JOINT THRESHOLDING AND QUANTIZER SELECTION FOR DECODER-COMPATIBLE BASELINE JPEG

Matthew Crouse and Kannan Ramchandran

Beckman Institute
University of Illinois at Urbana-Champaign
Urbana, IL 61801

ABSTRACT

This paper introduces a novel, image-adaptive, encoding scheme for the baseline JPEG standard [1, 2]. In particular, coefficient thresholding, JPEG quantization matrix (Q-matrix) optimization, and adaptive Huffman entropy-coding are jointly performed to maximize coded still-image quality within the constraints of the baseline JPEG syntax. Adaptive JPEG coding has been addressed in earlier works: in [3], where fast rate-distortion (R-D) optimal coefficient thresholding was described, and in [4, 5], where R-D optimized Q-matrix selection was performed. By formulating an algorithm which optimizes these two operations jointly, we have obtained performance comparable to more complex, "state-of-the-art" coding schemes: for the "Lenna" image at 1 bpp, our algorithm has achieved a PSNR of 39.6 dB. This result represents a gain of 1.7 dB over JPEG with customized Huffman entropy coder, and even slightly exceeds the published performance of Shapiro's wavelet-based scheme [6]. Furthermore, with the choice of appropriate visually-based error metrics, noticeable subjective improvement has been achieved as well.

1. INTRODUCTION

Recent times have seen an explosion in the popularity of the JPEG coding system [1, 2] due to its viability for diverse commercial applications in image compression. This has motivated the recent study of powerful adaptive coding schemes which remain faithful to the JPEG syntax [3, 4, 5], so as to permit the continued use of the baseline JPEG decoders that are of great commercial value. This paper is thus motivated by the practical considerations of abounding applications (like digital image libraries, centralized image storage banks, and image transfer over networks) where extra encoder complexity is allowable, but the continued compatibility and simplicity of current JPEG decoders is essential.

Let us briefly review the baseline JPEG syntax. First, each color component of the image is partitioned into 8x8 pixel blocks. Each block is then independently transformed by an 8x8 Discrete-Cosine-Transform (DCT), and quantized with an 8x8 matrix of uniform scalar quantizer stepsizes, the Q-matrix. The quantized blocks are subsequently entropy-coded in raster scan order, with the DC component differentially coded from block to block and the AC components zero run-length coded within each block. For the AC case, a zig-zag scan is used to order the coefficients, with a Huffman coding table assigning a codeword to each nonzero quantized DCT coefficient based on its amplitude and the number of zeros preceding it in the zig-zag scan. In the DC case, a Huffman codeword is assigned based on the difference between the DC coefficient of the current block and the DC coefficient of the previous block.

The JPEG Q-matrix, whose 64 integer elements range from 1 to 255, largely determines the quality and compression of the JPEG-coded image. Although the JPEG syntax allows the Q-matrix to be customized at the encoder, typically a scaled version of the "example" JPEG Q-matrix (which has become the "de facto") is used, with the scale-size trading compression for quality. This scaling method is potentially suboptimal, since image-adaptive, R-D tradeoffs are not fully explored. As an alternative to scaling, R-D optimized Q-matrix algorithms have been proposed in [4, 5]. However, a fundamental restriction of the Q-matrix syntax is that the Q-matrix cannot be locally adapted. Hence, although Q-matrix optimization performs well in an "average" sense, potential gain can be accrued by exploiting discrepancies between the local and "average" image statistics. This problem has been addressed in [3], where a fast, R-D optimal coefficient thresholding "kernel" was developed. The idea of thresholding is that inefficiently coded coefficients (i.e. coefficients whose contribution to reducing coding distortion is not worth their cost in bits) may be thresholded, their bits being allocated to coding "more worthwhile" coefficients. Although the thresholding kernel itself is optimal given a fixed Huffman table and Q-matrix, the potential suboptimality of [3] occurs in Q-matrix selection, which is done by the scaling method. Logically, R-D optimal techniques should be applied to jointly choose the Q-matrix and set of coefficients to threshold. Furthermore, since the JPEG syntax provides for customization of the Huffman coding table, for improved performance this parameter should also be optimized. These facts motivate the following joint optimization problem that we will attempt to solve.

2. PROBLEM STATEMENT

Let c^b_{ij} represent the image DCT coefficient at block b and spatial frequency (i, j) of an NxN image. Then, our optimization parameters consist of the following (Figure 1): an

This work was supported in part by the National Science Foundation under grant NSF 92-122 (RIA-94).

0-7803-2431-5/95 $4.00 © 1995 IEEE

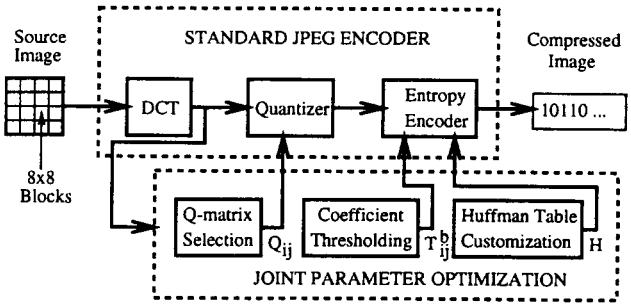

Figure 1: Optimizing the encoding of an NxN image within the constraints of the JPEG syntax. The allowable parameters, over each block b and DCT spatial frequency (i,j), are the entries of the 8x8 Q-matrix, Q_{ij}, a set of coefficient thresholding parameters, T^b_{ij}, and the Huffman table, H.

8x8 Q-matrix with stepsizes Q_{ij} that quantize coefficients at spatial frequency (i,j), a Huffman coding table H, and a set of binary thresholding parameters $T^b{}_{ij}$ that signal whether to threshold coefficient $c^b{}_{ij}$, i.e.

$$T^b{}_{ij} = \begin{cases} 1 & \Rightarrow \text{Transmit } c^b{}_{ij} \\ 0 & \Rightarrow \text{Zero-out } c^b{}_{ij} \end{cases}$$

Let R and D represent the bit rate and coding distortion of the coded image for a certain set of parameters. Then, for $i,j = 0,\ldots,7$ & $b = 1,\ldots,\frac{N^2}{64}$, the optimization problem becomes

$$\min_{T^b_{ij},Q_{ij},H} [D(T^b_{ij},Q_{ij})] \text{ s.t. } R(T^b_{ij},Q_{ij},H) < R_{budget} \quad (1)$$

3. PREVIOUS WORK

3.1. Adaptive quantizer selection

R-D optimized Q-matrix selection algorithms have been presented in [4, 5]. For our discussion, we concentrate on [4] because it makes no assumptions on the probability distributions of the DCT coefficients. The scheme of [4] may be interpreted as minimizing (1) as a function of Q_{ij}, with H constant and $T^b_{ij} = 1, \forall i,j,b$. Due to the AC run-length coding, it is difficult to obtain an optimal solution to this problem using classical bit allocation techniques. Therefore, the algorithm in [4] uses a greedy, steepest-descent optimization, starting with a "coarse" Q-matrix and making "finer" one quantizer entry at a time until a desired rate constraint is met. The location (i_o,j_o) of the best quantizer to update (i.e. which of the 64 Q-matrix entries to update) and what integer stepsize $Q_{i_oj_o} \rightarrow q, 1 \leq q < Q_{i_oj_o}$, to use are calculated in a greedy fashion. That is, over all possible AC indices (i,j) and all possible updates $Q_{ij} \rightarrow q$, the set of indices and updates are chosen to maximize $\frac{\Delta D}{\Delta R}$, where ΔD is the image-wide improvement in distortion and ΔR is the image-wide increase in rate accrued by changing the quantizer value.

3.2. Optimal Thresholding

A fast thresholding "kernel" has been developed in [3] that is complementary to quantizer selection. For a fixed Q-matrix and Huffman table, this kernel finds the optimal set of AC coefficients to threshold. The key to the algorithm is using Lagrange multipliers to convert the constrained problem in (1) to the following unconstrained minimization

$$\min_{T^b_{ij},Q_{ij},H} [J(\lambda) = D(T^b_{ij},Q_{ij}) + \lambda * R(T^b_{ij},Q_{ij},H)] \quad (2)$$

where as before, $i,j = 0,\ldots,7$ & $b = 1,\ldots,\frac{N^2}{64}$. $J(\lambda)$ is the Lagrangian cost for fixed quality factor λ. With appropriate choice of λ, achieved by iteration, solutions to (2) provide solutions to (1). The power of Lagrange multipliers is that, since the AC coefficients are coded independently from block to block, the problem of minimizing $D + \lambda R$ for the entire image can be solved by independently minimizing $D + \lambda R$ for individual blocks, with each block-sized minimization quickly solved via a Dynamic Programming-based algorithm. Intuitively, this Dynamic Programming algorithm works in recursive fashion, beginning by calculating the Lagrangian cost of transmitting only the DC for the block, and then recursively computing the incremental costs of transmitting the AC coefficients (in zig-zag scan order), the recursions being performed through fast pruning techniques. Once the kernel finds the best coefficient to end the scan (hence thresholding all further coefficients), it "backtracks" to find the optimal set of predecessors to keep.

Note that in [3], the thresholding kernel is applied at various scales of the "example" JPEG Q-matrix in order to find the Q-matrix and set of thresholded coefficients with lowest Lagrangian cost $D+\lambda R$. This is equivalent to solving (1) with a fixed Huffman table and the added constraint that $Q = k * Q_o$, where Q_o is the "example" JPEG matrix.

4. QUANTIZER SELECTION VIA LAGRANGE MULTIPLIERS

In order to integrate quantizer selection with the thresholding kernel, we attack the quantizer selection problem using Lagrange multipliers. Thus, we seek to minimize (2) as a function of the Q-matrix entries Q_{ij} for fixed Huffman table H. Functional dependency on the thresholding parameters T^b_{ij} will be incorporated later, but for now we will ignore the thresholding parameters (or assume as before $T^b_{ij} = 1, \forall i,j,b$). Conceptually this is equivalent to the approach in [4], with a few implementational differences.

Taking an approach similar to [4] we iteratively update the quantizer entries one at a time under the constraint that all other quantizers remain constant, choosing the quantizer update which minimizes $D + \lambda R$. We select a zig-zag scan order that covers all AC entries of the Q-matrix, and then update each Q-matrix entry $Q_{ij} \rightarrow q, 1 \leq q \leq 255$, to minimize $D+\lambda R$ given that all other Q-matrix entries are fixed. This allows 63 Q-matrix updates per scan iteration, with the scans being repeated until a local minimum is reached.

A further implementational point is that for a fixed (i_o,j_o) in the scan order, only the *relative* Lagrangian costs for each possible update $Q_{i_oj_o} \rightarrow q$ require calculation, since all other Q-matrix entries are assumed constant. For example, in a given block b_o the choice of a single quantizer affects the coding of at most two coefficients, the current coefficient

being quantized, $c^{b_o}_{i_o j_o}$, and the next nonzero coefficient in the zig-zag scan. The effect on coding the latter is simply through a change in run-length if the current coefficient is quantized to zero. This limited dependency allows efficient techniques for comparing the Lagrangian cost of different quantizer stepsizes. For example, within a block, the cost is the same for all stepsizes which quantize the current coefficient to zero (i.e. all q's such that $q > 2|c^{b_o}_{i_o j_o}|$), implying computational reduction since DCT blocks of typical images are sparse.

5. JOINT OPTIMIZATION

Given the constraint of producing a JPEG-decodable bit stream, one may ask what is the absolute best performance that can be achieved. Because the space of encoder optimization parameters is finite, once can search for the optimal encoder using integer programming. However, an exhaustive search is not only computationally absurd, but lacking in insight. We therefore assimilate the contributions of previous work in posing a tractable joint optimization algorithm. Recall that the thresholding kernel finds the best set of DCT coefficients to threshold in order to minimize the Lagrangian cost, *given that the Huffman table and Q-matrix are fixed.* Obviously, for joint optimization, the best possible Q-matrix to use with thresholding should be tuned to minimize the same Lagrangian cost function. If this were not the case, then by perturbing the quantizers one would achieve superior rate-distortion performance. Moreover, the optimization to tune the quantizers should be with respect to the coefficients *that are to be transmitted* , rather than the unthresholded DCT proper, requiring a generalization of the quantizer selection algorithm.

The original quantizer selection takes as input the DCT coefficients c^b_{ij}, and finds an optimized Q-matrix, under the assumption the Huffman table is fixed. In order to optimize the Q-matrix for quantizing the coefficients to be transmitted, we modify the input to the algorithm. Instead of the actual coefficients c^b_{ij}, the input becomes the set of thresholded coefficients $\hat{c}^b_{ij}$, where $\hat{c}^b_{ij} = T^b_{ij} c^b_{ij}$. This relation formalizes the quantizer selection dependency on the thresholding parameters T^b_{ij}. Under this interpretation quantizer selection may be viewed as *minimizing the Lagrangian cost (2) as a function of Q_{ij} for fixed thresholding parameters T^b_{ij} and Huffman table H.*

Note also that customizing the Huffman table according to the statistics of the currently thresholded, quantized image, not only decreases the current coded bit cost, but also provides more accurate entropy/rate estimates for future thresholding and quantization steps. Therefore, the Huffman table should be updated to reflect changes in statistics from thresholding and quantization. The above ideas intuitively motivate the following algorithm, which at each step decreases the Lagrangian cost. Note that a more complete derivation for the joint optimization algorithm using concepts from Entropy Constrained Vector Quantization is provided in [7].

5.1. Algorithm

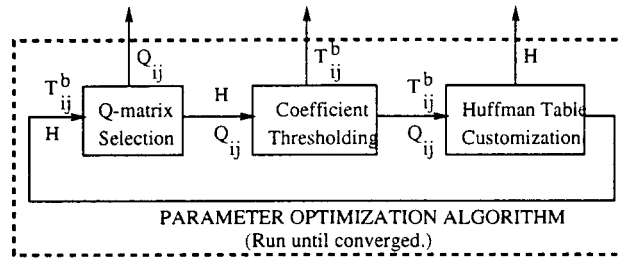

Figure 2: Encoder parameter optimization. Starting from some initial conditions on Q_{ij}, H, and T^b_{ij}, the algorithm iteratively updates each parameter given that the other parameters are fixed. The algorithm continues until the net improvement in Lagrangian cost J is minimal, at which point the parameters are sent.

1. Initialize Huffman table H, quantizers Q_{ij}, T^b_{ij} and convergence factor $\kappa(\lambda)$.
2. Quantizer selection - minimize Lagrangian cost (2) as a function of Q_{ij}, with T^b_{ij} and H constant.
3. Threshold - minimize Lagrangian cost (2) as a function of T^b_{ij}, with Q_{ij} and H constant.
4. Customize Huffman table H - minimize Lagrangian cost (2) as a function of H, with T^b_{ij} and Q_{ij} constant.
5. Return to 2 until convergence criterion is met $(\Delta J(\lambda) < \kappa(\lambda))$.

The algorithm guarantees convergence, since the parameter searches are within a finite space and each operation decreases the Lagrangian cost. Also, to obtain solutions to the rate-constrained problem (1), a root-solver requiring minimizations at different λ's may be used to find λ such that the rate constraint is best satisfied. To enhance speed, a reasonable assumption that has been verified in our experiments is that the Q-matrix is a monotonic function of λ. This allows solutions from previous λ's to tightly bound the quantizer search for future iterations.

Distortion metric: The preceding algorithm can be used to optimize a flexible range of distortion metrics. Although for reference we quote results using the the conventional mean-squared error (L_2) metric, this metric may not optimize perceived picture quality. For example, in adaptive quantizer selection at lower rates, the L_2 metric results in finer coding of less high-frequency information at the cost of annoying low-frequency blocking artifacts. For best results, a complex metric on different parameters such as the display conditions, spatial frequency of the error, and error masking [8] should be used.

An in-depth discussion of perceptual weightings is beyond the scope of this paper. We use a very simple model, noting from [8] that, at least for a simplistic model, error in different frequencies should be normalized by dividing by the perceptibility of the error in those frequencies. Because the "example" Q-matrix provides an estimate of perceptibility thresholds (though this perceptibility varies depending on viewing conditions), one may use it to weight the errors in different frequencies. The low bit rate example we present will incorporate these weightings, with a decrease

in the normalizing factor for the four lowest frequency coefficients by two, to reflect the perceptibility of "blocking" artifacts.

6. RESULTS

The following tables reflect the PSNR improvements due to thresholding, adaptive quantization, and joint optimization, with a customized Huffman table being used for all cases.

JPEG coding of 512x512 **Lenna** PSNR (dB)				
Rate (bpp)	Custom Baseline	Adaptive Quant.	Adaptive Thresh	Joint Opt.
.25	31.6	31.9	32.1	32.3
.5	34.9	35.5	35.3	35.9
.75	36.6	37.5	37.2	38.1
1.0	37.9	38.8	38.4	39.6

JPEG coding of 512x512 **Barbara** PSNR (dB)				
Rate (bpp)	Custom Baseline	Adaptive Quant.	Adaptive Thresh	Joint Opt.
.25	25.2	26.0	25.9	26.7
.50	28.3	30.1	29.3	30.6
.75	31.0	33.0	31.9	33.6
1.00	33.1	35.2	34.1	35.9

Figure 3: JPEG w/ custom Huffman table (0.23 bpp)

Two major factors explain why the PSNR coding gains of adaptive quantizer selection and thresholding are nearly additive. First, their operations are nearly orthogonal, since adaptive quantizer selection exploits global image statistics, while thresholding exploits local statistics. Furthermore, as noted in [4], adaptive quantizer selection results in finer quantization of high frequency DCT coefficients than in baseline JPEG [4]. Although optimal on a global scale, this operation may do poorly on a local scale, since certain higher frequency coefficients become extremely costly to run-length encode depending upon the statistics of their block. Thresholding exploits these local statistics by removing these coefficients to help provide the substantial overall PSNR gain.

Figure 4: R-D optimized JPEG (0.23 bpp)

We present a subjective example of the "Lenna" image coded at .23 (bpp) using standard JPEG with customized Huffman table (Figure 2) and using our adaptive coder with a weighted error criterion (Figure 3). Retaining full JPEG compatibility, the adapted version has significantly reduced blockiness, since quantizer selection favors the DC component more than "scaling" and thresholding of certain costly (in an R-D sense) high-frequency information allows finer coding of low-frequency information.

7. REFERENCES

[1] G. K. Wallace, "The JPEG still-picture compression standard," *Communications of the ACM*, vol. 34, pp. 30–44, April 1991.

[2] W. Pennebaker and J. Mitchell, *JPEG Still Image Data Compression Standard*. New York: Van Nostrand Reinhold, first ed., 1993.

[3] K. Ramchandran and M. Vetterli, "Rate–distortion optimal fast thresholding with complete JPEG/MPEG decoder compatability," *IEEE Trans. on Image Processing special issue on image compression*, vol. 3, pp. 700–704, September 1994.

[4] S. Wu and A. Gersho, "Rate-constrained picture-adaptive quantization for JPEG baseline coders," in *Proc. Inter. Conf. Acoustics, Speech and Signal Processing*, vol. 5, pp. 389–392, April 1993.

[5] A. Hung and T. Meng, "Optimal quantizer step sizes for transform coders," in *Proc. Inter. Conf. Acoustics, Speech and Signal Processing*, pp. 2621–2624, April 1991.

[6] J. Shapiro, "Embedded image coding using zerotrees of wavelet coeffcients," *IEEE Trans. on Signal Processing*, vol. 41, pp. 3445–3462, December 1993.

[7] M. Crouse and K. Ramchandran, "Entropy-constrained quantization framework for JPEG encoder-optimization." To be submitted IEEE Trans. on Image Processing.

[8] J. A. Solomon, A. B. Watson, and A. Ahumada, "Visibility of DCT basis functions: Effects of contrast masking," in *Proc. Data Compression Conference*, pp. 361–370, March 1994.