Journal of Electronic Imaging 4(2), 144–150 (April 1995).

Design of image-adaptive quantization tables for JPEG

Hei Tao Fung
Kevin J. Parker
University of Rochester
Department of Electrical Engineering
Rochester, New York 14627
E-mail: parker@ee.rochester.edu

Abstract. *The rate-distortion trade-off in the discrete cosine transform-based coding scheme in ISO/JPEG is determined by the quantization table. To permit a different quality to be selected by a user, a common practice is to scale the standard quantization tables that have been empirically determined from psychovisual experiments. In this paper, an algorithm is presented to generate a quantization table that is optimized for a given image and for a given distortion. The computational complexity of this algorithm is reduced compared to other techniques. The optimized, image-adaptive quantization table typically yields an improvement of 15% to 20% in bit rate compared to the use of standard, scaled quantization tables. Once an optimized quantization table has been generated for a specific image, it can also be applied to other images with similar content with a small sacrifice in bit rate.*

1 Introduction

The ISO/JPEG standard is an international standard for continuous-tone still-image compression.[1] Its goal is to support a wide range of applications. Its discrete cosine transform (DCT)-based encoding steps include forward DCT (FDCT), quantization, and entropy coding. The decoding steps consist of entropy decoding, dequantization, and inverse DCT (IDCT). The image to be encoded is first divided into 8×8 blocks. On each block a FDCT is performed. The 64 DCT coefficients are uniformly quantized in accordance with a quantization table. An example of a quantization table is shown in Fig. 1. The purpose of quantization is to achieve a higher compression ratio (or lower bit rate) by representing the coefficients with lesser precision. Because quantization introduces distortion in the decoded image, naturally, there is a rate-distortion trade-off. This trade-off is determined by the quantization table. Each element in the quantization table represents the step size of the uniform quantizer for its corresponding coefficient. The index values to be coded are calculated as

$$Z_n(k) = \mathrm{NINT}[z_n(k)/Q(k)] \quad \text{for } k = 0, \ldots, 63 , \tag{1}$$

where $z_n(k)$ is the DCT coefficient, $Q(k)$ is the corresponding value in the quantization table, NINT is the nearest integer function, and n the block index. The DCT coefficients are reordered in a zigzag manner. The dc coefficients are coded using differential pulse code modulation (DPCM). For the ordered ac coefficients, pairs of the runlength of zeros and the magnitude of the following nonzero coefficients are formed. These pairs are entropy coded. At the decoder, the quantization table is either extracted from the data stream or known as a default. After the $Z_n(k)$'s are decoded, the dequantized values of the DCT coefficients are given by

$$z'_n(k) = Z_n(k) \times Q(k) \quad \text{for } k = 0, \ldots, 63 . \tag{2}$$

Finally, after reordering, an IDCT is performed to reconstruct the image block.

Psychovisual experiments described in Ref. 2 have led to a set of quantization tables, which are documented in Ref. 1 (Fig. 1). These tables are based on the visibility of the 8×8

Paper 94-023 received July 8, 1994; revised manuscript received Nov. 14, 1994; accepted for publication Nov. 29, 1994.
1017-9909/95/$6.00 © 1995 SPIE and IS&T

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

Fig. 1 The quantization table for the luminance component documented in the ISO/JPEG draft international standard.

DCT basis functions measured under certain conditions. To achieve different bit rates and fidelity, a popular practice is to scale the tables according to a scaling curve. The quantization tables generated by this practice are by no means optimal. By "optimal," we mean that for a given image and a given distortion value under an image quality measure, the bit rate is minimized, or for a given image and a given bit rate, the distortion value is minimized.

Theoretically speaking, one can determine the optimal quantization table by exhaustive search over the set of possible quantization tables. The values of the quantization table elements are specified to lie in the range [1,255]. Since there is one uniform quantizer for each DCT coefficient, there are 255^{64} possible quantization tables. For a given image and for each quantization table, a specific distortion value and bit rate can be determined. If the pairs of the distortion values and the corresponding bit rates are plotted to form a 2-D graph, the optimal quantization table can be determined by looking for the quantization table corresponding to the lowest point for a given bit rate or by looking for the quantization table corresponding to the left-most point for a given distortion value (Fig. 2). Obviously, this process of an exhausted search over all possible $Q(k)$'s is impractical.

In Ref. 3, an algorithm was presented to generate optimized quantization tables. Starting from an initial quantization table of very coarse quantizers, the algorithm decreases the step sizes of the quantizers one at a time until a given bit rate is reached. At each time, the element of the quantization table to be updated is chosen to be the one that gives the largest ratio of decrease in distortion to increase in bit rate. The quantization table found from this algorithm is again suboptimal, yet it gives a larger peak SNR than that resulting from the quantization table determined by psychovisual experiments. However, its main disadvantage is its computational complexity. With the DC quantizer step size fixed, the algorithm requires $\sum_{k=1}^{63}[Q(k)-1]$ evaluations of the distortion value and bit rate at each iteration.

In this paper, we present an algorithm for optimizing an image-adaptive quantization table that is aimed at reducing the computational complexity. First, we allow the quantization table element to go up or down depending on whether a coarser or a finer quantization table is needed to achieve the target distortion value at each iteration. We also limit the magnitude of change in the quantization table element values so that our method requires only 63 evaluations of the distortion value and bit rate at each iteration. Second, we propose the use of a novel entropy estimator to estimate the bit rate. The statistics required to calculate the entropy can be more easily updated than the statistics required to calculate the bit

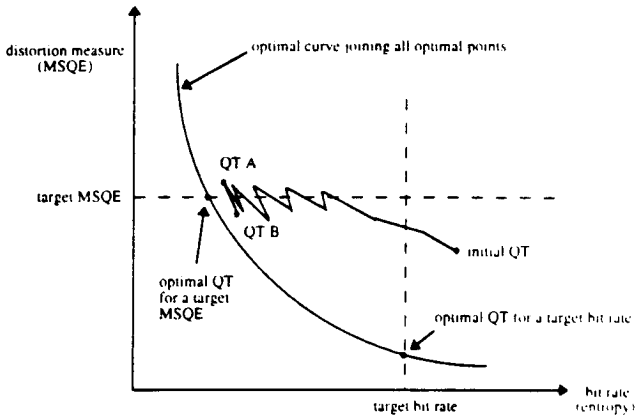


Fig. 2 A pictorial representation of the procedure to find the optimal quantization table for a target distortion value (or a target MSQE value). The coordinates of each point on the graph represent the distortion value (or MSQE value) and bit rate (or defined entropy) associated with a corresponding quantization table. At each iteration, we vary the quantization table element such that the point is moved up the gentlest slope if it is below the target line or down the steepest slope if it is above the target line. The procedure terminates when the downslope of one quantization table (QT A) equals the upslope of another quantization table (QT B). The optimized quantization table is QT A or QT B.

rate. As a result, this further reduces the computation complexity.

The algorithm presented in this paper is efficient in obtaining an optimized quantization table for a given image so that the image can be efficiently encoded using JPEG for a given distortion value. This paper also shows that the quantization table determined for a particular image is good for the set of images with similar content, i.e., a whole set of similar images can be efficiently encoded using the same quantization table. This observation implies that the algorithm may be run just once for the whole set or class of images. This fact enhances the value of the algorithm.

2 Algorithm

As mentioned earlier, for a given image, the distortion value and bit rate are associated with a quantization table. Points with the coordinates representing the distortion value and bit rate pairs can be plotted to form a 2-D graph. We illustrate our algorithm pictorially with this graph. Before we discuss the procedure, however, we first discuss the distortion measure we used in our experiment and the use of an entropy estimator to be defined as a bit rate estimate.

2.1 *Distortion Measure*

We use the mean square quantization error (MSQE) of the DCT coefficients as the distortion measure. MSQE is given by

$$\text{MSQE} = \sum_{n=0}^{N-1}\sum_{k=0}^{63}[z'_n(k)-z_n(k)]^2 \ , \tag{3}$$

where N is the number of blocks. The two reasons for using the MSQE are, first, MSQE can be readily calculated from the unquantized and the dequantized DCT coefficients. The alternative, which is to use the pixel error as the distortion measure, is less attractive because it is computationally ex-

pensive to convert the DCT coefficients to pixel values at each iteration. Second, the MSQE can be weighted by a human visual system (HVS) response function without much complication so that the weighted MSQE can correspond more to the perceptual quality. Details can be found in Refs. 4, 5, and 6. However, in our experiments, we have used unweighted MSQE for simplicity.

2.2 *Bit Rate Measure*

Determining the bit rate of an encoded image generally requires completing the whole encoding process. This would be computationally expensive in an iterative procedure. Instead of measuring the exact bit rate, we estimate it by an entropy of the quantized values. Empirically determined, the following definition of entropy reflects the bit rate well:

$$E=\sum_{x=0}^{M}\sum_{s=0}^{1}[-P(x)P(s|x)\log_2 P(s|x)]\ , \tag{4}$$

where $P(s|x)$ is the conditional probability of $s=S_n(k)$ given $x=X_n(k)$, i.e.,

$$P(s|x)=\frac{\sum_{n=0}^{N-1}\sum_{k=1}^{63}\delta[s-S_n(k)]\cdot\delta[x-X_n(k)]}{\sum_{n=0}^{N-1}\sum_{k=1}^{63}\delta[x-X_n(k)]}\ ; \tag{5}$$

$P(x)$ is the probability of $x=X_n(k)$, i.e.,

$$P(x)=\sum_{n=0}^{N-1}\sum_{k=1}^{63}\delta[x-X_n(k)]/(63N)\ ; \tag{6}$$

$\delta(\cdot)$ is the Dirac delta function; $S_n(k)$ is the order of magnitude of the quantized values, i.e.,

$$S_n(k)=\begin{cases}t,2^{t-1}\le|Z_n(k)|<2^t \text{ and } Z_n(k)\neq 0\\ 0,Z_n(k)=0\end{cases} \tag{7}$$

$$\text{for } k=1,\ \ldots,\ 63\ ;$$

$X_n(k)$ determines whether the previous quantized value is zero or nonzero.

$$X_n(k)=\begin{cases}1,S_n(k-1)=0\\ 0,S_n(k-1)\neq 0\end{cases}\quad \text{for } k=2,\ \ldots,\ 63\ : \tag{8}$$

$$X_n(1)=1\ :$$

N is the number of 8×8 blocks; and M is the maximum t in Eq. (7).

The key point in the empirical estimate of the entropy is that the important role of zeros and runs of zeros is captured in the simple definition of $X_n(k)$. The entropy value versus the actual bit rate is shown in Fig. 3. An expanded view is given in Fig. 4. A perfect bit rate measure would align all the curves in Fig. 4 on a straight line. Figures 3 and 4 thus illustrate the entropy's effectiveness as a bit rate measure, and based on the almost linear relationship, we assume

$$\Delta B=C\times\Delta E\ , \tag{9}$$

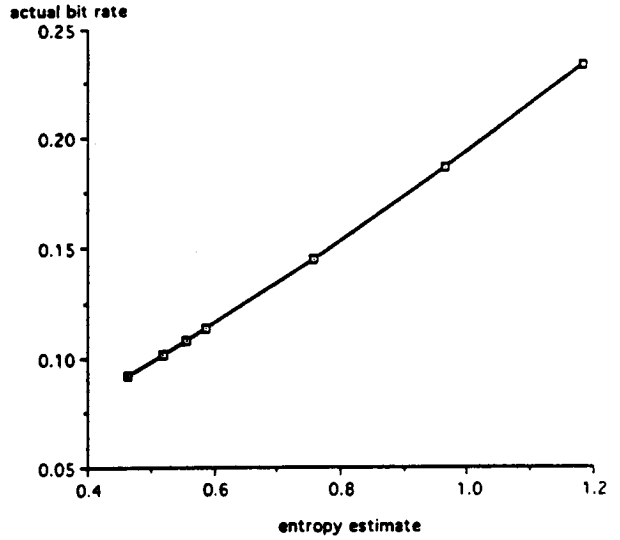


Fig. 3 A typical curve of bit rate of an encoded image vs. our entropy estimate using the quantized DCT values. It shows an almost linear relationship.

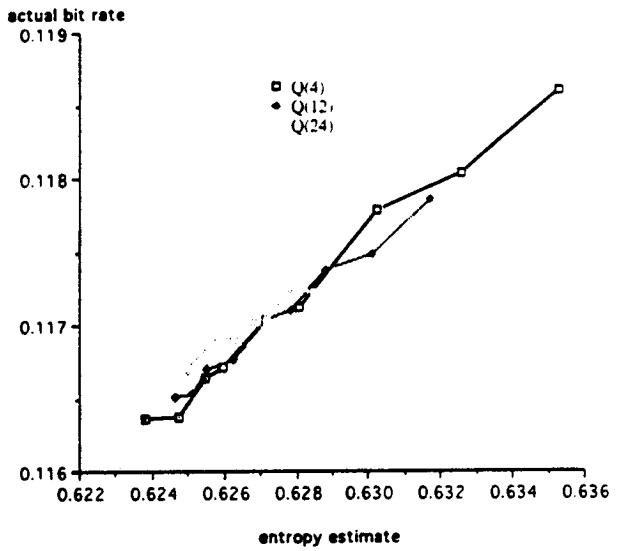


Fig. 4 An expanded view of typical curves of bit rate versus our entropy estimate using the quantized DCT values. For each curve, only one $Q(k)$ varies: $Q(4)$ varies from 8 to 16, $Q(12)$ from 12 to 20, and $Q(24)$ from 25 to 33. The relationships are complicated but can be reasonably approximated by a linear equation.

where ΔE denotes the change of the entropy of the quantizer output values, C is a constant, and ΔB is the change of bit rate. Our definition of entropy captures the role of long run-lengths of zeros in bit rate reduction using entropy encoding. The advantage of using this model as a bit rate measure is that the bit rate is estimated without actually going through the entropy coding, and calculating ΔE requires fewer operations. This is because the statistics required to calculate the entropy can be more easily updated than the statistics required to calculate the true bit rate because ΔE involves only second-order statistics, with no encoding steps.

2.3 Procedure

The procedure can be depicted as shown in Fig. 2. Qualitatively, we start with a point on the graph corresponding to an initial quantization table. The coordinates of the point represent the values of the MSQE and the entropy. Our algorithm changes the quantization table element one at a time such that the next point, corresponding to another quantization table, moves down the most negative slope or up the least negative slope, wiggling about the target MSQE value. With iterations, convergence is achieved when two points oscillate from one to the other as the downslope and upslope become essentially the same. The main steps of the algorithm are as follows:

1. Specify an initial quantization table with $Q(0) = 16$ and $Q(k)$ in $[1, \ldots, 255]$ for $k = 1, \ldots, 63$.
2. Calculate the MSQE. If the calculated MSQE is larger than the target MSQE, for each k from $\{1, \ldots, 63\}$, find the change in the defined entropy value, ΔE, and the change in the MSQE, ΔMSQE. Select the k that gives the smallest $-\Delta E/\Delta$MSQE ratio if $Q(k)$ is replaced by $Q(k) - S$, where S is a positive integer. Call that k k_{min} and the corresponding ratio R_{min}. Replace $Q(k_{min})$ by $Q(k_{min}) - S$.
3. Calculate the MSQE. If the calculated MSQE is smaller than or equal to the target MSQE, for each k from $\{1, \ldots, 63\}$, find ΔE and ΔMSQE. Select the k that gives the largest $-\Delta E/$MSQE ratio if $Q(k)$ is replaced by $Q(k) + S$. Call that k k_{max} and the corresponding ratio R_{max}. Replace $Q(k_{max})$ by $Q(k_{max}) + S$.
4. Repeat steps 2, 3, and 4 until $R_{max} \le R_{min}$.

The quantization table element for dc coefficients is set to 16 for simplicity. This corresponds to 7-bit precision. It can be specified to other values. The greedy bit allocation technique can be applied well to obtain the initial quantization table. It assigns more bits to represent the DCT coefficients with a greater variance over the ensemble of blocks and fewer bits to represent those with a smaller variance. The variances of the 63 ac coefficients, σ_k^2, are first found. The geometric mean ρ is given by

$$\rho = (\Pi \sigma_k^2)^{1/63} \quad \text{for } k = 1, \ldots, 63 \; . \tag{10}$$

Let B denote the total number of bits assigned to all 63 quantizers, b the average number of bits for each quantizer, and $Q(k)$ the quantizer step size for the k'th set of ac coefficients. Then the results are

$$b = B/63 \; . \tag{11}$$

$$b_k = b + (1/2)[\log_2(\sigma_k^2/\rho)] \; . \tag{12}$$

$$Q(k) = 8 \times 27/(2b_k - 1) \; . \tag{13}$$

and the $Q(k)$'s have to be clipped to fall between $[1,255]$.

The S is the magnitude of change in the quantizer step size at each iteration. Usually, the change in quantizer step size for small k's can more significantly affect the ratio than that for large k's. This is because the energy is concentrated in the low frequencies because it is the energy compaction property of the FDCT. Therefore, S can be made small for small k's and large for large k's. We may also vary S, using large S for the early stage of iteration and small S at the later stage, so that the convergence is approached faster at the early stage yet without losing precision at the later stage.

Convergence to oscillation is guaranteed on the condition that the current quantization table is allowed to swing back to its most recent parent. If this condition is not allowed, the algorithm terminates with $R_{max} < R_{min}$. To improve the rate of convergence, instead of changing one element in the quantization table at a time, we may change several elements after searching the k's for the best ratios.

The algorithm guarantees convergence independent of the initial quantization table, as opposed to the strict requirement on the initial quantization table in Ref. 3. We, therefore, may wisely pick an initial quantization table to reduce the number of iterations needed. Note also that this algorithm can be modified to find the optimum quantization table for a specified bit rate instead of specifying a given distortion.

3 Simulation Results

To present the results of our experiments, we use two test images: "Lena" and a text image. Both are 8-bit 256×256 images. We set S to 7, 3, and 1 sequentially for all k's so that computation is cut down at the early stage, and the process is refined at the later stage.

The gain from using the optimized quantization tables over using the scaled, standard tables is shown in Figs. 5 and 6. Figures 5 and 6 show the mean square pixel error (MSE) versus bit rate graphs for "Lena" and the text image, respectively. There are two sets of points for comparison: one for the quantization tables optimized through the algorithm at different target MSQEs, and the other for the quantization tables from scaling the quantization table in Fig. 1. On these graphs showing the final results, the MSE is used instead of the MSQE because the MSE can be readily converted to the SNR. Also the actual bit rate is used instead of the entropy value. Both Figs. 5 and 6 show that a reduction of 14% to 20% in bit rate can be achieved for a given MSE.

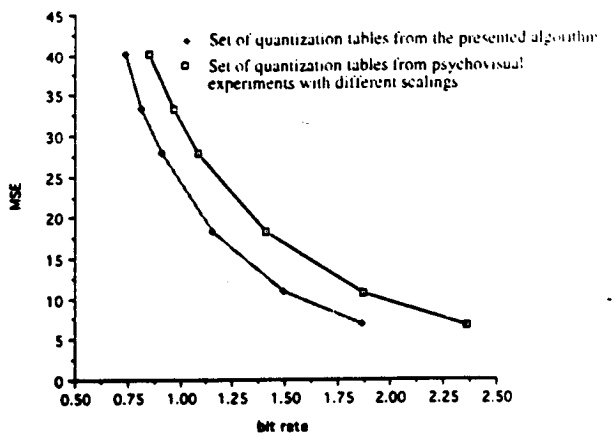


Fig. 5 MSE versus bit rate curves for "Lena." The points are determined using the set of optimized quantization tables obtained from the algorithm and the set of scaled, standard quantization tables. There is about 13% to 21% reduction in bit rate using the optimized quantization tables at the same distortion value.

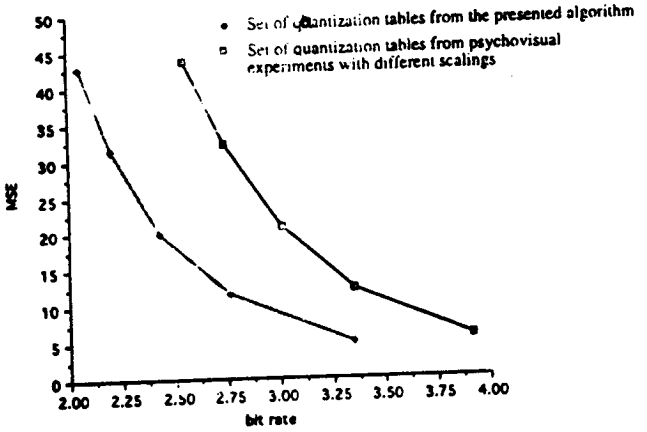


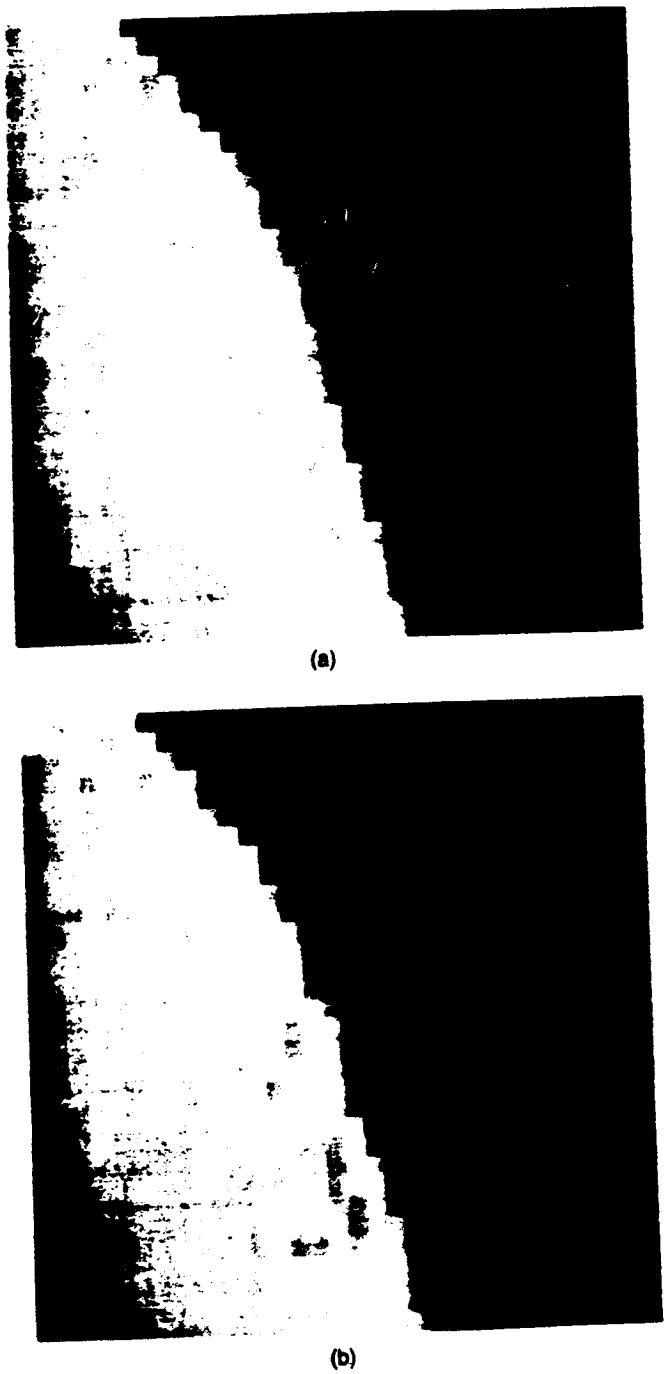
Fig. 6 MSE versus bit rate curves for a text image. The points are determined using the set of optimized quantization tables obtained from the algorithm and the set of scaled, standard quantization tables. There is about a 14% to 20% reduction in bit rate using the optimized quantization tables at the same distortion value.

The quality of the reconstructed images using the optimized quantization tables and the scaled, standard tables are compared in Figs. 7 and 8. Figures 7 and 8 show the expanded portions of the reconstructed images. For the same bit rates (about 1.1 bits/pixel for "Lena," and about 2.7 bits/pixel for the text image), the reconstructed images using quantization tables from our algorithm are less noisy than those using the standard, scaled quantization tables.

The effect of the initial quantization table on the final results is illustrated in Fig. 9, which shows the paths leading toward the local minima starting from different initial quantization tables. Depending on the initial quantization table, the final quantization table may terminate at a different local minimum. The greatest difference in bit rate for the same distortion in Fig. 9 is about 3.5%. This demonstrates that the result of convergence is not sensitive to the initial quantization table.

The algorithm currently takes about 13 minutes on a Sparc 10 workstation to obtain an optimized quantization table. Note also that the optimized quantization tables (represented in Figs. 5 and 6), produced by our algorithm for different target MSQEs, are not simply scaled versions of each other. The ratio of coefficients of any two tables varies significantly over k. This fact underscores the value of an adaptive algorithm.

The optimized quantization table for a particular image can be applied well to a set of images with similar content, whose spectral characteristics are more or less the same. We used six medical ultrasound images from two different B-scan imaging systems. One image was arbitrarily chosen as the "training" image. The optimized quantization table for the "training" image was obtained, with the MSE equal to 6.158 and 1.952 bits/pixel. This quantization table was applied to the other five images to obtain pairs of MSE values and bit rates (Fig. 10). Then optimized quantization tables were obtained for each of the five images such that the resulting MSE values were matched to the previous ones. In this way, the bit rates associated with the individually optimized quantization tables can be compared with the bit rates associated with the quantization table optimized for the one image (Fig. 10). It is shown that there is only about a few



(a)

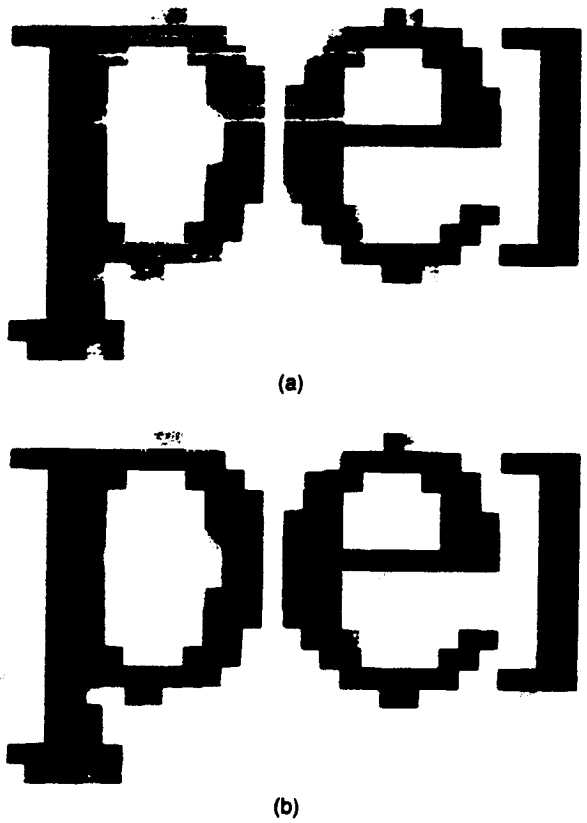
(b)

Fig. 7 Two blocks of 32×32 pixels showing part of the reconstructed "Lena." Both images are compressed at about 1.1 bits/pixel: (a) The image is obtained using an optimized table, and (b) a scaled, standard table is used.

percent increase in bit rate in the case of using one optimized table from a "training" image on the whole set of images. This implies that the optimization algorithm may be run just once. The optimized quantization table can then be used on the whole set of similar images with a little sacrifice in bit rate. This fact enhances the value of the algorithm.

4 Conclusion

An algorithm is presented to design a quantization table for a given image to be used in the ISO/JPEG DCT-based coders.



(a) (b)

Fig. 8 Two blocks of 32 × 32 pixels showing part of the reconstructed text images. Both images are compressed at about 2.7 bits/pixel: (a) The image is obtained using an optimized table, and (b) a scaled, standard table is used.

A reduction of 15% to 20% in the bit rate resulting from using a scaled version of the standard quantization table[1] for a given distortion is typical. The merit of the algorithm lies in the reduction of the computational complexity. Furthermore, the optimized quantization table for a particular image is found to be applicable to other images with similar content with a little sacrifice in bit rate. This fact implies that the algorithm may be run just once for a set of similar images.

Acknowledgments

This work was supported by the NSF/NYS/Center for Electronic Imaging Systems. Support from Eastman Kodak and Xerox is gratefully acknowledged. The authors are indebted to Dr. Majid Rabbani and Dr. John Hamilton for their suggestions and insights.

References

1. *Digital Compression and Coding of Continuous-tone Still Images, Part I, Requirements and Guidelines*, ISO/IEC JTC1 Draft International Standard 10918-1 (1991).
2. H. Lohscheller, "Subjectively adapted image communication system," *IEEE Trans. Commun.* **COM-32**(12), 1316–1322 (1984).
3. S. Wu and A. Gersho, "Rate-constrained picture-adaptive quantization for JPEG baseline coders," *Proc. ICASSP '93*, pp. 389–392, IEEE, Piscataway, NJ (1993).
4. N. B. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.* **COM-33**(6), 551–557 (1985).
5. A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression*, Plenum Press, New York (1988).

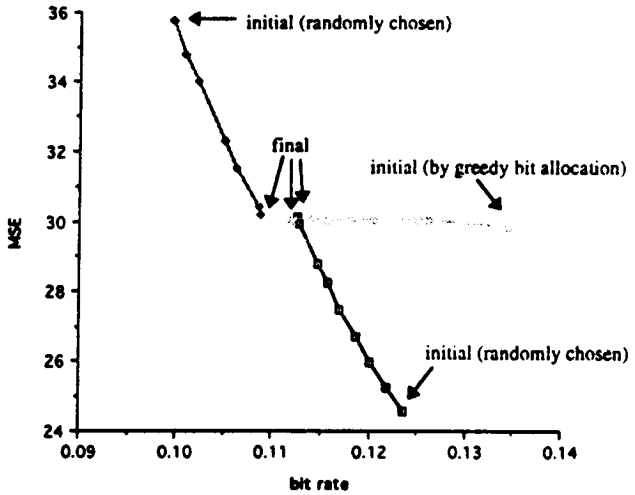


Fig. 9 The convergence of the algorithm to a specified MSE (= 30) from different initial states. The algorithm terminates at local minima. The lines join some of the intermediate states, skipping the ones between. There is about a 3.5% difference in bit rates for the final states, showing the small degree of sensitivity of the final state to the initial state.

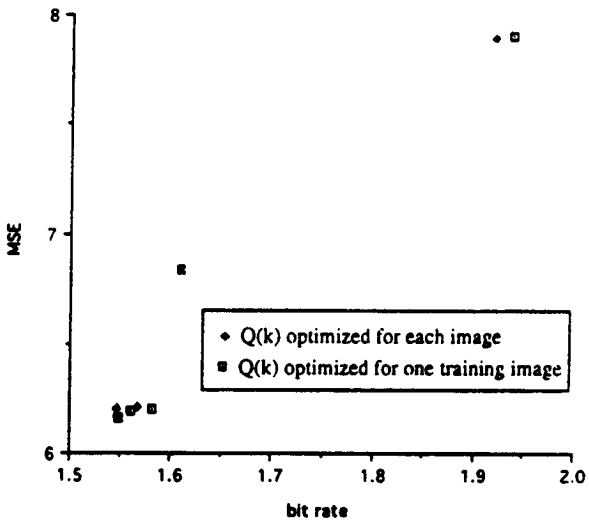


Fig. 10 The MSE versus bit rate graph for a set of five ultrasound images. A set of points is obtained from applying the optimized quantization table for another ultrasound image to the five images. The other set is obtained from using the quantization tables optimized for each of the five images. The MSE values are deliberately matched for comparison of bit rates. The first set of points has a few percent larger bit rates. This shows that an optimized quantization table for an image can be applied to other similar images with a little sacrifice in bit rate.

6. D. L. McLaren and D. T. Nguyen, "Removal of subjective redundancy from DCT-coded images," *IEEE Proc.* **138**(5), 345–350 (1991).

Hei Tao Fung received his BS and MS degrees in electrical engineering from the University of Rochester, New York, in 1992. He has received the Genesee Scholarship and the Chu Foundation Scholarship for four years since 1988. His interests include image processing and integrated circuit design. Mr. Fung is a member of Tau Beta Pi and Phi Beta Kappa. He is also a student member of the IEEE and the IS&T.

Kevin J. Parker received the BS degree in engineering science, summa cum laude, from the State University of New York at Buffalo in 1976. His graduate work in electrical engineering was done at the Massachusetts Institute of Technology, with MS and PhD degrees received in 1978 and 1981. From 1981 to 1985 he was an assistant professor of electrical engineering at the University of Rochester; currently he holds the title of professor of electrical engineering and radiology. Dr. Parker has received awards from the National Institute of General Medical Sciences (1979), the Lilly Teaching Endowment (1982), the IBM Supercomputing Competition (1989), and the World Federation of Ultrasound in Medicine and Biology (1991). He is a member of the IEEE Sonics and Ultrasonics Symposium Technical Committee and serves as reviewer and consultant for a number of journals and institutions. He is also a fellow of the IEEE and the American Institute of Ultrasound in Medicine. Dr. Parker's research interests are in medical imaging, linear and nonlinear acoustics, and digital halftoning.