Approximate Convolution Using DCT Coefficient Multipliers

Neri Merhav*and Renato Kresch
VCD and HP Israel Science Center†

Abstract

We develop a method for designing DCT coefficient multipliers in order to approximate the operation of 2D convolution of an image with a given kernel. The method is easy to implement on compressed formats of DCT-based compression methods (JPEG, MPEG, H.261) by using decoding quantization tables that are different from the encoding quantization tables.

Keywords: DCT-domain processing, filtering, convolution, image enhancement, quantization tables.

*Currently on sabbatical leave at Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA.
†Address: HP Israel Science Center, Technion City, Haifa 32000, Israel. E-mail: [merhav,renato]@hp.technion.ac.il

1 Introduction

This work addresses the problem of efficient 2D linear filtering in the discrete cosine transform (DCT) domain, which is an important problem in the area of processing and manipulation of images and video streams compressed in DCT-based methods, such as JPEG, MPEG, H.261, and others (see, e.g., [1-9]).

Most of the previously reported work on DCT domain processing in general, and 2D filtering in particular, focuses on exact algorithms, that provide the precise desired results. In [10] Bhaskaran *et al.* proposed a method for sharpening scanned text and picture images by multiplying the DCT coefficients of the image by fixed multipliers that were designed using statistical considerations. Specifically, these multipliers were designed so as to match the variances of the DCT coefficients of the scanned image to desirable reference variances corresponding to a computer-generated synthetic image. Clearly, DCT domain element-by-element multiplication does not exactly correspond to spatial domain convolution (see, e.g., [1], [2], [3], and [4] for convolution-multiplication properties of the DCT), but the motivation for this approximate filtering approach is clear: Once a set of DCT coefficient multipliers is available, the DCT domain element-by-element multiplication is easy to implement on compressed streams of DCT-based compression methods with no additional computational cost. One simply uses a decoding quantization table that is different from the encoding quantization table, so that the dequantization table includes the appropriate gains.

In this work, we further study the idea of using DCT domain coefficient multipliers in order to mimic a certain image enhancement operation. Unlike the variance matching approach of Bhaskaran *et al.*, however, we aim at approximating a given convolution kernel. Specifically, the problem we address is the following: Given a 2D separable, symmetric convolution kernel in the spatial domain, we seek a set of DCT coefficient multipliers that best approximate the operation of filtering by the given kernel in the least squares sense. We provide two variants of the solution to this problem, and demonstrate their performance.

The DCT domain multiplication approach is useful in several applications where the image is distorted by a certain mechanism before being compressed and stored, and one would like to embed the multipliers in the decoding quantization table in order to compensate for this distortion. One example is a color scanner which suffers from limited modulation transfer function (MTF) and misregistration problems [10]. Another example is the digital camera whose CCD sensors typically suffer from several sources of noise: photo-electric Poisson noise due to photon-electron conversion, electronic circuitry noise, and quantization noise of the digitization phase. The reconstruction process of digital pictures also suffers from artifacts due to the fact that every pixel carries one color only. Other image and video recording media are subjected to various types of distortion and noise as well due to technological limitations.

The outline of this document is as follows. In Section 2, we provide the formulation of the problem. Section 3 contains the mathematical derivation of two methods for designing the DCT domain coefficient multipliers. In Section 4, we demonstrate the performance and discuss the properties of these design methods. Finally, in Section 5, several conclusions are drawn along with directions for further research.

2 Preliminaries and Problem Description

The 8-point 2D DCT transforms an 8×8 block $\{x(n,m)\}_{n,m=0}^{7}$ in the spatial domain into a matrix of DCT coefficients $\{X(k,l)\}_{k,l=0}^{7}$, according to the following equation [12]:

$$X(k,l) = \frac{c(k)}{2}\frac{c(l)}{2}\sum_{n=0}^{7}\sum_{m=0}^{7} x(n,m)\cos(\frac{2n+1}{16}\cdot k\pi)\cos(\frac{2m+1}{16}\cdot l\pi), \tag{1}$$

where $c(0) = 1/\sqrt{2}$ and $c(i) = 1$ for $i > 0$. The inverse transform is given by:

$$x(n,m) = \sum_{k=0}^{7}\sum_{l=0}^{7} \frac{c(k)}{2}\frac{c(l)}{2} X(k,l)\cos(\frac{2n+1}{16}\cdot k\pi)\cos(\frac{2m+1}{16}\cdot l\pi). \tag{2}$$

In a matrix form, let $\boldsymbol{x} = \{x(n,m)\}_{n,m=0}^{7}$ and $\boldsymbol{X} = \{X(k,l)\}_{k,l=0}^{7}$, and define the 8-point DCT matrix $S = \{S(k,n)\}_{k,n=0}^{7}$, where

$$S(k,n) = \frac{c(k)}{2}\cos(\frac{2n+1}{16}\cdot k\pi). \tag{3}$$

We then have

$$\boldsymbol{X} = S\boldsymbol{x}S^t \tag{4}$$

where the superscript t denotes matrix transposition, and so

$$\boldsymbol{x} = S^{-1}\boldsymbol{X}S = S^t\boldsymbol{X}S, \tag{5}$$

where the second equality follows from the unitarity of S.

Filtering, or convolution, of an input image $\{I(i,j)\}$, (where i and j are integers taking on values in ranges that correspond to the size of the image), by a filter with impulse response $\{f(i,j)\}$ (also called kernel), results in an output image $\{J(i,j)\}$ given by:

$$J(i,j) = \sum_{i'}\sum_{j'} f(i',j')I(i-i',j-j') \tag{6}$$

where the range of summation over i' and j' is, of course, according to the support of the impulse response $\{f(i,j)\}$. In this work, we assume that the filter $\{f(i,j)\}$ is *separable*, that is, $f(i,j)$ can be factorized as

$$f(i,j) = v_i h_j, \tag{7}$$

for some one-dimensional sequences $\{v_i\}$ and $\{h_j\}$. The supports of $\{v_i\}$ and $\{h_j\}$ are $-M \le i \le M$ and $-N \le j \le N$, respectively, meaning that $f(i,j) = 0$ outside a $(2M+1)\times(2N+1)$ rectangle.

Incorporating the separability assumption into eq. (6), we get

$$J(i,j) = \sum_{i'=-M}^{M} v_{i'} \sum_{j'=-N}^{N} h_{j'} I(i-i',j-j'), \tag{8}$$

namely, one can first perform a one-dimensional convolution on each row with the *horizontal filter component* (HFC) $\{h_j\}$, and then another one-dimensional convolution on each resulting column with the *vertical filter component* (VFC) $\{v_i\}$. Of course, the order can be interchanged and the vertical convolutions can be carried out first without affecting the final result.

An important special case, assumed frequently in previously reported work (see, e.g., [1, 3, 8]) as well as in this work, is that of symmetric filter components, namely, $v_i = v_{-i}$ and $h_i = h_{-i}$ for all i.

The input image $\{I(i,j)\}$ is given in the compressed domain, that is, we are given a sequence of 8×8 matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ of DCT coefficients corresponding to spatial domain 8×8 spatial domain blocks $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$ that together form the input image $\{I(i,j)\}$. Our task is to calculate a good approximation of the sequence of 8×8 matrices $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ of DCT coefficients of the spatial domain blocks $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots$ associated with the filtered image $\{J(i,j)\}$, directly from $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$, without going via the spatial domain and performing spatial domain convolution. We further assume that M and N do not exceed 8 (that is, the filter size is always smaller than 17×17), so that every DCT block of the filtered image $\{J(i,j)\}$ depends on the corresponding DCT block of the input image $\{I(i,j)\}$ and the eight immediate neighbors of $\boldsymbol{X}$.

Specifically, let

$$V = \begin{pmatrix} v_8 & v_7 & \cdot & \cdot & \cdot & v_1 & v_0 & v_1 & \cdot & \cdot & \cdot & v_7 & v_8 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & v_8 & v_7 & \cdot & \cdot & \cdot & v_1 & v_0 & v_1 & \cdot & \cdot & \cdot & v_7 & v_8 & 0 & \cdot & \cdot & 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & v_8 & v_7 & \cdot & \cdot & \cdot & v_1 & v_0 & v_1 & \cdot & \cdot & \cdot & v_7 & v_8 \end{pmatrix} \tag{9}$$

and

$$H = \begin{pmatrix} h_8 & h_7 & \cdot & \cdot & \cdot & h_1 & h_0 & h_1 & \cdot & \cdot & \cdot & h_7 & h_8 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & h_8 & h_7 & \cdot & \cdot & \cdot & h_1 & h_0 & h_1 & \cdot & \cdot & \cdot & h_7 & h_8 & 0 & \cdot & \cdot & 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 & \cdot & \cdot & \cdot & 0 & h_8 & h_7 & \cdot & \cdot & \cdot & h_1 & h_0 & h_1 & \cdot & \cdot & \cdot & h_7 & h_8 \end{pmatrix} \tag{10}$$

Let $\boldsymbol{x}$ denote a spatial domain input block of size 24×24, subdivided into nine 8×8 blocks as follows:

$$\boldsymbol{x} = \begin{pmatrix} \boldsymbol{x}_{11} & \boldsymbol{x}_{12} & \boldsymbol{x}_{13} \\ \boldsymbol{x}_{21} & \boldsymbol{x}_{22} & \boldsymbol{x}_{23} \\ \boldsymbol{x}_{31} & \boldsymbol{x}_{32} & \boldsymbol{x}_{33} \end{pmatrix} \tag{11}$$

The 8×8 output block $\boldsymbol{y}$ that corresponds to the central input block $\boldsymbol{x}_{22}$ is given by

$$\boldsymbol{y} = V\boldsymbol{x}H^t. \tag{12}$$

Our problem is the following: Given H and V, we seek a fixed 8×8 matrix G of DCT domain multipliers such that element-by-element multiplication of G by $\boldsymbol{X}_{22}$ (the DCT of $\boldsymbol{x}_{22}$), henceforth denoted by $\hat{\boldsymbol{Y}} = G \bullet \boldsymbol{X}_{22}$, would have an IDCT $\hat{\boldsymbol{y}}$ that is as close as possible to $\boldsymbol{y}$, namely, the error $\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ is "small" in some reasonable sense. The most common measure of the error magnitude is its energy $\epsilon^2 = |\boldsymbol{e}|^2$, i.e., the sum of squares of the elements of $\boldsymbol{e}$. Since $\epsilon^2 = \epsilon^2(\boldsymbol{x})$ depends also on the input $\boldsymbol{x}$, and we wish that G would be fixed and independent of $\boldsymbol{x}$, there are two possible approaches at this point. One approach, henceforth referred to as the *minimum mean squared error* (MMSE) approach, is to minimize the expectation of $\epsilon^2(\boldsymbol{x})$ w.r.t $\boldsymbol{x}$, which requires some estimates or assumptions about the second order statistics of $\boldsymbol{x}$. The second approach, which will be referred to as the *minimax* approach, minimizes $\max_{\boldsymbol{x}} \epsilon^2(\boldsymbol{x})$ subject to a constraint on the energy of $\boldsymbol{x}$. The latter approach is somewhat more pessimistic but it to avoids the dependence upon the second order statistics of $\boldsymbol{x}$. Both approaches will be discussed in the next section.

3 Mathematical Derivation

By Parseval's theorem and the unitarity of the DCT, the spatial domain error energy ϵ^2 remains unchanged under the DCT, i.e., $\boldsymbol{e}$ and its DCT $\boldsymbol{E} = S\boldsymbol{e}S^t$ have the same energy. Therefore, we can seek the best multiplier matrix G directly in the DCT domain by minimizing the energy of $\boldsymbol{E}$.

To this end, let us partition the matrix V into three 8×8 matrices $V = [V_1, V_2, V_3]$, where

$$V_1 = \begin{pmatrix} v_8 & v_7 & v_6 & v_5 & v_4 & v_3 & v_2 & v_1 \\ 0 & v_8 & v_7 & v_6 & v_5 & v_4 & v_3 & v_2 \\ 0 & 0 & v_8 & v_7 & v_6 & v_5 & v_4 & v_3 \\ 0 & 0 & 0 & v_8 & v_7 & v_6 & v_5 & v_4 \\ 0 & 0 & 0 & 0 & v_8 & v_7 & v_6 & v_5 \\ 0 & 0 & 0 & 0 & 0 & v_8 & v_7 & v_6 \\ 0 & 0 & 0 & 0 & 0 & 0 & v_8 & v_7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & v_8 \end{pmatrix} \tag{13}$$

$$V_2 = \begin{pmatrix} v_0 & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & v_7 \\ v_1 & v_0 & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \\ v_2 & v_1 & v_0 & v_1 & v_2 & v_3 & v_4 & v_5 \\ v_3 & v_2 & v_1 & v_0 & v_1 & v_2 & v_3 & v_4 \\ v_4 & v_3 & v_2 & v_1 & v_0 & v_1 & v_2 & v_3 \\ v_5 & v_4 & v_3 & v_2 & v_1 & v_0 & v_1 & v_2 \\ v_6 & v_5 & v_4 & v_3 & v_2 & v_1 & v_0 & v_1 \\ v_7 & v_6 & v_5 & v_4 & v_3 & v_2 & v_1 & v_0 \end{pmatrix} \tag{14}$$

and $V_3 = V_1^t$. In the same fashion, H is partitioned into $[H_1, H_2, H_3]$ with similar definitions of H_1, H_2, and H_3.

The ideal convolution can now be expressed as

$$\boldsymbol{y} = \sum_{i=1}^{3}\sum_{j=1}^{3} V_i \boldsymbol{x}_{ij} H_j^t. \tag{15}$$

Since the DCT is unitary, it is distributive w.r.t matrix multiplication, and so the last equation can be written in the DCT domain as

$$\boldsymbol{Y} = \sum_{i=1}^{3}\sum_{j=1}^{3} \boldsymbol{V}_i \boldsymbol{X}_{ij} \boldsymbol{H}_j^t, \tag{16}$$

where $\boldsymbol{Y}$, $\boldsymbol{V}_i$, $\boldsymbol{X}_{ij}$, and $\boldsymbol{H}_j$ are the 2D-DCT's of $\boldsymbol{y}$, V_i, $\boldsymbol{x}_{ij}$, and H_j, respectively, $i, j = 1, 2, 3$. Now, let

$$\boldsymbol{E} = \boldsymbol{Y} - \hat{\boldsymbol{Y}} = \sum_{i=1}^{3}\sum_{j=1}^{3} \boldsymbol{V}_i \boldsymbol{X}_{ij} \boldsymbol{H}_j^t - G \bullet \boldsymbol{X}_{22}. \tag{17}$$

In order to express the element-by-element multiplication $G \bullet \boldsymbol{X}_{22}$ in terms of ordinary algebraic matrix multiplication, it will be convenient to represent the data $\{\boldsymbol{X}_{ij}\}$ in a one-dimensional representation by *column stacking* [11, Sections 5.3-5.4]. The column-stacked version $\bar{Z}$ of an $m \times n$ matrix Z is a (mn)-dimensional column vector formed by concatenating the columns of Z from left to right. The basic fact that will be used hereafter with regard to column stacking is that if $W = AZB^t$ where Z is as above, and A and B are matrices of dimensions $k \times m$ and $l \times n$, respectively, then $\bar{W} = (B \otimes A)\bar{Z}$, where $B \otimes A$ is the Kroenecker tensor product of B and A, defined as

$$B \otimes A = \begin{pmatrix} b_{11}A & b_{12}A & \cdot & \cdot & \cdot & b_{1n}A \\ b_{21}A & b_{22}A & \cdot & \cdot & \cdot & b_{2n}A \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{l1}A & b_{l2}A & \cdot & \cdot & \cdot & b_{ln}A \end{pmatrix}. \tag{18}$$

Returning to the approximate filtering problem, we can now rewrite eq. (17) in the column-stacked representation as follows:

$$\bar{\boldsymbol{E}} = \sum_{i=1}^{3}\sum_{j=1}^{3} F_{ij} \bar{\boldsymbol{X}}_{ij} - D\bar{\boldsymbol{X}}_{22} \tag{19}$$

where $F_{ij} = H_j \otimes V_i$, $i, j = 1, 2, 3$, and $D = \text{diag}\{\bar{G}\}$.

3.1 The MMSE Approach

In this approach, we would like to minimize the expectation of $\epsilon^2 = \bar{\boldsymbol{E}}^t \bar{\boldsymbol{E}} = \text{tr}\{\bar{\boldsymbol{E}}\bar{\boldsymbol{E}}^t\}$ over D (or, equivalently G). Let $R_{ij,kl} = E\{\bar{\boldsymbol{X}}_{ij}\bar{\boldsymbol{X}}_{kl}^t\}$. Then,

$$\mathbf{E}\epsilon^2 = \mathbf{E}\text{tr}\left[(\sum_{i=1}^{3}\sum_{j=1}^{3} F_{ij}\bar{\boldsymbol{X}}_{ij} - D\bar{\boldsymbol{X}}_{22})(\sum_{k=1}^{3}\sum_{l=1}^{3} F_{kl}\bar{\boldsymbol{X}}_{kl} - D\bar{\boldsymbol{X}}_{22})^t\right]$$

$$
\begin{aligned}
= \quad & \mathrm{tr}\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3} F_{ij}R_{ij,kl}F_{kl}^t - D\sum_{k=1}^{3}\sum_{l=1}^{3} R_{22,kl}F_{kl}^t \right. \\
& \left. -\sum_{i=1}^{3}\sum_{j=1}^{3} F_{ij}R_{ij,22}D + DR_{22,22}D\right). \qquad (20)
\end{aligned}
$$

By taking partial derivatives w.r.t the diagonal elements $\{d_i\}_{i=1}^{64}$ and setting to zero, one obtains a set of 64 decoupled linear equations with 64 unknowns whose solutions are given by

$$
d_k^* = \frac{\sum_{i=1}^3 \sum_{j=1}^3 (F_{ij}R_{ij,22})(k,k)}{R_{22,22}(k,k)}, \qquad (21)
$$

where $A(i,j)$ is understood as the ijth element of a matrix A. Therefore, the optimal DCT domain gain factor $g_{ij} = G(i,j)$, $i,j = 1,...,8$ is given by d_k^*, where $k = 8(j-1)+i$, which corresponds to the column stacking order.

As can be seen, the optimal solution depends not only on the given convolution kernel (via $\{F_{ij}\}$), but also on the covariance matrices $\{R_{ij,22}\}$ of the DCT domain data. Therefore, in order to use this solution, one must estimate these covariance matrices from sample images, or to assume a certain form. We will adopt the second approach.

Before doing that, we note that by substituting eq. (21) into eq. (20), we get the following expression for the MMSE.

$$
(\mathbf{E}\epsilon^2)_{\min} = \mathrm{tr}\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3} F_{ij}R_{ij,kl}F_{kl}^t\right) - \sum_{m=1}^{64} \frac{[\sum_{i=1}^3 \sum_{j=1}^3 (F_{ij}R_{ij,22})(m,m)]^2}{R_{22,22}(m,m)}. \qquad (22)
$$

This expression, that provides a measure of the goodness of fit, gives a guideline about the conditions under which a given filter can be well approximated by DCT coefficient multipliers. The ratio between the first term of eq. (22) and $(\mathbf{E}\epsilon^2)_{\min}$ is the signal-to-noise ratio corresponding to the approximation. As expected, when $\boldsymbol{H}_2$ and $\boldsymbol{V}_2$ and hence also F_{22} are diagonally dominant, the MMSE is relatively small.

For the sake of simplicity in implementing eq. (21), we shall adopt a spatial domain, separable, first order Markov model [11, Sect. 5.6]. According to this model, the spatial domain covariance between two pixel intensities $x(n_1, m_1)$ and $x(n_2, m_2)$ is given by

$$
r(n_1,m_1,n_2,m_2) \stackrel{\Delta}{=} \mathbf{E}[x(n_1,m_1)x(n_2,m_2)] = \sigma^2\rho^{|n_1-n_2|+|m_1-m_2|}, \qquad (23)
$$

where ρ is a parameter in the range $(-1,1)$, and σ^2 is a scaling factor whose value is immaterial for eq. (21) and hence will be assumed unity. The covariance matrices $\{R_{ij,22}\}$ in this case, are obtained as follows. Let r_0 and r_1 be 8×8 Toeplitz matrices whose ijth elements are $\rho^{|i-j|}$ and $\rho^{|8+j-i|}$, respectively. Let R_0 and R_1 denote the 2D-DCT's of r_0 and r_1, respectively, i.e., $R_0 = Sr_0S^t$ and $R_1 = Sr_1S^t$. Then,

$$
R_{ij,22} = \boldsymbol{R}_j \otimes \boldsymbol{R}_i, \qquad (24)
$$

where

$$
\boldsymbol{R}_i = \begin{cases} R_1 & \text{if } i = 1 \\ R_0 & \text{if } i = 2 \\ R_1^t & \text{if } i = 3 \end{cases} \qquad (25)
$$

Thus, the numerator of eq. (21) degenerates to

$$\begin{aligned}
\sum_{i=1}^{3}\sum_{j=1}^{3} F_{ij} R_{ij,22} &= \sum_{i=1}^{3}\sum_{j=1}^{3} (\boldsymbol{H}_j \otimes \boldsymbol{V}_i)(\boldsymbol{R}_j \otimes \boldsymbol{R}_i) \\
&= \sum_{i=1}^{3}\sum_{j=1}^{3} (\boldsymbol{H}_j \boldsymbol{R}_j) \otimes (\boldsymbol{V}_i \boldsymbol{R}_i) \\
&= (\sum_{l=1}^{3} \boldsymbol{H}_l \boldsymbol{R}_l) \otimes (\sum_{l=1}^{3} \boldsymbol{V}_l \boldsymbol{R}_l) \\
&\triangleq \boldsymbol{H}_R \otimes \boldsymbol{V}_R. \qquad (26)
\end{aligned}$$

Hence, for $k = 8(j-1)+i$, we get

$$d_k^* = g_{ij} = \frac{\boldsymbol{V}_R(i,i)}{R_0(i,i)} \cdot \frac{\boldsymbol{H}_R(j,j)}{R_0(j,j)}. \qquad (27)$$

In other words, the matrix G in this case is just the outer product of two vectors formed by the diagonals of $\boldsymbol{V}_R$, $\boldsymbol{H}_R$, and R_0, which means that the optimum two dimensional MMSE solution separates into the combination of the two optimum one dimensional solutions corresponding to the horizontal convolution and the vertical convolution. In the special case where $\rho = 0$, i.e., $R_0 = r_0 = I$ and $R_1 = r_1 = 0$, we simply get $g_{ij} = \boldsymbol{V}_2(i,i)\boldsymbol{H}_2(j,j)$.

Incorporating the Quantization Error

Since this work is primarily motivated by embedding the multipliers in the quantization tables, as explained in the Introduction, a natural refinement of this method would be to incorporate the effect of quantization errors, and to optimize the DCT-domain gains so as to minimize the combined effect of approximation error and quantization error. In this subsection, we examine the effect of quantization error on the design of the multipliers.

If we consider the JPEG algorithm, then at the encoder, every DCT coefficient $\boldsymbol{X}_{22}(i,j)$ is first divided by the encoding step-size $\delta_e(i,j)$, and then rounded to the closest integer. At the decoder, the resultant integer is multiplied by the decoding step-size $\delta_d(i,j)$ (which is traditionally identical to $\delta_e(i,j)$), and so the decoded DCT coefficient is given by

$$\hat{\boldsymbol{X}}_{22}(i,j) = \frac{\delta_d(i,j)}{\delta_e(i,j)} \cdot \boldsymbol{X}_{22}(i,j) + \delta_d(i,j)Q(i,j) \qquad (28)$$

where $-0.5 \le Q(i,j) < 0.5$ is the roundoff error at the encoder. If we identify the ratio $\delta_d(i,j)/\delta_e(i,j)$ as g_{ij}, then the first term is the desired term and the second is an error term. Thus, we rewrite eq. (28) as

$$\hat{\boldsymbol{X}}_{22}(i,j) = g_{ij}\boldsymbol{X}_{22}(i,j) + g_{ij}\delta_e(i,j)Q(i,j). \qquad (29)$$

Assuming that the encoding quantization table $\delta_e = \{\delta_e(i,j)\}$ is fixed and only the decoding table $\delta_d = \{\delta_d(i,j)\}$ absorbs the multipliers (so as to avoid any effects on the compressibility), then eq. (19) is now rewritten as

$$\bar{\boldsymbol{E}} = \sum_{i=1}^{3}\sum_{j=1}^{3} F_{ij}\bar{\boldsymbol{X}}_{ij} - D\bar{\boldsymbol{X}}_{22} + D\Delta_e\bar{Q} \qquad (30)$$

where $\Delta_e = \text{diag}\{\bar{\delta}_e\}$, and $\bar{Q}$ is the column stacked version of the roundoff error matrix $Q = \{Q(i,j)\}$. The error signal now has two components. The first component is the approximation error, which is given by the first two terms as before. The second component is the quantization error given by the third term. If we assume that these two components are uncorrelated (which is a reasonable assumption when $\{\delta_e(i,j)\}$ are fairly small), then similarly as in (20), we obtain

$$\begin{aligned}\mathbf{E}\epsilon^2 \;=\; \text{tr}\Bigg[&\sum_{i=1}^{3}\sum_{j=1}^{3}\sum_{k=1}^{3}\sum_{l=1}^{3} F_{ij}R_{ij,kl}F_{kl}^t - D\sum_{k=1}^{3}\sum_{l=1}^{3} R_{22,kl}F_{kl}^t \\ &- \sum_{i=1}^{3}\sum_{j=1}^{3} F_{ij}R_{ij,22}D + D(R_{22,22} + \Delta_e R_Q \Delta_e)D\Bigg], \end{aligned} \tag{31}$$

where R_Q is the covariance matrix of $\bar{Q}$. If we further assume that R_Q is diagonal (i.e., the roundoff errors are uncorrelated), then the optimal gains are as in eq. (21) except that the denominator is replaced by $R_{22,22}(k,k) + \bar{\delta}_e^2(k)R_Q(k,k)$. This means that the gain factors are reduced by a factor of $R_{22,22}(k,k)/[R_{22,22}(k,k) + \bar{\delta}_e^2(k)R_Q(k,k)]$, which (similarly as in the Wiener solution), is the best compromise between the desired response and noise suppression.

In order to obtain a rough assessment on the order of magnitude of this attenuation factor, let us assume that each $Q(i,j)$ is uniformly distributed in $[-0.5, 0.5)$, and so $R_Q(k,k) = 1/12$ for all k. Now, for the recommended JPEG quantization table, the step-sizes $\delta_e(i,j)$ for the low (and typically important) frequency components (say, $i + j \le 5$) are all less than or equal to 16. Thus, $\bar{\delta}_e^2(k)R_Q(k,k)$ does not exceed $16^2/12 = 21.333$. On the other hand, the variances of these low frequency DCT coefficients $R_{22,22}(k,k)$ are typically of the order of magnitude of 10^3 or 10^4, namely, at least 2 or 3 orders of magnitude larger than the quantization error term. Thus, at least for the important frequency components, we do not expect the gain factors to be affected significantly by the quantization error.

3.2 The Minimax Approach

As an alternative to the MMSE approach, one might consider the more conservative minimax approach, where instead of minimizing $\mathbf{E}\epsilon^2(\boldsymbol{x})$, one minimizes the maximum of $\epsilon^2(\boldsymbol{x})$ where the input $\boldsymbol{x}$ has a given energy.

To this end, we will rewrite eq. (19) in a slightly different manner. Let $\bar{\boldsymbol{X}}$ denote the 576-dimensional column vector formed by the concatenation of $\bar{\boldsymbol{X}}_{11}, \bar{\boldsymbol{X}}_{12}, ..., \bar{\boldsymbol{X}}_{33}$ in a block column stacking order. Let $\boldsymbol{V} = [\boldsymbol{V}_1, \boldsymbol{V}_2, \boldsymbol{V}_3]$, $\boldsymbol{H} = [\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3]$, and let

$$\boldsymbol{D} = [\boldsymbol{O}\ \boldsymbol{O}\ \boldsymbol{O}\ \boldsymbol{O}\ D\ \boldsymbol{O}\ \boldsymbol{O}\ \boldsymbol{O}\ \boldsymbol{O}], \tag{32}$$

where $\boldsymbol{O}$ is the 8×8 all-zero matrix and D is as above. Then, eq. (19), can be rewritten as

$$\boldsymbol{E} = [(\boldsymbol{H} \otimes \boldsymbol{V}) - \boldsymbol{D}]\bar{\boldsymbol{X}} \tag{33}$$

and therefore

$$\epsilon^2(\bar{\boldsymbol{X}}) = \boldsymbol{E}^t\boldsymbol{E} = \bar{\boldsymbol{X}}^t[(\boldsymbol{H} \otimes \boldsymbol{V}) - \boldsymbol{D}]^t[(\boldsymbol{H} \otimes \boldsymbol{V}) - \boldsymbol{D}]\bar{\boldsymbol{X}}. \tag{34}$$

Minimizing over D the maximum of $\epsilon^2(\bar{\boldsymbol{X}})$ subject to an input energy constraint $\bar{\boldsymbol{X}}^t\bar{\boldsymbol{X}} \leq A$ is equivalent to minimizing the largest eigenvalue of the matrix $[(\boldsymbol{H}\otimes\boldsymbol{V})-\boldsymbol{D}]^t[(\boldsymbol{H}\otimes\boldsymbol{V})-\boldsymbol{D}]$, which is a 576×576 matrix. This in turn is equivalent to minimizing the largest eigenvalue of the 64×64 matrix $[(\boldsymbol{H}\otimes\boldsymbol{V})-\boldsymbol{D}][(\boldsymbol{H}\otimes\boldsymbol{V})-\boldsymbol{D}]^t$, which is still a large matrix dimension for any iterative search for the optimum D.

To alleviate this difficulty, we adopt a suboptimal solution that separates the two dimensional problem into two one-dimensional problems of the vertical convolution and the horizontal convolution. For the one-dimensional vertical convolution, consider three 8-dimensional column vectors of 1D-DCT coefficients $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$. The desired convolution result corresponding to $\boldsymbol{X}_2$ is given by

$$\boldsymbol{Y} = \boldsymbol{V}_1\boldsymbol{X}_1 + \boldsymbol{V}_2\boldsymbol{X}_2 + \boldsymbol{V}_3\boldsymbol{X}_3 \tag{35}$$

and the approximation is given by $\hat{\boldsymbol{Y}} = D_v\boldsymbol{X}_2$, where D_v is a diagonal matrix corresponding to the VFC. The error is given by $\boldsymbol{E} = [\boldsymbol{V}_1, \boldsymbol{V}_2 - D_v, \boldsymbol{V}_3]\boldsymbol{X}$, where $\boldsymbol{X}$ denotes the 24-dimensional column vector formed by concatenating $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, and $\boldsymbol{X}_3$. Therefore, the one-dimensional minimax problem is that of minimizing w.r.t D_v the largest eigenvalue of the 8×8 matrix

$$W(D_v) = \boldsymbol{V}_1\boldsymbol{V}_1^t + (\boldsymbol{V}_2 - D_v)^2 + \boldsymbol{V}_3\boldsymbol{V}_3^t \tag{36}$$

where we have used the symmetry of $\boldsymbol{V}_2$ and D_v. A natural initial guess for an iterative search for the optimum D_v would be to set the diagonal elements of D_v to be the same as the corresponding diagonal elements of $\boldsymbol{V}_2$. If $|v_0|$ is considerably larger than all $|v_i|$ for all $i \neq 0$, then $\boldsymbol{V}_2$ is diagonally dominant, and this initial guess is already fairly close to the optimum solution. (Note also that this is equivalent to the MMSE solution for $\rho = 0$ as described above.) The iterative optimization algorithm that we have used was the Nedler-Meade simplex search for unconstrained optimization, which is implemented by the MATLAB library function `fmins`.

The proposed sub-optimum minimax procedure is to find the optimum diagonal matrix D_v^* for vertical convolution, and similarly, the optimum diagonal matrix D_h^* for horizontal convolution, and then to approximate $\boldsymbol{Y}$ as $D_v^*\boldsymbol{X}_{22}D_h^*$. This means that g_{ij} is given by the product of the iith element of D_v^* and the jjth element of D_h^*.

4 Experimental Results and Discussion

We have simulated both the MMSE approach and the minimax approach (without quantization) and examined their performance on real images in comparison to the true convolution. As will be seen, the approximate convolution method works well for convolution kernels where the central coefficient (h_0 or v_0) is considerably larger than other coefficients, (e.g., by a factor of 2 or 3 at least). For kernels that do not have this property, e.g., the 5×5 uniform weight averaging kernel, we have witnessed blocky-ness effects in the resulting image, due to error discontinuities at the boundaries between blocks.

We first examined the design of a lowpass filter for noise cleaning applications. The desired lowpass filter is given by $h_0 = v_0 = 0.5$, $h_1 = v_1 = 0.25$ and $h_i = v_i = 0$ for all $i > 1$. Fig. 1 illustrates the original image, fig. 2 is a noisy version, fig. 3 is the noisy image after exact convolution with the above filter, fig. 4 is the result of DCT domain multiplication, where the multipliers were designed using the MMSE approach with $\rho = 0.9$, and fig. 5

is associated with DCT domain multipliers designed by the minimax approach. As can be seen, the approximate methods give images that are visually equivalent to that of the exact convolution image. We have also examined the MMSE approach with various values of ρ in the range $[0, 0.99]$ but since h_0 and v_0 dominate the other coefficients, the resulting multipliers were not very sensitive to ρ and the resultant images looked quite the same. (There are merely minor changes in the multiplier values when ρ varies in that range.)

In a second experiment, we examined the design of an approximate highpass filter for edge sharpening applications. The desired highpass filter is given by $h_0 = v_0 = 3$, $h_1 = v_1 = -1$, and $h_i = v_i = 0$ for all $i > 1$. Fig. 6 is an original image of scanned text, fig. 7 is the resulting image after exact convolution with the above filter, fig. 8 is the result of DCT domain multiplication, where the multipliers were designed using the MMSE approach with $\rho = 0.9$, and fig. 9 is associated with the minimax approach. As can be seen, the MMSE approach gives a result similar to that of the exact convolution, that is, sharpening the text at the expense of noticeable background noise. (Again, the results of the former were not very sensitive to ρ.) The minimax approach, on the other hand, also enhances the text, but the background is significantly cleaner.

In other experiments, with different kernels and different images, we always found that both the MMSE and the minimax approach provide results that are perceptually equivalent to that of the exact convolution, where sometimes the minimax approach, which does not depend on the image statistics, is somewhat better.

5 Conclusion and Extensions

We have developed two methods for designing DCT domain coefficient multipliers, the MMSE approach and the minimax approach. The first method depends on the second order statistics of the image, or the class of images under consideration. If the covariance of the image is assumed separable, the two dimensional problem breaks, without loss in optimality, into two separate one dimensional problems corresponding to the vertical convolution and the horizontal convolution. If, in addition, the central kernel coefficient is considerably large compared to the other coefficients, then the resulting multipliers are relatively insensitive to the spatial domain correlation between pixels. The second method does not depend on the statistics of the image. Although we were unable to prove that the minimax problem, splits without loss of optimality, into separate row and column problems, we have adopted this approach for reasons of simplicity. Nevertheless, the suboptimal minimax approach provided results which are equivalent or even better than the MMSE approach in approximating the exact convolution.

It should be kept in mind that no matter what is the design criterion, DCT coefficient multiplication can efficiently approximate symmetric kernels only. For example, if the kernel is antisymmetric then V_2 and hence also $\boldsymbol{V}_2$ is an antisymmetric matrix, which means that it cannot be diagonally dominant (as the main diagonal is all-zero), and so there is no hope to approximate $\boldsymbol{V}_2$ efficiently by a diagonal matrix D even in the one dimensional case. Separability, however, is not a mandatory condition, as at least the MMSE approach can be extended to the nonseparable case.

Another possible interesting extension is that of using a *minimum weighted mean squared error* rather than the ordinary MMSE criterion. The weighting can be attributed either to the spatial domain or to the DCT domain. In the former case, one has control of the tradeoff between errors at block boundaries and errors at internal pixels, which might help in reducing possible blocky-ness effects. In the latter case, one may want to assign higher weights to the more important frequency components, e.g., the DC component. A parallel extension is possible for the minimax approach.

6 Acknowledgement

We would like to thank Vasudev Bhaskaran for his assistance in the experimental part of this work.

7 References

[1] W. H. Chen and S. C. Fralick, "Image enhancement using cosine transform filtering," *Image Sci. Math. Symp.*, Monterey, CA, November 1976.

[2] K. N. Ngan and R. J. Clarke, "Lowpass filtering in the cosine transform domain," *Int. Conf. on Commun.*, Seattle, WA, pp. 37.7.1-37.7.5, June 1980.

[3] B. Chitprasert and K. R. Rao, "Discrete cosine transform filtering," *Signal Processing*, Vol. 19, pp. 233-245, 1990.

[4] S. A. Martucci, "Symmetric convolution and discrete sine and cosine transforms," *IEEE Trans. on Signal Processing*, Vol. SP-42, no. 5, pp. 1038-1051, May 1994.

[5] J. B. Lee and B. G. Lee, "Transform domain filtering based on pipelining structure," *IEEE Trans. on Signal Processing*, Vol. SP-40, no. 8, pp. 2061-2064, August 1992.

[6] S.-F. Chang and D. G. Messerschmitt, "Manipulation and compositing of MC-DCT compressed video," *IEEE J. Selected Areas in Communications*, Vol. 13, no. 1, pp. 1-11, January 1995.

[7] A. Neri, G. Russo, and P. Talone, "Inter-block filtering and down-sampling in DCT domain," *Signal Processing: Image Communication*, Vol. 6, pp. 303-317, 1994.

[8] N. Merhav and V. Bhaskaran, "A fast algorithm for DCT domain filtering," HPL Technical Report #HPL-95-56, May 1995.

[9] R. Kresch and N. Merhav, "Fast DCT Domain filtering using the DCT and the DST," HPL Technical Report, in preparation.

[10] V. Bhaskaran, G. Beretta, and K. Konstantinides, "Test and Image Sharpening of JPEG compressed Images in the Frequency Domain," HPL Technical Report #HPL-94-90, October 1994.

[11] W. K. Pratt, *Digital Image Processing*, John Wiley & Sons, second edition, 1991.

[12] K. R. Rao, and P. Yip, *Discrete Cosine Transform: Algorithms, Advantages, Applications*, Academic Press 1990.

Figure 1: The original image.

Figure 2: The noisy image.

Figure 3: Exact convolution with a noise cleaning filter.

Figure 4: MMSE approach with $\rho = 0.9$.

Figure 5: The minimax approach.

One of the greatest scientific minds of all time, Albert Einstein is best known for his contributions to the field of physics. Born in Germany in 1879, Einstein received his diploma from the Swiss Federal Polytechnic School in Zurich, where he trained as a teacher in physics and mathematics. In 1905, he received his Ph.D. and published four research papers, the most significant being the creation of the special theory of relativity. He became internationally famous when he was awarded the Nobel Prize for Physics in 1922.

The important military implications of the discovery of the fission of uranium in 1939 led Einstein to appeal to President Franklin Roosevelt. Einstein's letter to the president led to the development of the atomic bomb.

Einstein left the field of physics greatly changed through his brilliant contributions. His discoveries provided the impetus for future research into understanding the mysteries of the universe.

Figure 6: The original image.

One of the greatest scientific minds of all time, Albert Einstein is best known for his contributions to the field of physics. Born in Germany in 1879, Einstein received his diploma from the Swiss Federal Polytechnic School in Zurich, where he trained as a teacher in physics and mathematics. In 1905, he received his Ph.D. and published four research papers, the most significant being the creation of the special theory of relativity. He became internationally famous when he was awarded the Nobel Prize for Physics in 1922.

The important military implications of the discovery of the fission of uranium in 1939 led Einstein to appeal to President Franklin Roosevelt. Einstein's letter to the president led to the development of the atomic bomb.

Einstein left the field of physics greatly changed through his brilliant contributions. His discoveries provided the impetus for future research into understanding the mysteries of the universe.

Figure 7: Exact convolution.

One of the greatest scientific minds of all time, Albert Einstein is best known for his contributions to the field of physics. Born in Germany in 1879, Einstein received his diploma from the Swiss Federal Polytechnic School in Zurich, where he trained as a teacher in physics and mathematics. In 1905, he received his Ph.D. and published four research papers, the most significant being the creation of the special theory of relativity. He became internationally famous when he was awarded the Nobel Prize for Physics in 1922.

The important military implications of the discovery of the fission of uranium in 1939 led Einstein to appeal to President Franklin Roosevelt. Einstein's letter to the president led to the development of the atomic bomb.

Einstein left the field of physics greatly changed through his brilliant contributions. His discoveries provided the impetus for future research into understanding the mysteries of the universe.

Figure 8: The MMSE approach with $\rho = 0.9$.

One of the greatest scientific minds of all time, Albert Einstein is best known for his contributions to the field of physics. Born in Germany in 1879, Einstein received his diploma from the Swiss Federal Polytechnic School in Zurich, where he trained as a teacher in physics and mathematics. In 1905, he received his Ph.D. and published four research papers, the most significant being the creation of the special theory of relativity. He became internationally famous when he was awarded the Nobel Prize for Physics in 1922.

The important military implications of the discovery of the fission of uranium in 1939 led Einstein to appeal to President Franklin Roosevelt. Einstein's letter to the president led to the development of the atomic bomb.

Einstein left the field of physics greatly changed through his brilliant contributions. His discoveries provided the impetus for future research into understanding the mysteries of the universe.

Figure 9: The minimax approach.