{N_a} a_{i,j} y_{m-i,n-j} \tag{1}$$

where $\{x_{m,n}\}$ and $\{y_{m,n}\}$ are the input and output sequences, respectively.

Consider the input sequence $\{x_{m,n}\}$ to be partitioned into nonoverlapping blocks of dimension K by L, where

$$K \geqslant \max(M_a, M_b), \quad L \geqslant \max(N_a, N_b). \tag{2}$$

Now, if these blocks are arranged as vectors with element subscripts ordered lexicographically, (1) can be written in a matrix form. Using this formulation, the following "2-D block recursive equation" can be obtained [5], [6].

$$\sum_{m=0}^{1} \sum_{n=0}^{1} D_{mn} X(i-m, j-n) - \sum_{m=0}^{1} \sum_{n=0}^{1} C_{mn} Y(i-m, j-n) = 0. \tag{3}$$

In this equation, $X(i,j) \triangleq X_{i,j}$ may be defined as

$$X_{i,j} = [\hat{X}_{iK}^{(j)} \hat{X}_{iK+1}^{(j)} \cdots \hat{X}_{(i+1)K-1}^{(j)}]^t$$

where

$$\hat{X}_l^{(j)} = [x_{l,jL}\, x_{l,jL+1} \cdots x_{l,(j+1)L-1}]^t \tag{4}$$

and similarly for $Y(i,j) \triangleq Y_{i,j}$. Matrices C_{ij} are either lower or upper triangular block Toeplitz, i.e.,

$$C_{00} = \begin{bmatrix} A_0 & & & & & \\ A_1 & A_0 & & & \bigcirc & \\ \vdots & & \ddots & & & \\ A_{M_a} & A_{M_a-1} & \cdots & A_0 & & \\ 0 & A_{M_a} & \cdots & & A_0 & \\ \vdots & \bigcirc & & \ddots & & \ddots \\ & & & A_{M_a} & \cdots & A_0 \end{bmatrix}_{(KL \times KL)} \tag{5a}$$

and

$$C_{10} = \begin{bmatrix} 0 & 0 \cdots A_{M_a} & \cdots & A_1 \\ & & A_{M_a} & \cdots A_2 \\ & & & \ddots \vdots \\ & & & A_{M_a} \\ & \bigcirc & & 0 \\ & & & \vdots \\ & & & 0 \end{bmatrix}_{(KL \times KL)} \tag{5b}$$

This also holds for C_{01} and C_{11}, where the constituent block matrices are A_i''s. The block matrices A_i and A_i' are also lower and upper triangular Toeplitz, respectively, and may be defined similarly in terms of the coefficients $a_{i,j}$ of the difference equation [5], [6]. Matrices D_{ij} can be defined, in a simi-

0096-3518/83/1200-1570$01.00 © 1983 IEEE

lar manner, in terms of B_i and B_i', where these latter matrices may be defined in terms of $b_{i,j}$ similar to A_i and A_i'.

III. Error Analysis

As a consequence of input, coefficient, and product quantization, the actual filter implemented by a finite wordlength machine is represented by

$$\sum_{m=0}^{1}\sum_{n=0}^{1}[\tilde{D}_{mn}\tilde{X}(i-m,j-n)]_r - \sum_{m=0}^{1}\sum_{n=0}^{1}[\tilde{C}_{mn}\tilde{Y}(i-m,j-n)]_r = \mathbf{0} \tag{6}$$

where $[\cdot]_r$ indicates rounding and superscript $(\tilde{\ })$ denotes the actual quantized value of the matrices or vectors, i.e.,

$$\tilde{D}_{mn} = D_{mn} + \Delta D_{mn}, \quad \tilde{C}_{mn} = C_{mn} + \Delta C_{mn}$$

$$\tilde{X}(i-m,j-n) = X(i-m,j-n) + \Delta X(i-m,j-n)$$

$$\tilde{Y}(i-m,j-n) = Y(i-m,j-n) + \Delta Y(i-m,j-n). \tag{7}$$

Subtracting (3) from (6) and using (7) yields

$$\begin{aligned}&\sum_{m=0}^{1}\sum_{n=0}^{1}\tilde{D}_{mn}\Delta X(i-m,j-n)\\ &-\sum_{m=0}^{1}\sum_{n=0}^{1}\tilde{C}_{mn}\Delta Y(i-m,j-n)\\ &+\sum_{m=0}^{1}\sum_{n=0}^{1}\Delta D_{mn}X(i-m,j-n)\\ &-\sum_{m=0}^{1}\sum_{n=0}^{1}\Delta C_{mn}Y(i-m,j-n)\\ &+\sum_{m=0}^{1}\sum_{n=0}^{1}\beta_{mn}(i-m,j-n)\\ &-\sum_{m=0}^{1}\sum_{n=0}^{1}\alpha_{mn}(i-m,j-n) = \mathbf{0}.\end{aligned} \tag{8}$$

In (8), the first term represents the effect of input quantization, the third and fourth terms are due to coefficient quantization, and the fifth and sixth terms give the errors due to the quantization of the product; the effect of these on the output blocks is determined by the second term of this equation.

In the following analysis, attention has been focused only on product quantization error. The effects of input and coefficient quantization have been neglected.

The roundoff error vectors $\alpha_{mn}(i-m,j-n)$ result from the multiplications of C_{mn} by $Y(i-m,j-n)$. For example,

$$\alpha_{00}(i,j) = [C_{00}Y(i,j)]_r - C_{00}Y(i,j)$$

$$= \begin{bmatrix} \alpha_0(iK,j) \\ \alpha_1(ik,j) + \alpha_0(iK+1,j) \\ \vdots \\ \sum_{m=0}^{M_a} \alpha_m(iK+M_a-m,j) \\ \vdots \\ \sum_{m=0}^{M_a} \alpha_m(iK+K-1-m,j) \end{bmatrix}. \tag{9}$$

Error vectors $\alpha_{10}(i-1,j)$, $\alpha_{01}(i,j-1)$, and $\alpha_{11}(i-1,j-1)$ may be defined in a similar manner. The constituent vectors $\alpha_m(l,j)$ and $\alpha_m'(l,j-1)$ which result from the multiplications of A_m and A_m' by $\hat{Y}_l^{(j)}$ and $\hat{Y}_l^{(j-1)}$ may be defined as

$$\alpha_m(l,j) = [A_m\hat{Y}_l^{(j)}]_r - A_m\hat{Y}_l^{(j)}$$

$$= \begin{bmatrix} \delta_{m,0}(l,jL) \\ \delta_{m,1}(l,jL) + \delta_{m,0}(l,jL+1) \\ \vdots \\ \sum_{n=0}^{N_a} \delta_{m,n}(l,jL+N_a-n \\ \vdots \\ \sum_{n=0}^{N_a} \delta_{m,n}(l,jL+L-1-n) \end{bmatrix} \tag{10}$$

where

$$\delta_{m,n}(p,q) = [a_{m,n}y_{p,q}]_r - a_{m,n}y_{p,q}$$

and

$$|\delta_{m,n}(p,q)| < 2^{-t}$$

for a word containing t bits. Vectors $\alpha_m'(l,j-1)$ may be defined in a similar manner.

A similar procedure can be repeated for the roundoff error vectors $\beta_{mn}(i-m,j-n)$.

Now define an error vector $E(i,j)$ as

$$E(i,j) = \sum_{m=0}^{1}\sum_{n=0}^{1}\beta_{mn}(i-m,j-n) - \sum_{m=0}^{1}\sum_{n=0}^{1}\alpha_{mn}(i-m,j-n). \tag{11}$$

The effect of this error on the output may be written as $V(i,j) \triangleq \Delta Y(i,j)$ such that

$$E(i,j) = \sum_{m=0}^{1}\sum_{n=0}^{1} C_{mn}V(i-m,j-n). \tag{12}$$

Vectors $V_{i,j} \triangleq V(i,j)$ and $E_{i,j} \triangleq E(i,j)$ may be defined similarly to $X_{i,j}$ in (4).

If the input array is of dimensions $P \times Q$, it may be represented by a sequence of period P, Q; it is composed of blocks of size $K \times L$, and hence, the blocks have a periodicity of M, N where

$$M \triangleq \frac{P}{K}, \quad N \triangleq \frac{Q}{L}.$$

Similarly, $V(i,j)$ and $E(i,j)$ blocks will be periodic with period M, N. The (z, w)-transforms of these block sequences may be defined as

$$\begin{aligned}\hat{V}(z,w) &= (z,w)\{V_{i,j}\}\\ &\triangleq [V_{0,0}(z,w)\ V_{0,1}(z,w)\\ &\quad \cdots V_{0,L-1}(z,w)\ V_{1,0}(z,w) \cdots V_{1,L-1}(z,w)\\ &\quad \cdots V_{K-1,L-1}(z,w)]^t\end{aligned} \tag{13}$$

where $V_{r,s}(z,w)$, $r \in [0, K-1]$, and $s \in [0, L-1]$ is the (z, w)-transform of the sampled version sequence formed by the (r, s)th elements of all the blocks, i.e.,

$$V_{r,s}(z, w) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} \tilde{v}_{iK+r, jL+s} z^{-i} w^{-j}. \quad (14)$$

From (13) and (14), we may write

$$\hat{V}(z, w) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} V_{i,j} z^{-i} w^{-j}. \quad (15)$$

The (z, w)-transform of the sequence $\{E_{i,j}\}$ may be defined likewise. Now, taking the (z, w)-transform of both sides of (12) yields

$$\hat{E}(z, w) = \hat{G}(z, w) \cdot \hat{V}(z, w) \quad (16)$$

where $\hat{G}(z, w)$ is the (z, w)-transform of $\{C_{mn}\} = \{C_{00}, C_{10}, C_{01}, C_{11}\}$, i.e.,

$$\hat{G}(z, w) = C_{00} + C_{10} z^{-1} + C_{01} w^{-1} + C_{11} z^{-1} w^{-1}. \quad (17)$$

Note that matrix $\hat{G}^{-1}(z, w)$ is the block matrix transfer function of the corresponding all-pole filter.

IV. A Bound on the Norm of Error

Since $V(i, j)$ and $E(i, j)$ are considered to be periodic, their rms values are

$$\langle V \rangle \triangleq \frac{1}{MN} \left[\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \|V(i, j)\|^2 \right]^{1/2}$$

and

$$\langle E \rangle \triangleq \frac{1}{MN} \left[\sum_{i=0}^{M-1} \sum_{j=0}^{N-1} \|E(i, j)\|^2 \right]^{1/2} \quad (18)$$

where

$$\|X\| \triangleq (X^t \cdot X)^{1/2} = \left(\frac{1}{N} \sum_{i=1}^{N} |x_i|^2 \right)^{1/2}.$$

Theorem: If $\xi_{h,l}\{\exp(\jmath\phi_1), \exp(\jmath\phi_2)\}$, $h \in [0, K-1]$, $l \in [0, L-1]$ are the eigenvalues of $\hat{G}^{-1}\{\exp(\jmath\phi_1), \exp(\jmath\phi_2)\}$, then

$$\min_{\phi_1, \phi_2, h, l} \left| \xi_{h,l}\{\exp(\jmath\phi_1), \exp(\jmath\phi_2)\} \right| \langle E \rangle \leqslant \langle V \rangle \leqslant \max_{\phi_1, \phi_2, h, l} \left| \xi_{h,l}\{\exp(\jmath\phi_1), \exp(\jmath\phi_2)\} \right| \langle E \rangle \quad (19)$$

where

$$\phi_1 = \frac{2\pi m}{M}, \quad \phi_2 = \frac{2\pi n}{N}, \quad m \in [0, M-1] \; n \in [0, N-1].$$

Proof: The proof is straightforward and, thus, is omitted here [8].

Now, since C_{00}, C_{10}, C_{01}, and C_{11} are block Toeplitz matrices, it can be shown that $\hat{G}\{\exp(\jmath\phi_1), \exp(\jmath\phi_2)\}$ is a $\{\theta\}$-block circulant matrix [5] where $\theta \triangleq \exp(-\jmath\phi_1)$; also, the elemental blocks of each C_{ij} are themselves Toeplitz, and so the blocks of $\hat{G}$ are $\{\Omega\}$-circulant where $\Omega \triangleq \exp(-\jmath\phi_2)$. Using the properties of these matrices [5], [10], it can be shown that the eigenvalues of $\hat{G}$ are the elements of the 2-D DFT of size $P \times Q$ of sequence $\{a_{p,q}\}$. As a result, the eigenvalues of $\hat{G}^{-1}$ are

$$\xi_{h',l'} = 1 \Big/ \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} a_{p,q} W_P^{ph'} W_Q^{ql'} \quad (20)$$

where

$$W_P \triangleq \exp\left(-j \frac{2\pi}{P}\right), \; W_Q \triangleq \exp\left(-j \frac{2\pi}{Q}\right)$$

and

$$\begin{aligned} h' &= hM + m; & h' &\in [0, P-1] \\ p &= iK + r; & p &\in [0, P-1] \\ l' &= lN + n; & l' &\in [0, Q-1] \\ q &= jL + s; & q &\in [0, Q-1]. \end{aligned}$$

Therefore, (19) becomes

$$\langle V \rangle \leqslant \max_{h', l'} \left[\frac{1}{\sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} a_{p,q} \exp\left(-2\pi j \left[\frac{h'p}{P} + \frac{l'q}{Q}\right]\right)} \right] \langle E \rangle. \quad (21)$$

Now, in order to express $\langle E \rangle$, take the norm of E in (11). Thus,

$$\|E(i, j)\| \leqslant \sum_{m=0}^{1} \sum_{n=0}^{1} \|\beta_{mn}(i-m, j-n)\| + \sum_{m=0}^{1} \sum_{n=0}^{1} \|\alpha_{mn}(i-m, j-n)\|. \quad (22)$$

Using the definitions of α_{mn}'s and β_{mn}'s and their constituent vectors α_m, α'_m, β_m, and β'_m in (9) and (10), and also considering that $a_{0,0} = 1$, the upper bound on the norm of error vector $E(i, j)$ may be found as

$$\|E(i, j)\|_{\text{upper}} < 2^{-t} [\tfrac{3}{2}(\hat{M}_a \hat{N}_a + \hat{M}_b \hat{N}_b - 1)] \quad (23a)$$

where

$$\hat{M}_a = M_a + 1, \quad \hat{N}_a = N_a + 1$$
$$\hat{M}_b = M_b + 1, \quad \hat{N}_b = N_b + 1.$$

This upper bound occurs when $K = 2M_a + 1 = 2M_b + 1$ and $L = 2N_a + 1 = 2N_b + 1$. The lower bound, on the other hand, occurs when $K = M_a = M_b$ and $L = N_a = N_b$, i.e., for minimum values of the block sizes. This bound is found to be

$$\|E(i, j)\|_{\text{lower}} \leqslant 2^{-t} [\tfrac{4}{3}(\hat{M}_a \hat{N}_a + \hat{M}_b \hat{N}_b - 1)]. \quad (23b)$$

It is interesting to note that the computational time and computer memory allocations for 2-D block processing would also become minimum when the minimum values for block sizes are chosen [6].

Now, in order to demonstrate the advantage of the block implementation technique over the direct recursion method using 2-D difference equations, it is necessary to show that the upper bound on the mean-square value of the corresponding scalar error sequence is less than that of direct method. Combine (18), (21), and (23) and divide $\langle V \rangle$ by $(KL)^{1/2}$ to obtain the rms values of the relevant scalar error sequence $\langle v \rangle$. This would result in

$$\langle v \rangle_{\text{upper}} < \frac{3}{2} \frac{\langle e \rangle}{(KL)^{1/2}} \quad (24)$$

where $\langle e \rangle$ is the bound on the roundoff error for the direct filtering process [7, expression (32)] when the sequences are assumed to be periodic, i.e.,

$$\langle e\rangle \leqslant \max_{\psi_1,\psi_2} \left[1 \Big/ \sum_{p=0}^{P-1} \sum_{q=0}^{Q-1} a_{p,q} \cdot \exp\left[-j(p\psi_1 + q\psi_2)\right] \right] \langle e^{(d)}\rangle \tag{25}$$

where

$$\langle e^{(d)}\rangle \leqslant 2^{-t}(\hat{M}_a\hat{N}_a + \hat{M}_b\hat{N}_b - 1)$$

and

$$\psi_1 = \frac{2\pi h'}{P}, \quad \psi_2 = \frac{2\pi l'}{Q}.$$

As a result, the upper bound on the roundoff error obtained in this correspondence for 2-D block implemented filter is reduced approximately by a factor of $(KL)^{1/2}$ when compared with the result given by Aggarwal [7] for the filter implemented by an ordinary 2-D difference equation.

In the implementation, the choice of blocks with minimum dimensions becomes much more attractive, due to the resultant efficient filtering operation. Additionally, this would also result in an optimum roundoff error characteristic, as shown in (23b). As a consequence, the block implementation technique provides a very efficient and accurate means of recursive filtering operation, even when compared with the direct method using the 2-D difference equation.

V. Conclusion

The bound on the norm and the mean-square value of the error produced due to roundoff of multiplications for a 2-D block implemented digital filter is obtained employing fixed-point arithmetic. This bound is shown to be smaller than that available when the filter is implemented using ordinary 2-D difference equations.

Several 2-D block structures may be determined which exhibit different performances with regard to roundoff error. The exact form of these structures requires further investigation.

References

[1] S. K. Mitra and R. Gnanasekaran, "Block implementation of recursive digital filters—New structures and properties," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 200–207, Apr. 1978.

[2] C. W. Barnes and S. Shinnaka, "Block-shift invariance and block implementation of discrete-time filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 667–672, Aug. 1980.

[3] —, "Finite word effects in block-state realizations of fixed-point digital filters," *IEEE Trans. Circuits Syst.*, vol. CAS-27, pp. 345–349, May 1980.

[4] S. Y. Hwang, "Roundoff noise in state-space digital filtering: A general analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 256–262, June 1976.

[5] M. R. Azimi-Sadjadi, "Block implementation of two-dimensional digital filters," Ph.D. dissertation, Imperial College Sci. Technol., Univ. London, London, England, Oct. 1981.

[6] —, "Block implementation of two-dimensional digital filters," in *Proc. MEDCOMP '82*, Philadelphia, PA, Sept. 1982, pp. 160–169.

[7] M. D. Ni and J. K. Aggarwal, "Two-dimensional digital filtering and its error analysis," *IEEE Trans. Comput.*, vol. C-23, pp. 942–954, Sept. 1974.

[8] K. P. Shipp, J. R. Aggarwal, and J. K. Aggarwal, "An error analysis for a vector model of two-dimensional recursive filters," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 339–341, Aug. 1976.

[9] M. R. Azimi-Sadjadi, "A 2-D block-state realization model," in *Proc. IEEE Int. Symp. Circuits, Syst.*, Newport Beach, CA, May 1983, pp. 919–922.

[10] P. J. Davis, *Circulant Matrices.* New York: Wiley, 1979.

Comments on "A Recursive Kalman Window Approach to Image Restoration"

J. BIEMOND AND R.H.J.M. PLOMPEN

Abstract—**In a recent paper,[1] a recursive Kalman window estimation procedure for image restoration was claimed to be at least nearly optimal. Here, we show that it is not and point out some basic model errors.**

I. Introduction

The Kalman window approach to image restoration, introduced by Dikshit[1] to reduce processing time and storage requirements by processing images in overlapping strips and moving a processing window within these strips, contains some basic model errors, which severely affect the optimality of the resulting Kalman window algorithm.

Based on the assumption of a two-dimensional (2-D), separable, exponentially decaying autocovariance function of the original undistorted image, Dikshit introduces white noise driven, semicausal image models. These models are transformations of noncausal 4-point and 8-point nearest-neighbor models.

By observing that the models postulated by Dikshit are representations of a general linear 2-D autoregressive type of model for homogeneous images, in Section II we calculate the model coefficients in a linear MSE fitting procedure. We will show that the proposed transformation does not yield semicausal models with MSE coefficients (minimum-variance models), and further, that even in the case of minimum-variance model representations, semicausal and noncausal minimum-variance models are not driven by white noise.

In Section III, we will show that the erroneous assumption of a white noise model input results in an inadequate description of the dynamic model representing the original image and, ultimately, in a nonoptimal Kalman filter solution.

II. Image Modeling and Model Input

In the paper[1], it is assumed that the original undistorted image can be represented by a zero-mean homogeneous $m \times n$ random field, and that the 2-D ensemble autocorrelation function is separable and exponentially decaying, i.e.,

$$E[x(i,j)\,x(i-k,j-l)] = r(k,l) = \sigma^2 \exp(-\gamma_1|k| - \gamma_2|l|), \tag{1}$$

where k and l are the vertical and horizontal displacements, respectively. For convenience, we set $\sigma^2 = 1$.

The linear models postulated in the paper[1] to describe this

Manuscript received November 4, 1982; revised March 23, 1983.

The authors are with the Information Theory Group of the Electrical Engineering Department, Delft University of Technology, Delft, The Netherlands.

[1] S. S. Dikshit, *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-30, pp. 125–140, Apr. 1982.

0096-3518/83/1200-1573$01.00 © 1983 IEEE