TWO-DIMENSIONAL FIR FILTER DESIGN USING NONLINEAR PROGRAMMING

Jeng-Jong Pan*
Senior Applications Scientist
TGS Technology, Inc.
EROS Data Center
Sioux Falls, SD 57198
U. S. A.
WG III/4

ABSTRACT

Two nonlinear programming techniques, unconstrained and constrained, can be used for designing two-dimensional finite-duration impulse response (FIR) digital filters. The objective function in nonlinear programming is the summation of the squares of difference between desired frequency and computed frequency and the squares of a FIR filter having an exponential-type weighting function. The frequency response of a FIR filter, thus, can be smoothed by forcing the magnitude of the impulse response to decrease gradually from the center of the filter to the edges. The use of constrained nonlinear programming provides additional constraints to control filter coefficients in the spatial domain. A two-dimensional differentiator design using these techniques is presented. These techniques, however, can be applied to a wide range of filter design problems.

INTRODUCTION

In digital image processing, two-dimensional (2D) finite-duration impulse (FIR) filters (also called nonrecursive filters) are often used to improve the utility of a given image for a specific application. Techniques used to design a filter for one application may not be suitable for another. It is desirable, therefore, to develop filter design techniques that are flexible enough to encompass a large number of applications without sacrificing performance for any individual filtering problem. A filter design tool that can be easily modified to accommodate a wide range of applications, such as smoothing or edge sharpening, would be of great value to a filter designer.

For a given FIR filter, there is a unique frequency response that is the discrete Fourier transform (DFT) of the filter. The FIR filter can be obtained by performing the inverse Fourier transform on desired frequency response and truncating the result. This truncation, however, distorts the desired frequency response.

It is not possible to construct a linear, discrete FIR filter with a DFT that exactly matches all desired frequencies. Thus,

---

Publication authorized by the Director, U.S. Geological Survey.
*Work performed under U.S. Geological Survey contract 14-08-001-22521.

techniques used in digital filter design attempt to compromise the trade-off between matching desired frequency responses with computed frequency response, and controlling the effects of spatial truncation.

In this paper, nonlinear programming techniques are used to solve the above problem in FIR filter design. Here only odd-length symmetric (zero-phase) filters are discussed. Nonlinear programming is a mathematical technique used to minimize a nonlinear function called the "objective function". The objective function defined here is a combination of a mean-square error term of frequency responses and an exponential-type term of impulse responses. The mean-square error term, a constraint applied in the frequency domain, is used to control the difference between desired and computed frequency responses of the filter. The exponential-type term, a constraint applied in the spatial domain, is used to decrease the magnitude of the FIR filter coefficient gradually from the center of the filter to the edges.

TWO-DIMENSIONAL FIR FILTER DESIGN

In a two-dimensional linear, discrete system, the frequency response of an odd-length FIR filter is written as

$$H(u,v) = \sum_{m=-N}^{N} \sum_{n=-N}^{N} h(m,n)\, e^{-j(mu+nv)} \qquad (1)$$

where $j = \sqrt{-1}$, $2N+1$ is the filter size in each dimension, and $h(m,n)$ is the FIR filter (or the filter coefficient). As shown in equation (1), the frequency response can be written as a linear function of the FIR filter. The FIR filter design problem consists of determining $h(m,n)$, when N is predefined, which produces the desired frequency response $H(u,v)$.

If the impulse response is restricted to be real, then it follows that

$$H(u,v) = H^{*}(-u,-v) \qquad (2)$$

where the asterisk means complex conjugation. If we are only concerned with a filter having zero-phase, that means the frequency response is also assumed to be real, then

$$h(m,n) = h(-m,-n) \qquad (3)$$

and equation (1) can be rewritten as

$$H(u,v) = h(0,0) + 2 \sum_{m=1}^{N} h(m,0)\cos(mu) + 2 \sum_{n=1}^{N} \sum_{m=1}^{N} h(m,n)\cos(mu+nv) \qquad (4)$$

In order to construct a suitable FIR filter for the continuous frequency response of equation (4), the computed frequency

response $H(u,v)$ must have the best approximation to the desired frequency response $H_d(u,v)$ in the Nyquist range of the frequency domain. Let the error between desired frequency response and computed frequency response be given by

$$E(u,v) = H_d(u,v) - H(u,v) \tag{5}$$

The approach for determining a FIR filter is to minimize some functions of this error, such as the L_2-norm (mean-square value), e_2,

$$e_2 = [\iint E(u,v)^2 \, du \, dv]^{1/2} \tag{6}$$

or the Chebyshev norm, e_∞,

$$e_\infty = \max |E(u,v)| \tag{7}$$

(De Meyer, 1974). Mathematical optimization techniques such as linear programming have been employed extensively to minimize the Chebyshev norm (Herrmann, 1970; Rabiner, 1972; Rabiner and others, 1975; Fiasconaro, 1979). The magnitudes of the FIR filter obtained from linear programming, however, are unable to be decreased gradually from the center of filter to the edges (Lewin and Telljohann, 1984).

The least-squares method can be applied to determine $h(m,n)$, by minimizing equation (6). The filter design obtained by this way, is equivalent to the filter design using a rectangular windowing method (Oppenheim and Schafer, 1975), which leads to adverse behavior if discontinuities exist in $H_d(u,v)$. Instead of using a rectangular windowing method to truncate a fiter directly, an exponential-type weighting function can be used to force the magnitudes of the filter coefficients to decrease gradually to zero from the center of the filter to the edges. The filter coefficients can be determined by minimizing a nonlinear function, $F(X)$, as

$$\begin{aligned} F(X) &= F(h(m,n)) \\ &= r\,P(h(m,n)) + q\,Q(h(m,n)) \end{aligned} \tag{8}$$

where

X is a vector with parameters $h(m,n)$, $m = 0,1,\ldots,N$, $n = 0,1,\ldots,N$

$$P(h(m,n)) = \iint E(u,v)^2 \, du \, dv,$$

$$Q(h(m,n)) = -h(0,0)^2 + \sum_{n=0}^{N} \sum_{\substack{m=n \\ m \neq 0}}^{N} h(m,n)^2 \, e^{z\sqrt{m^2+n^2}},$$

z, r and q are positive constants.

The first term, $P(h(m,n))$, in equation (8) is a least-squares term that forces the solution into satisfying the desired frequency response by treating the equation as a penalty constraint weighted by a factor r. The second term, $Q(h(m,n))$, incorporates the spatial constraints weighted by factor q. The larger the ratio r/q, the greater the emphasis on exerting the

accuracy of frequency response, relative to the magnitude variation of the FIR filter. The value of $r/q=1000$ is recommended as about optimum for a wide range of filter design. If q is set to zero, the equation (8) can be solved by least-squares method.

The exponential-type function in the second term weights filter coefficients more when the greater their distance from the central coefficient. The central coefficient, $h(0,0)$, is not weighted. The weighting factor z in the exponential-type function is selected by the filter designer and can be adjusted to force the magnitudes of filter coefficients at the edges to nearly zero.

If some further constraints on the FIR filter are desired, then each constraint can be formed into as

$$L \leq h(m,n) \leq U \qquad (9)$$

where L is the lower bound and U is the upper bound of the filter coefficient. L and U can be either constants or functions of the other filter coefficients.

NONLINEAR PROGRAMMING

The mathematical technique used to minimize the nonlinear function of equation (8) is nonlinear programming. The purpose of nonlinear programming is to minimize a nonlinear objective function $F(X)$, where $X, X=(x_1, x_2, \ldots, x_k)$, is a point in the k-dimensional parameter space. The space is defined by k mutually orthogonal axes, each represented by parameter x_i, $i=1, 2, \ldots, k$. Nonlinear programming is either unconstrained or constrained, depending on the ability to manipulate the restrictions of the parameters x_i in the optimization process.

Unconstrained nonlinear programming, where no constraints are placed on parameters, can be classified as either direct search methods or gradient methods. Direct search methods do not require explicit evaluations of partial derivatives of the objective function, but instead only require computing the objective function values, plus information obtained from earlier iterations. In some direct search methods (for example, Powell's method, 1964), objective function values are used to obtain numerical approximations of their derivatives. These direct methods are most suitable for investigating simple models that have relatively few parameters.

Gradient methods are those in which the search direction for finding a minimum point in each iteration is selected on the basis of the gradient of the objective function. This involves computing the partial derivative of the function with respect to each parameter. The negative of the gradient vector can be utilized to find the minimum point rapidly. The gradient of an objective function can be computed either analytically or numerically.

Constrained nonlinear programming techniques fall into two

broad categories: feasibility checking and modified objective function approaches. A point X that satisfies all known constraints is called a feasible point. Feasibility checking methods are similar to unconstrained methods, except that a checking procedure is made to see if a constraint is violated. If this occurs, the current point is relocated inside the feasible region in a prescribed manner. The modified objective function technique incorporates the constraints into the objective function to produce an unconstrained problem. Penalty functions are applied to the objective function at nonfeasible points, forcing the search back into the feasible region.

Details of the mathematical techniques used in the nonlinear programming can be found in Luenberger (1973). In this paper, Powell's method, a direct search method in unconstrained problems, and Rosenbrock's constrained method, a feasibility checking approach in constrained problems, are used to study the FIR filter design. The optimization algorithms of Powell's and Rosenbrock's constrained methods can be found in Powell (1964) and Rosenbrock and Storey (1966).

APPLICATIONS AND DISCUSSIONS

A 2D differentiator is used to illustrate the applications of nonlinear programming in FIR filter design. The frequency response of a differentiator is given, without normalization, by

$$F(u,v) = \sqrt{u^2 + v^2} \qquad (10)$$

The symmetry conditions $h(m,n) = h(|m|,|n|)$ and $h(m,n) = h(n,m)$ apparently exist. Therefore, equation (4) can be expressed as

$$H(u,v) = h(0,0) + 2 \sum_{m=1}^{N} h(m,0)(\cos(mu) + \cos(mv))$$

$$+ 4 \sum_{m=1}^{N} h(m,m) \cos(mu) \cos(mv)$$

$$+ 4 \sum_{n=1}^{N} \sum_{m=n+1}^{N} h(m,n) (\cos(mu) \cos(mv) + \cos(nu) \cos(nv)) \qquad (11)$$

Since all four quadrants of the filter are symmetrical, only the upper-right quadrant of the FIR filter and the corresponding frequency response will be shown.

The frequency response, however, is a continuous function of frequency in the frequency domain, there are an infinite number of points (u,v) over which the best approximation is desired. It is necessary to choose some finite subset of the infinite set when applying nonlinear programming to solve filter design problems. In order to numerically integrate $P(h(m,n))$ of equation (8), we need to change the integration to a summation-type as

$$\sum_{i=1}^{M}\sum_{j=1}^{M} E^2(u_i,v_j) \qquad (12)$$

where M, the data-size, is the number of data samples with equal space along each axis in the upper-right quadrant.

In table 1, the errors of six filters with different weighting parameters are compared. The mean-square error is obtained from equation (12) and the maximum error is the maximum difference between each desired frequency and computed frequency. The impulse response of each filter is shown in table 2 and the corresponding frequency response is shown in figure 1.

From table 2, it is found, in this example, that value z assumed between 2.5 and 3.0 can force the magnitudes of filter coefficients to nearly zero at the filter edge. The mean-square error and the maximum error, however, relatively increased when z increased.

In table 3, the mean-square errors of different data sizes with four different parameter z values are compared when N=5, r=1000, and q=1. Error increases with increasing M when N is fixed. This implies that it is not necessary to sample the desired frequency too densely.

In table 4, the mean-square errors of different filter sizes with four different z values are compared when M=11, r=1000, and q=1. It is evident that the mean-square error decreases with increasing filter size while the data size is fixed. It proves that the computed frequency response at a desired frequency response will be quantitatively better for a larger filter size. However, larger filter size entails an increase in the computation time for the numerical summation. The designer has to make a compromise between obtaining sufficient accuracy with a relatively small filter size and reducing the computation time.

For some applications in digital image processing, it may be required to reduce the ringing effect caused by alternating signs of the filter coefficients. In order to minimize the ringing problem, one may restrict all coefficients, except the central one, h(0,0), to have the same sign. Also, for getting an even smoother response, the magnitude of each filter coefficients may be constrained to be

$$\begin{aligned} &h(m,n) \le 0 && \text{for all } (m,n), \text{ except } (0,0), \\ &h(m',n') \le h(m,n) && \text{if } |m'| + 1 = |m| \text{ and } n' = n, \\ &\qquad\qquad \text{or} && \text{if } |n'| + 1 = |n| \text{ and } m' = m \end{aligned} \qquad (13)$$

Since the Rosenbrock's constrained method has to check, and relocate if necessary, the point inside the feasible region at each iteration, the computation time used in this method is therefore greater than those based on unconstrained optimization methods. Here, a smaller filter, filter 7 (N=3), is designed under the above constraints when M=4. The frequency response of filter 7 is shown in figure 1(h) and the

Table 1. Error analysis of 2D differentiator design using nonlinear programming

Name	r	q	z	mean-square error	maximum error
Filter 1	1000	0	0.0	0.016	0.050
Filter 2	1000	1	1.0	0.016	0.054
Filter 3	1000	1	1.5	0.016	0.054
Filter 4	1000	1	2.0	0.019	0.057
Filter 5	1000	1	2.5	0.056	0.082
Filter 6	1000	1	3.0	0.118	0.103

Table 3. Mean-square errors of different data sizes (M) when N=5, r=1000, and q=1.

z	M=7	M=8	M=9	M=10	M=11
0.0	0.016	0.047	0.071	0.091	0.109
1.0	0.016	0.047	0.071	0.091	0.109
2.0	0.019	0.049	0.072	0.093	0.110
3.0	0.118	0.150	0.176	0.200	0.221

Table 4. Mean-square errors of different filter sizes (N) when M=11, r=1000, and q=1.

z	N=5	N=6	N=7	N=8	N=9
0.0	0.109	0.076	0.039	0.025	0.007
1.0	0.109	0.076	0.039	0.025	0.007
2.0	0.110	0.083	0.046	0.032	0.015
3.0	0.221	0.136	0.089	0.060	0.044

Table 2. Impulse responses of seven different FIR filters. Filter number is refered to table 1. Filter 7 is derived using constrained nonlinear programming.

Filter 1

5	-0.02309	-0.00244	-0.00211	-0.00270	-0.00192	-0.00216
4	0.02298	-0.00144	-0.00205	-0.00233	-0.00300	-0.00192
3	-0.04573	-0.00948	-0.00407	-0.00162	-0.00233	-0.00270
2	0.05766	-0.00990	-0.00982	-0.00407	-0.00205	-0.00211
1	-0.44191	-0.08149	-0.00990	-0.00948	-0.00144	-0.00244
n=0	2.41943	-0.44191	0.05766	-0.04573	0.02298	-0.02309
m=	0	1	2	3	4	5

Filter 2

5	-0.02307	-0.00244	-0.00211	-0.00269	-0.00192	-0.00214
4	0.02296	-0.00145	-0.00205	-0.00233	-0.00300	-0.00192
3	-0.04573	-0.00947	-0.00407	-0.00163	-0.00233	-0.00269
2	0.05765	-0.00990	-0.00982	-0.00407	-0.00205	-0.00211
1	-0.44190	-0.08150	-0.00990	-0.00947	-0.00145	-0.00244
n=0	2.41953	-0.44190	0.05765	-0.04573	0.02296	-0.02307
m=	0	1	2	3	4	5

Filter 3

5	-0.02276	-0.00250	-0.00217	-0.00274	-0.00183	-0.00166
4	0.02291	-0.00149	-0.00206	-0.00235	-0.00291	-0.00183
3	-0.04573	-0.00945	-0.00410	-0.00167	-0.00235	-0.00274
2	0.05763	-0.00988	-0.00984	-0.00410	-0.00206	-0.00217
1	-0.44194	-0.08153	-0.00988	-0.00945	-0.00149	-0.00250
n=0	2.41943	-0.44194	0.05763	-0.04573	0.02291	-0.02276
m=	0	1	2	3	4	5

Filter 4

5	-0.01960	-0.00259	-0.00255	-0.00220	-0.00088	-0.00022
4	0.02218	-0.00192	-0.00217	-0.00252	-0.00184	-0.00088
3	-0.04620	-0.00957	-0.00432	-0.00202	-0.00252	-0.00220
2	0.05754	-0.00971	-0.00991	-0.00432	-0.00217	-0.00255
1	-0.44269	-0.08180	-0.00971	-0.00957	-0.00192	-0.00259
n=0	2.41898	-0.44269	0.05754	-0.04620	0.02218	-0.01960
m=	0	1	2	3	4	5

Filter 5

5	-0.00755	-0.00146	-0.00148	-0.00038	-0.00010	-0.00001
4	0.01856	-0.00337	-0.00204	-0.00176	-0.00020	-0.00010
3	-0.04805	-0.01030	-0.00599	-0.00247	-0.00176	-0.00038
2	0.05792	-0.00946	-0.01008	-0.00599	-0.00204	-0.00148
1	-0.44498	-0.08247	-0.00946	-0.01030	-0.00337	-0.00146
n=0	2.42009	-0.44498	0.05792	-0.04805	0.01856	-0.00755
m=	0	1	2	3	4	5

Filter 6

5	-0.00094	-0.00019	-0.00018	-0.00002	-0.00001	0.00000
4	0.00925	-0.00285	-0.00066	-0.00036	0.00000	-0.00001
3	-0.04772	-0.00997	-0.00676	-0.00125	-0.00036	-0.00002
2	0.05885	-0.01029	-0.00983	-0.00676	-0.00066	-0.00018
1	-0.44688	-0.08382	-0.01029	-0.00997	-0.00285	-0.00019
n=0	2.42438	-0.44688	0.05885	-0.04772	0.00925	-0.00094
m=	0	1	2	3	4	5

Filter 7

3	-0.00311	-0.00162	-0.00030	-0.00030
2	-0.05656	-0.05651	-0.00031	-0.00030
1	-0.34428	-0.11917	-0.05651	-0.00162
n=0	2.56946	-0.34428	-0.05656	-0.00311
m=	0	1	2	3

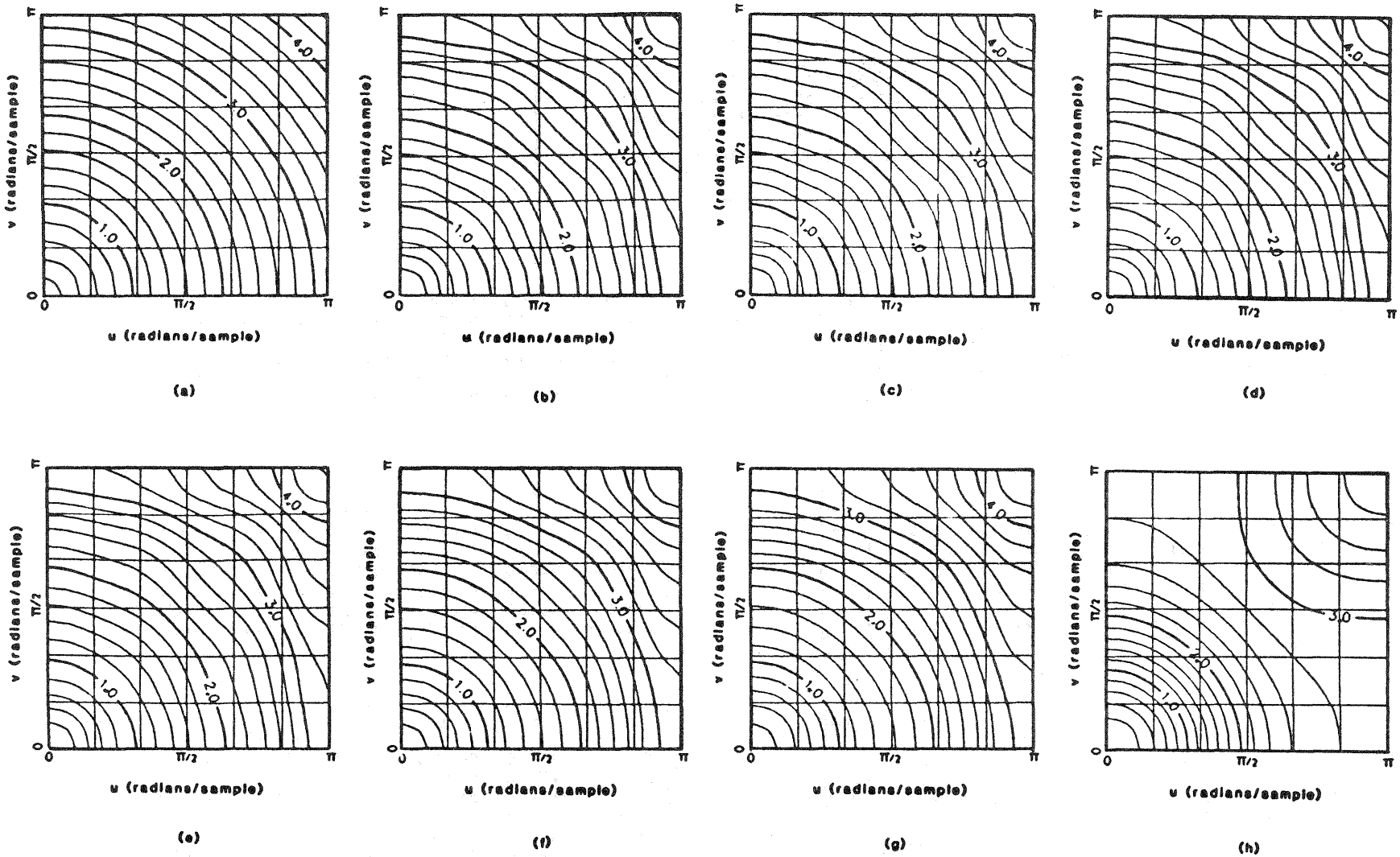

Figure 1. Frequency responses of 2D differentiators. (a) Theoretical frequency response. (b) to (h) are frequency responses of filters 1 to 7, respectively.

corresponding impulse response is shown in table 2.

It is not always suitable to impose additional constraints on the filter coefficients. Once these additional constraints are added, the value of the first term in the objective function (equation 8) will increase significally. This means the specified frequency response will probably be unable to be matched exactly. There is always a trade-off between the smoothness of the FIR filter and the error between the desired frequency and the computed frequcency.

In figure 2, a synthetic image is used to evaluate the performances of several different filters designed using nonlinear programming. Comparing the width of each boundary, particularly at the central small square block, filter 6 (z=3.0) and filter 7 give better resolution for edge-sharpening. In figure 3, LANDSAT Thematic Mapper near-infrared reflectance (band 4) data is also processed using the same filters . Note the sharpness introduced by filter 6. Since the filter 7 is designed at the expense of frequency accuracy, some unexpected information could be introduced in figure 7(f).

CONCLUSIONS

Design of two-dimensional FIR filters from frequency response specifications has been discussed. The examples shown in this paper indicate that nonlinear programming can be used quite successfully to design FIR digital filters. The frequency response of a FIR filter, obtained by minimizing equation (8), can be smoothed by forcing the magnitude of the impulse response to decrease gradually from the center of the filter to the edges.

The factors r and q in equation (8) are used to control the restrictions in the frequency and spatial domains simultaneously. The larger the ratio r/q, the greater the emphasis on exerting the accuracy of frequency response, relative to the magnitude variation of the FIR filter. The factor z in equation (8) is selected by the filter designer, based on the character of the FIR filter required and can be adjusted to force the magnitudes of filter coefficients to nearly zero at the filter edges.

If further constraints on the FIR filter are required, constrained nonlinear programming may be used to solve such problems. The result of imposing additional constraints, however, could reduce the accuracy of the FIR filter frequency response.

REFERENCES

De Meyer, F., 1974, Filter techniques in gravity interpretation: Advances in Geophysics, v.17, p.187-256.

Fiasconaro, J. G., 1979, Two-dimensional nonrecursive filters, in Picture Processing and Digital Filtering, ed. T. M. Huang, Topics in Applied Physics Series, v.6: Springer-Verlag, New York, p.69-129.

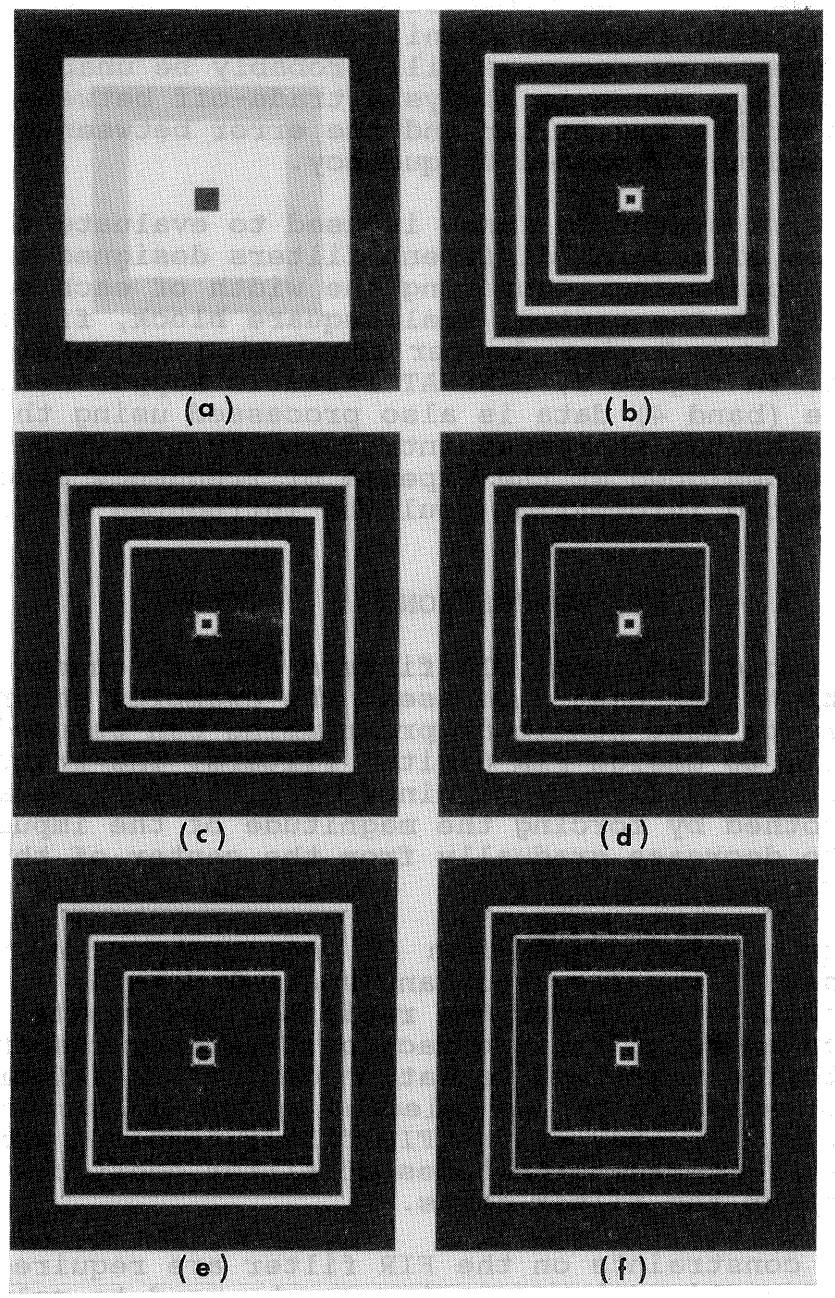

Figure 2. Synthetic image test of 2D differentiators. (a) Synthetic image. (b) to (f) are results using filters 1, 2, 4, 6, and 7.

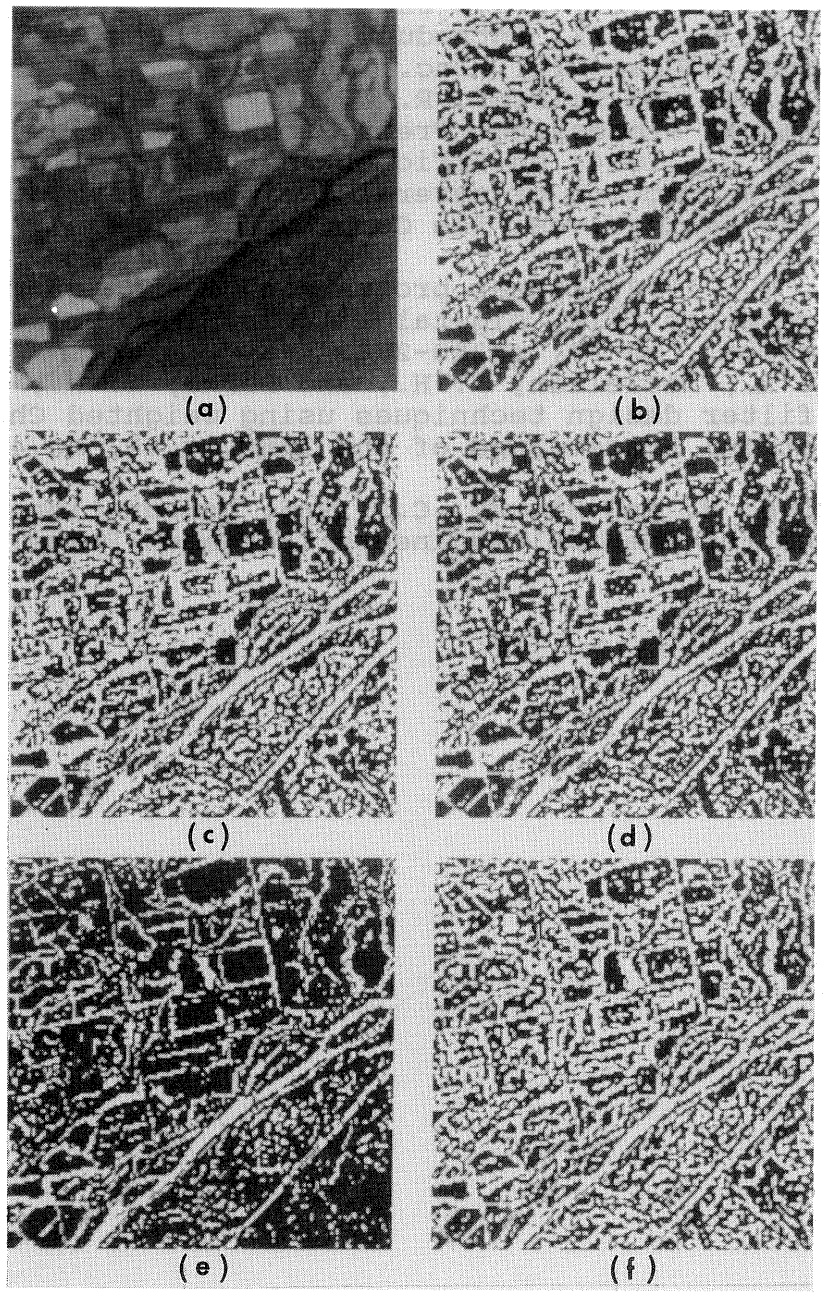

Figure 3. LANDSAT image test of 2D differentiators. (a) Original TM band 4 image. (b) to (f) are results using filters 1, 2, 4, 6, and 7.

Herrmann, O., 1970, On the design digital filters with linear phase: Electron. Lett., v.6, no.11, p.328-329.
Lewin, J. and Telljohann, M. A., 1984, Digital filter design using linear programming: SPIE v.504, Application of Digital Image Processing VII, p.221-228.
Luenberger, D. G., 1973, Introduction to linear and nonlinear programming: McGraw-Hill, Inc., New York.
Oppenheim, A. V. and Schafer, R. W., 1975, Chapter 5 of Digital Signal Processing: Prentice Hall, New Jersey.
Powell, M. J. D., 1964, An efficient method for finding the minimum of a function of several variables without calculating derivatives: The Computer Journal, v.7, p.155-162.
Rabiner, L. R., 1972, Linear programming design of finite impulse response (FIR) digital filters: IEEE Trans. Audio Electroacoust., AU-20, p.280-288.
Rabiner, L. R., McClellan, J. H., and Parks, T. W., 1975, FIR digital filter design techniques using weighted Chebyshev approximation: Proceedings of the IEEE, v.63, No.4, p.595-609.
Rosenbrock, H. H. and Storey, C., 1966, Computational techniques for chemical engineers: Pergamon Press, New York.