



ELSEVIER

Signal Processing 55 (1996) 305–311

**SIGNAL
PROCESSING**

Finite-wordlength design of 2-D FIR digital filters for sampling structure conversion

Jong-Jy Shyu^{a,*}, Soo-Chang Pei^b, Yuan-Chih Lin^a

^a*Department of Computer Science and Engineering, Tatung Institute of Technology, 40 Chungshan N. Rd.,
3rd Sec., Taipei, Taiwan, Republic of China*

^b*Department of Electrical Engineering, National Taiwan University, Taipei, Taiwan, Republic of China*

Received 19 May 1995; revised 23 July 1996

Abstract

In this paper, an effective approach is proposed for designing discrete coefficient 2-D FIR digital filters for sampling structure conversion. After obtaining the initial continuous solution, the conventional Lagrange multiplier approach associated with an appropriate tree search algorithm is used iteratively to optimize the remaining unquantized coefficients of the designed filter in the least-squares sense when one or more of the coefficients take on discrete values, till all of the filter coefficients are quantized. The method is simple and the performance is comparable with that of the existing methods.

Zusammenfassung

In diesem Beitrag wird ein wirksamer Ansatz zum Entwurf zweidimensionaler FIR-Filter mit diskreten Koeffizienten zur Abtastraster-Umsetzung vorgeschlagen. Nach dem Erhalt einer kontinuierlichen Anfangslösung wird der übliche Ansatz mit einem Lagrange-Multiplikator zusammen mit einer geeigneten Baumsuche iterativ angewandt, um die verbliebenen unquantisierten Koeffizienten des entworfenen Filters im Kleinste-Quadrate-Sinn zu optimieren; dabei nehmen einzelne Koeffizienten diskrete Werte an, bis schließlich alle Filterkoeffizienten quantisiert sind. Das Verfahren ist einfach, und es leistet gleich viel wie die bekannten Methoden.

Résumé

Dans cet article, on propose une approche efficace de conception de filtres FIR 2-D à coefficients discrets pour la conversion de la structure d'échantillonnage. Après avoir obtenu la solution continue initiale, l'approche par multiplicateur de Lagrange conventionnel associé à un algorithme de recherche par arbre est utilisée itérativement afin d'optimiser au sens des moindres carrés les coefficients non quantifiés du filtre lorsqu'un ou plus des coefficients prend des valeurs discrètes, et ce jusqu'à ce que tous les coefficients du filtre soient quantifiés.

Keywords: Sampling structure conversion; Lagrange multiplier approach; Tree search algorithm

* Corresponding author. Fax: +886-2-5925252, ext. 2288; e-mail: jshyu@cse.tit.edu.tw.

1. Introduction

In digital signal processing, the conversion between different periodic sampling structures is an important problem, especially for the conversion between quincunx and rectangular structures in television image processing and HDTV applications [2, 7, 9, 10]. The sampling matrices of the rectangular and quincunx sampling structures are generally given by

$$S_R = \begin{bmatrix} T_1 & 0 \\ 0 & T_2 \end{bmatrix} \quad (1)$$

and

$$S_{L,K} = \begin{bmatrix} LT_1 & LT_1 \\ KT_2 & -KT_2 \end{bmatrix}, \quad (2)$$

respectively, where T_1 and T_2 represent the horizontal and vertical sampling periods, and L and K are quincunx sampling parameters.

The conversion from the rectangular sampling structure to the (L, K) quincunx sampling structure, and the inverse operation, correspond, respectively, to decimation and interpolation process. To avoid aliasing error and to remove redundant images, appropriate decimation filters and interpolation filters must be used, respectively. Moreover, because human eye is particularly sensitive to the distortion on the flat areas of the reconstructed pictures in the interpolation processing, certain constraints for designing interpolation filters should be considered.

The conversion system's operation requires numerous multiplications and additions. Multiplication, in particular, is extremely time consuming. So if a multiplication operation could be replaced by only a few shift operations, then the complexity of the entire conversion system could be reduced quite dramatically, such that a fast real-time system becomes feasible.

In this paper, an effective method is proposed for designing 2-D discrete coefficient FIR digital filters for sampling structure conversion systems. The method associates the Lagrange multiplier approach and a tree search algorithm, hence the constraints stated above can be incorporated into the design procedures. For each branch of the tree, the Lagrange multiplier approach is used to optimize the remaining unquantized coefficients of the designed interpolation and decimation filters in the least-squares sense when one or more

of the coefficients takes on discrete values. The proposed approach can be applied for different discrete coefficient spaces including the evenly distributed finite wordlength space and the nonuniformly distributed powers-of-two space. In this paper, the former is used only for comparison. Comparing with the existing methods, the method is simple and the performance is comparable.

2. Review of the applications of the Lagrange multiplier approach for designing continuous coefficient 2-D FIR filters for sampling structure conversion

Suppose the frequency response of the designed filter with quadrantal symmetric coefficients $h(n_1, n_2)$ is given by

$$\begin{aligned} H(\omega_1, \omega_2) &= \sum_{n_1=-N_1}^{N_1} \sum_{n_2=-N_2}^{N_2} h(n_1, n_2) e^{-jn_1\omega_1} e^{-jn_2\omega_2} \\ &= \sum_{n_1=0}^{N_1} \sum_{n_2=0}^{N_2} a(n_1, n_2) \cos(n_1\omega_1) \cos(n_2\omega_2), \end{aligned} \quad (3)$$

where $a(n_1, n_2)$ is related to $h(n_1, n_2)$ by

$$a(n_1, n_2) = \begin{cases} h(n_1, n_2) & \text{if } n_1 = 0 \text{ and } n_2 = 0, \\ 2h(n_1, n_2) & \text{if } (n_1 = 0 \text{ and } n_2 \neq 0) \\ & \text{or } (n_1 \neq 0 \text{ and } n_2 = 0), \\ 4h(n_1, n_2) & \text{otherwise.} \end{cases} \quad (4)$$

Eq. (3) can be represented in vector form by

$$H(\omega_1, \omega_2) = \mathbf{A}^T \mathbf{C}(\omega_1, \omega_2) = \mathbf{C}^T(\omega_1, \omega_2) \mathbf{A}, \quad (5)$$

where

$$\mathbf{A} = [\mathbf{A}_0^T \ \mathbf{A}_1^T \ \dots \ \mathbf{A}_{M_1}^T]^T \quad (6)$$

and

$$\begin{aligned} \mathbf{C}(\omega_1, \omega_2) &= [\mathbf{C}_0^T(\omega_1, \omega_2) \ \mathbf{C}_1^T(\omega_1, \omega_2) \ \dots \ \mathbf{C}_{N_1}^T(\omega_1, \omega_2)]^T, \end{aligned} \quad (7)$$

in which

$$A_i = [a(i, 0) \ a(i, 1) \ \dots \ a(i, N_2)]^T, \quad 0 \leq i \leq N_1, \quad (8)$$

and

$$C_i(\omega_1, \omega_2) = [\cos(i\omega_1) \ \cos(i\omega_1) \cos(\omega_2) \ \dots \ \cos(i\omega_1) \cos(N_2\omega_2)]^T, \quad 0 \leq i \leq N_1. \quad (9)$$

Using these notations, the integrated square error is

$$e = \iint_R W(\omega_1, \omega_2) |D(\omega_1, \omega_2) - H(\omega_1, \omega_2)|^2 d\omega_1 d\omega_2 = s + P^T A + A^T Q A, \quad (10)$$

where R represents the designed bands, $W(\omega_1, \omega_2)$ is the weighting function,

$$s = \iint_R W(\omega_1, \omega_2) D^2(\omega_1, \omega_2) d\omega_1 d\omega_2, \quad (11)$$

$$P = \iint_R -2W(\omega_1, \omega_2) D(\omega_1, \omega_2) \times C(\omega_1, \omega_2) d\omega_1 d\omega_2 \quad (12)$$

and

$$Q = \iint_R W(\omega_1, \omega_2) C(\omega_1, \omega_2) \times C^T(\omega_1, \omega_2) d\omega_1 d\omega_2. \quad (13)$$

Because human eye is particularly sensitive to the distortion on the flat areas of the reconstructed pictures in the interpolation processing, the following frequency constraints in the first quarter plane should be considered [8, 10]:

$$H(0, 0) = G, \quad (14)$$

$$H(\omega_{1l_1}, \omega_{2k_2}) = H(\hat{\omega}_{1\hat{l}_2}, \hat{\omega}_{2\hat{k}_2}) = 0,$$

where G is equal to 1 for the decimation filter design and equal to $2LK$ for interpolation processing,

$$(\omega_{1l_1}, \omega_{2k_2}) = \left(\frac{(2l_1 + 1)\pi}{L}, \frac{(2k_2 + 1)\pi}{K} \right),$$

$$0 \leq l_1 \leq \left\lfloor \frac{L-1}{2} \right\rfloor = \bar{L}, \quad 0 \leq k_2 \leq \left\lfloor \frac{K-1}{2} \right\rfloor = \bar{K}, \quad (15)$$

and

$$(\hat{\omega}_{1\hat{l}_2}, \hat{\omega}_{2\hat{k}_2}) = \left(\frac{2l_2\pi}{L}, \frac{2k_2\pi}{K} \right),$$

$$0 \leq l_2 \leq \left\lfloor \frac{L}{2} \right\rfloor = \hat{L}, \quad 0 \leq k_2 \leq \left\lfloor \frac{K}{2} \right\rfloor = \hat{K}$$

but $(l_2, k_2) \neq (0, 0)$, (16)

in which $\lfloor x \rfloor$ denotes the largest integer less than x . Eq. (14) can be represented in matrix form by

$$B^T A = G, \quad (17)$$

where

$$B = [C(0, 0) \ C(\omega_{10}, \omega_{20}) \ \dots \ C(\omega_{1\hat{L}}, \omega_{2\hat{K}}) \ C(\hat{\omega}_{10}, \hat{\omega}_{21}) \ \dots \ C(\hat{\omega}_{1\hat{L}}, \hat{\omega}_{2\hat{K}})] \quad (18)$$

and

$$G = [G \ 0 \ \dots \ 0 \ 0 \ \dots \ 0]^T. \quad (19)$$

Hence, the design of two-dimensional filters for sampling structure conversion can be formulated as a quadratic programming problem

$$\begin{aligned} \text{Minimize} \quad & e = s + P^T A + A^T Q A \\ \text{subject to} \quad & B^T A = G, \end{aligned} \quad (20)$$

which results in the closed-form solution [5, 8]

$$A = Q^{-1} B (B^T Q^{-1} B)^{-1} G + \frac{1}{2} Q^{-1} [B (B^T Q^{-1} B)^{-1} B^T Q^{-1} - I] P, \quad (21)$$

where I is an $(N_1 + 1)(N_2 + 1) \times (N_1 + 1)(N_2 + 1)$ identity matrix.

3. Design of finite-wordlength 2-D FIR filters for sampling structure conversion

Once the continuous coefficient filter is obtained, the key operation in the discrete optimization algorithm is to optimize the unquantized coefficients when some of the coefficients take on discrete values. Notice that the constraints of some coefficients to discrete values can be represented in a constrained

matrix form. For example, the coefficients $a(0,2)$, $a(0,5)$ and $a(0,3)$ should be constrained to discrete values $a_d(0,2)$, $a_d(0,5)$ and $a_d(0,3)$, respectively, which can be represented by the constrained equation

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots \end{bmatrix} \mathbf{A} = \begin{bmatrix} a_d(0,2) \\ a_d(0,5) \\ a_d(0,3) \end{bmatrix}. \quad (22)$$

Hence, the Lagrange multiplier approach can be applied iteratively as an effective discrete optimization algorithm.

As to the branch and bound methods [1], there are two main tree search strategies, i.e. the depth-first-search strategy and breadth-first-search strategy. The former has the advantage of producing an early improvement in the initial solution, and the latter usually leads to a better first solution but is ineffective which makes it rarely be used alone. In this paper, we use the hybrid tree search of the two [3]. The details of the tree search algorithm is described as follows. After obtaining the continuous coefficient filter by (21), the coefficient with maximum absolute value is selected to take on discrete values and then fixed at I discrete values in the vicinity of the continuous optimum value. Each of the discrete values is fixed and the Lagrange multiplier approach is used to find the continuous solution under the constrained equation like (22). Hence, there are I optimization problems if the chosen coefficient is fixed at I different discrete values, which result in I sets of continuous solutions under fixing the first quantized coefficient. For each of the I sets of continuous solutions, the coefficient with maximum absolute value is chosen to take on discrete values from the unquantized coefficients. So there are I^2 optimization problems when two coefficients are quantized, which also result in I^2 sets of continuous solutions under fixing two quantized coefficients. For each of the obtained solutions, certain measure error is computed such that the constraints in (14) can be incorporated into the optimization. Then only I of the I^2 problems, which provide the smallest error, are selected for further quantization of the coefficients (so that the number of problems should be processed would not increase step by step) and produce I^2 further optimization problems when a third coefficient is selected to take on discrete values. The

processes are continued until all the coefficients take on discrete values.

The procedures for designing finite-precision coefficient filters for sampling structure conversion by the proposed iterative Lagrange multiplier approach is summarized as follows:

Step 1. Find the initial optimal continuous solution by (21) and the given specifications.

Step 2. Select the coefficient with maximum absolute value, for example $a(r_1, r_2)$, and fix it at I discrete values in the vicinity of $a(r_1, r_2)$, say $a_d^i(r_1, r_2)$, $i = 1, \dots, I$.

Step 3. Establish the constrained equation for each of the I optimization problems under fixing the first quantized coefficient

$$\mathbf{B}_i^T \mathbf{A} = \mathbf{G}_i, \quad i = 1, \dots, I, \quad (23)$$

where \mathbf{B}_i , $i = 1, \dots, I$, are column vectors with zero elements except the $(r_1 + r_2(N_2 + 1) + 1)$ th element be unit and \mathbf{G}_i , $i = 1, \dots, I$, are one-element matrices with element $a_d^i(r_1, r_2)$, $i = 1, \dots, I$, respectively.

Step 4. Find the I sets of continuous solutions by

$$\mathbf{A} = \mathbf{Q}^{-1} \mathbf{B}_i (\mathbf{B}_i^T \mathbf{Q}^{-1} \mathbf{B}_i)^{-1} \mathbf{G}_i + \frac{1}{2} \mathbf{Q}^{-1} [\mathbf{B}_i (\mathbf{B}_i^T \mathbf{Q}^{-1} \mathbf{B}_i)^{-1} \mathbf{B}_i^T \mathbf{Q}^{-1} - \mathbf{I}] \mathbf{P}. \quad (24)$$

Step 5. Select the coefficient with maximum absolute value from unquantized coefficients for each of the I sets of continuous solutions and fix it at I different discrete values, for example, $a_d^{ij}(k_1, k_2)$, $i = 1, \dots, I$, $j = 1, \dots, I$, where i denotes the order of I discrete values in the previous quantized step and j denotes that in the latest quantized step.

Step 6. Establish the constrained equation for each of the I^2 further optimization problems under fixing the quantized coefficients by

$$\mathbf{B}_{ij}^T \mathbf{A} = \mathbf{G}_{ij}, \quad i = 1, \dots, I, \quad j = 1, \dots, I, \quad (25)$$

where

$$\mathbf{B}_{ij} = [\mathbf{B}_i \quad \mathbf{Z}_j], \quad i = 1, \dots, I, \quad j = 1, \dots, I, \quad (26)$$

and

$$\mathbf{G}_{ij} = \begin{bmatrix} \mathbf{G}_i \\ \mathbf{Y}_j \end{bmatrix}, \quad i = 1, \dots, I, \quad j = 1, \dots, I, \quad (27)$$

in which \mathbf{Z}_j , $j = 0, \dots, I$, are column vectors with zero elements except the element be unit whose position

agrees with that of the quantized coefficient in Step 5, and $Y_j, j = 0, \dots, I$, are one-element matrices with element $a_d^{ij}(k_1, k_2), j = 0, \dots, I$, for the corresponding i . Step 7. Find the I^2 sets of continuous solutions by (24) and calculate the error value

$$e_A = s + \mathbf{P}^T \mathbf{A} + \mathbf{A}^T \mathbf{Q} \mathbf{A} + W \cdot S(|\mathbf{B}^T \mathbf{A} - \mathbf{G}|) \quad (28)$$

for the I^2 optimization problems where \mathbf{B} and \mathbf{G} have been originally defined in (17)–(19), W is the weighting constant and $S(\mathbf{M})$ denotes the summation of the elements in the matrix \mathbf{M} . Then select I sets provide smallest value of e_A for further optimization. Notice that the error in (28) is so defined that the constraints in (14) can be incorporated into the discrete optimization.

Step 8. If all the coefficients are quantized, go to the next step; otherwise set

$$\mathbf{B}_l = \mathbf{B}_{ij}, \quad l = 1, \dots, I, \quad (29)$$

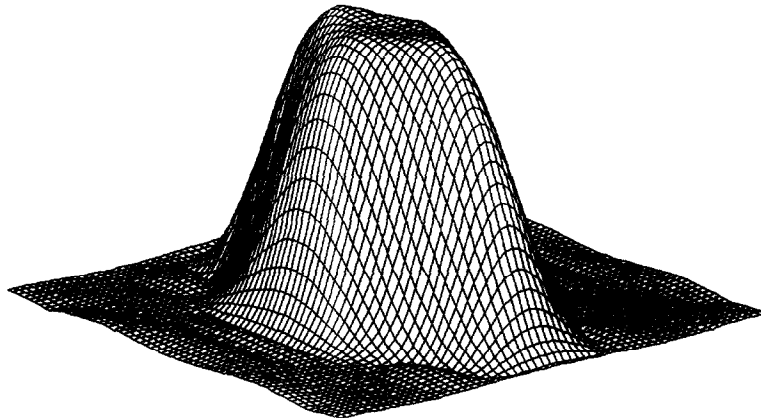
and

$$\mathbf{G}_l = \mathbf{G}_{ij}, \quad l = 1, \dots, I, \quad (30)$$

where the I sets of \mathbf{B}_{ij} and \mathbf{G}_{ij} are chosen in Step 7, then go to Step 5.

Step 9. Select the set provides the smallest error value e_A from the I sets of discrete solutions obtained in Step 7 as the desired solution.

Example. A $5 \times 9, (L, K) = (1, 2)$ interpolation filter is designed in this example, and the desired response



(a)

n_1	0	1	2
n_2			
0	18	25	0
1	38	18	-1
2	19	6	-4
3	4	-2	-4
4	0	-2	-2

(b)

Fig. 1. (a) The amplitude response of a $5 \times 9, (L, K) = (1, 2)$ interpolation filter. (b) The filter coefficients in the first quarter plane ($h(n_1, n_2)$ multiplied by 2^6).

is given by

$$D(\omega_1, \omega_2) = \begin{cases} 4, & \omega_1 + 2\omega_2 \leq 2\pi(0.2), \\ 0, & \omega_1 + 2\omega_2 \geq 2\pi(0.7). \end{cases} \quad (31)$$

The coefficients are coded with 7 bits (sign bit included). When $I = 4$ and $W = 1000$ are used, the resultant amplitude response is shown in Fig. 1(a) and the filter coefficients in the first quarter plane are listed in Fig. 1(b). The related results are tabulated in Table 1 accompanying with those of the algorithm in [9] (integer linear programming) and direct quantization. Notice that although the approach of direct quantization gives smaller integrated square error, but it results in significant constrained error, where the integrated square error is defined in (10) and the constrained error denotes the sum of square errors between constrained values and actual values over constrained points. Also, we present the results that the coefficient with minimum absolute value to be quantized first, but the performance is not so good. Moreover, through several design examples, we recommend that the quantization is processed with the

maximum absolute valued coefficient to be chosen first.

4. Discussions and conclusions

An effective method has been proposed for designing discrete coefficient FIR filters for sampling structure conversion. The method associates successfully the Lagrange multiplier approach and a tree search algorithm, which makes it attractive for designing finite-wordlength digital filters. A measure error is proposed, such that the frequency-domain constraints for sampling structure conversion system can be incorporated into the design procedures. Comparing with the existing methods, the proposed method is simple and the performance is comparable.

From Table 1 and through several examples, we generally choose the coefficient with maximum absolute value to quantize first. Comparing with other algorithms for the choice of the coefficients, the integrated square error is 0.0098 in the proposed method

Table 1

The obtained results of the designed 5×9 , $(L, K) = (1, 2)$ interpolation filter for different approaches

Method	Initial solution	Proposed approach I ^a	Proposed approach II ^b	Direct quantization	Algorithm in [1] (integer programming)
Integrated square error	0.0036	0.0098	0.0117	0.0084	0.0166
Passband peak error	0.1558	0.1599	0.1911	0.1562	0.1167
Stopband peak error	0.1411	0.1420	0.1829	0.1577	0.1214
Amplitude response in constrained points					
$H(0, 0)$	4	4	4	3.8906	4
$H(0, \pi)$	0	0	0	0.0156	0
$H(\pi, \frac{\pi}{2})$	0	0	0	0.0156	0
Design time on 486 PC (s)	6	20	20	7	

^a The proposed approach I chooses the maximum absolute-valued coefficient to quantize first.

^b The proposed approach II chooses the minimum absolute-valued coefficient to quantize first.

which is smaller than 0.0166 if the most sensitive coefficient is quantized first.

Generally, as the statements declared in Section 1, the proposed approach can be applied for different discrete coefficient spaces including the evenly distributed finite wordlength space and the nonuniformly distributed powers-of-two space. If the former is used, the coefficients can be implemented using both of floating or fixed point arithmetic, and the multiplication operation can be replaced by a few shift operation if the latter approach is applied.

For the choice of the number of bits, it is dependent on the design requirements. If the integrated square error cannot be satisfied, the number of bits of the quantized coefficient or the filter length should be increased.

As to the choice of I , the number of discrete values in the vicinity of the continuous optimum value to be quantized, the same results are obtained in the presented example when $I = 2, 4$ and 8 . By the experience of several designed examples, the performance of $I = 2$ is good enough, but we usually take $I = 4$.

References

- [1] R.S. Garfinkel and G.L. Nemhauser, *Integer Programming*, Wiley, New York, 1972.
- [2] A. Knoll, "Filter design for the interpolation of highly subsampled pictures", *Signal Processing: Image Communication*, Vol. 3, June 1991, pp. 239–248.
- [3] Y.C. Lim and S.R. Parker, "Discrete coefficient FIR digital filter design based upon an LMS criteria", *IEEE Trans. Circuits Systems*, Vol. 30, October 1983, pp. 723–739.
- [4] Y.C. Lim and S.R. Parker, "FIR filter design over a discrete powers-of-two coefficient space", *IEEE Trans. Acoust. Speech Signal Process.*, Vol. 31, June 1983, pp. 583–591.
- [5] G.W. Medlin, J.W. Adams and C.T. Leondes, "Lagrange multiplier approach to the design of FIR filters for multirate applications", *IEEE Trans. Circuits Systems*, Vol. 35, October 1988, pp. 1210–1219.
- [6] S.C. Pei and J.J. Shyu, "Design of two-dimensional FIR eigenfilters for sampling structure conversion", *IEEE Trans. Circuits Systems Video Tech.*, Vol. 3, April 1993, pp. 158–162.
- [7] M. Renfors, T. Huuhtanen, A. Nieminen and T. Koivunen, "Linear and nonlinear filters for sampling structure conversion of two-dimensional sequences", in: L. Chiariglione, ed., *Signal Processing of HDTV, II*, Elsevier, Amsterdam, 1990.
- [8] J.J. Shyu and S.C. Pei, "Lagrange multiplier approach to the design of two-dimensional FIR digital filters for sampling structure conversion", *IEEE Trans. Signal Process.*, Vol. 42, October 1994, pp. 2884–2886.
- [9] P. Siohan, "2-D FIR filter design for sampling structure conversion", *IEEE Trans. Circuits Systems Video Tech.*, Vol. 1, December 1991, pp. 337–350.
- [10] P. Siohan and D. Pelè, "Sampling pattern conversion with 2-D FIR filters", *Proc. 3rd Internat. Conf. on Image Proc. and its Applications*, IEE Conf. Publication No. 307, July 1989, pp. 275–279.
- [11] P. Siohan and A. Benslimane, "Finite precision design of optimal linear phase 2-D FIR digital filters", *IEEE Trans. Circuits Systems*, Vol. 36, January 1989, pp. 11–22.