# Lossy Compression of Grayscale Document Images by Adaptive-Offset Quantization

Kris Popat

Xerox Palo Alto Research Center, Palo Alto, California, USA

## ABSTRACT

This paper describes an adaptive-offset quantization scheme and considers its application to the lossy compression of grayscale document images. The technique involves scalar-quantizing and entropy-coding pixels sequentially, such that the quantizer's offset is always chosen to minimize the expected number of bits emitted for each pixel, where the expectation is based on the predictive distribution used for entropy coding. To accomplish this, information is fed back from the entropy coder's statistical modeling unit to the quantizer. This feedback path is absent in traditional compression schemes. Encouraging but preliminary experimental results are presented comparing the technique with JPEG and with fixed-offset quantization on a scanned grayscale text image.

**Keywords:** document image compression, quantization, entropy coding, arithmetic coding

## 1. INTRODUCTION

Grayscale images of text documents typically contain high-spatial-frequency components that prevent transform-based compression techniques such as JPEG from performing acceptably at high levels of compression. This is not unexpected, as transform-based compression techniques are usually optimized for images that are dominated by low-spatial-frequency components, such as natural scenes.[1] Despite its shortcomings when applied to document images, JPEG is often used for compression when the images have been scanned in grayscale, which makes it an appropriate method to compare against. Compression techniques designed specifically for document images, such as JBIG2,[2] typically require that the text regions be binarized prior to compression. Binarization can adversely affect readability as well as the accuracy of any subsequent optical character recognition (OCR), particularly when the image originates at low to moderate spatial resolution. Binarization prior to compression can therefore be inappropriate in many situations.

We consider an alternative approach that compresses without spatial transformation or binarization, by simply quantizing and entropy-coding each grayscale pixel sequentially in raster order. Elements of the basic technique, shown in Figure 1, were described in previous work[3] in the context of compressing natural scenes. Here we expand on the idea and apply it specifically to document images. In its operating directly on grayscale pixel values and in the type of statistical model it employs, the approach described here resembles the two-stage method described recently,[4] but departs significantly from it in other aspects, such as its tight integration of the quantizer, entropy coder, and statistical model. In particular, in the method proposed here the quantizer cooperates directly with the entropy coder by shifting, on a pixel-by-pixel basis, the entire lattice of decision and reproduction values to minimize the entropy of the specific conditional probability distribution used by the entropy coder in compressing the quantized pixel.

The paper is organized as follows. Section 2 discusses the quantizer, which is the lossy element in the compressor. It explains the motivation behind adapting the offset, and compares the proposed approach with the well-known

Author email address: popat@parc.xerox.com

[To appear in *Proceedings of IS&T/SPIE Electronic Imaging 2001: Document Recognition and Retrieval VIII* January 2001.]

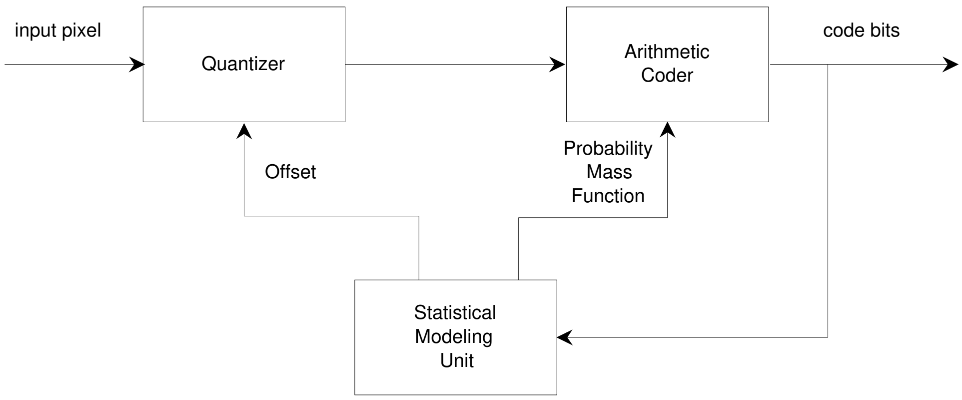


**Figure 1.** Compression system using adaptive-offset quantization.

and superficially similar technique of differential predictive coding. Section 3 discusses the problem of entropy coding; i.e., transforming the quantizer's output into a compact sequence of bits. This requires, in part, the predictive probabilistic modeling of the quantizer's output, on the basis of a subset of previously quantized pixels. Section 4 presents and discusses preliminary experimental results.

## 2. ADAPTIVE-OFFSET QUANTIZATION

The heart of the proposed system is an entropy-coded scalar quantizer that adapts on the basis of information fed back from the statistical modeling unit. Properly speaking, the term *entropy coded quantization* refers to the totality of operations shown in Figure 1 and not just to the quantizer. In this section we will focus on the role of the quantizer as the lossy element, with the understanding that entropy coding will later be performed. We defer detailed discussion of the latter to Section 3. Our quantization terminology generally follows that used in standard texts.[5,6]

### 2.1. Quantizer

In its full generality a *scalar quantizer* is a device that maps a real number to a representative number chosen from a countable set.* Here, we take the set of output values to be finite but large, and we take the quantization region corresponding to each representative output value to be an interval of the real line. Specifically, the non-extremal decision regions of the quantizer are taken to be of equal size $\Delta$, with the reproduction values at their midpoints. The choice of uniform stepsize is motivated both by a desire for simplicity and in recognition of the asymptotic optimality[7] and non-asymptotic near-optimality[8] of uniform-threshold entropy-constrained quantization in a variety of settings. The choice of midpoint reproductions is less sound theoretically, but is made for convenience. As will be clear in what follows, these choices are not essential for the proposed technique to function, but a large performance gain by generalizing on these particular aspects seems unlikely.

To simplify both the design and analysis of the system, the number of quantization regions $N$ is chosen to be sufficiently large to make overload rare, so that the extremal decision regions are used only for exceptional cases involving outliers. The ability to choose a large value of $N$ without directly paying a price in bit-rate is due to the entropy coding of the quantizer's output: for a fixed $\Delta$ and with $N$ sufficiently large to avoid distortion due to overload, increasing $N$ further does not increase entropy, since the probability (and hence contribution to both entropy and distortion) of each additional quantization region becomes negligible. In the present context of compressing grayscale images, the input pixel amplitude range is bounded, so that we can (and do) choose $N$ to prevent overload entirely.

*The textbook by Gersho and Gray[6] provides an excellent introduction to quantization, both scalar and multidimensional.

Besides $\Delta$ and $N$, the remaining quantization parameter is the absolute location or *offset* of the quantizing lattice, which we can specify by a parameter $\phi \in [0, \Delta]$ such that the non-overload input-output relation of the quantizer is

$$q(x) = \phi + \Delta \lfloor \frac{x - \phi}{\Delta} + 1/2 \rfloor \tag{1}$$

where $\lfloor x \rfloor$ denotes the greatest integer not greater than $x$. As two examples, when $\phi$ is fixed at zero, the quantizer is termed *midtread*, and when it is fixed at $\Delta/2$, it is termed *midrise*.[5] However, it is important to note that $\phi$ can range over a continuum of values.

At high rates the precise choice of offset has little effect on rate-versus-distortion performance,[7] but in low-rate entropy coded quantization the offset can be critical. For instance, in the well-studied case where the input distribution is symmetric around a single mode, choosing an offset that results in the mode being well-centered within a quantization region is essential for good performance.[5,9] If the offset is inadvertently chosen at the other extreme, so that the mode is evenly split by a decision boundary, then performance will suffer; in fact in this case the output entropy will be bounded below by one bit per quantized value, irrespective of how large a distortion is deemed tolerable. In the present application, we have no particular reason to believe that the distribution of the quantizer output will be either symmetric or unimodal. In fact, in text images pixels near character edges may be better described by a bimodal distribution. Nevertheless, there is ample reason to expect that the choice of offset, provided that it is allowed to vary from pixel to pixel, will still be critical at low bitrates. The reason is that just as in the symmetric unimodal case, in order for the bitrate to be low, probability mass must be concentrated in relatively few output levels. This in turn requires that modes not be split by decision boundaries when doing so is avoidable. Choosing an appropriate quantization offset for each pixel can provide a means of minimizing the needless splitting of modes. Hence we allow $\phi$ to vary from pixel to pixel, always choosing a value that minimizes quantizer output entropy with respect to the statistical model used by the entropy coder. In practice, this can be accomplished by searching over a suitably fine grid on the interval $[0, \Delta]$. For the results presented later in this paper, $\phi$ is determined for each pixel position by searching over the set $\{0, 1, \ldots, \Delta - 1\}$, choosing a value that minimizes the entropy of the estimated quantizer-output distribution.

The motivation for adaptation of $\phi$ can be clarified by considering the example in Figure 2. By choosing the offset to shift the quantizer lattice in (a) slightly to the left as shown in (b), both the mean-square and maximum distortions are nearly unchanged, but the entropy of the quantizer output is reduced substantially. The effect can be regarded as a generalization of the previously cited phenomenon of midtread quantizers outperforming midrise entropy-coded quantizers on unimodal symmetric distributions.

## 2.2. Relationship to Differential Predictive Coding

In principle, shifting the offset of the quantizer could be effected by subtracting out a value from the quantizer's input, and adding it back in to the quantizer's output. Viewed in this way, the technique resembles differential predictive coding,[5] wherein a predicted value is subtracted from a correlated signal both to whiten it (i.e., remove redundancy) and to reduce its variance, allowing it to be compressed more effectively by a simple quantizer. While the two techniques coincide in structure, they differ in their goals and details of their constituent elements, and these differences play out in their functioning and ultimately performance. The differences are substantial but subtle, and some discussion is appropriate.

The contrast between the two approaches can be made apparent by again considering a case where the predictive distribution for the pixel being quantized is bimodal. Such distributions are not uncommon in grayscale document images, even when conditioned on nearby pixels. The differential predictive coder focuses on the *input* to the quantizer; it typically seeks to minimize variance by subtracting out the estimated mean of the input, under the assumption that the quantizer will have an easier time coding lower-variance values than larger-variance ones. This assumption is typically valid when no entropy coding is used, or when the statistical model used in entropy coding

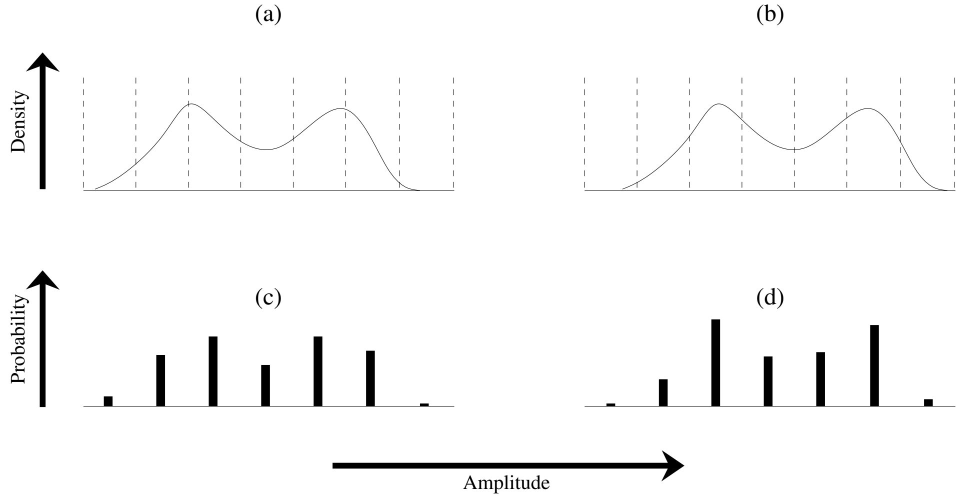


**Figure 2.** Example of an input pixel's estimated probability density relative to the quantization lattice for two different offsets in (a) and (b). The dashed vertical lines indicate the positions of the threshold boundaries between quantization regions. The corresponding probability masses for the quantized pixel are shown in (c) and (d). For a given input density, an offset is sought that minimizes the entropy of the resulting probability mass function. In these graphs, it is evident that (d) has a less uniform distribution of probability masses than (c), a relationship which is reflected in their entropies.

is severely constrained (e.g., a parametric model). When more extensive modeling and coding is being performed, subtracting out the mean as a preprocessing step does not necessarily improve either output entropy or quantization error. Consider the bimodal example. If the modes happen to be spaced apart in such a way that after subtracting out the mean, both are split by decision boundaries, then both quantizer output entropy and mean-square error would be worse than if a value, say, one-half stepsize to either side had been used instead as the "predictor." Traditional differential predictive coding has no way of detecting such conditions or accommodating them.

In contrast, adaptive-offset quantization focuses on the *output* of the quantizer, and makes no attempt to minimize input variance directly. Instead, the goal is to shift the input so that to the extent possible, probability mass is concentrated well within each quantization region instead of near the boundaries. The degree to which this is possible will depend on the specific predictive distribution, on a case-by-case basis. When it is possible, both entropy and mean-square error will be simultaneously improved. When it is not possible, no harm to either will have been done by trying. In most cases, an appropriate choice for the "predictor" value will require numerical search, as opposed to simply computing the mean of the predictive distribution.

Note that we might, in principle, adapt the stepsize as well as the offset by searching for joint values that minimize the entropy of the predictive distribution, under a constraint on the maximum allowable instantaneous error. Doing so would provide an extra degree of freedom that might allow for improved centering of modes within the quantization regions, thereby improving performance. However, the search would become much more complicated, and the variability of stepsize would induce a spatial variability in image quality that would have to be controlled. For these reasons, we keep the stepsize fixed in the present study and adapt only the offset. The ultimate practicality of adaptive-offset quantization, or of distribution-adaptive quantization in general, will depend largely on being able to carry out the search over quantizer parameters efficiently. Investigation of efficient search procedures is left for future work.

# 3. ENTROPY CODER AND STATISTICAL MODELING UNIT

The quantizer reduces the information content of the input, but it does not directly result in a reduction in the volume of data used in representing the image. To achieve compression, we employ a subsequent stage of lossless, variable-rate coding, also known as entropy coding. There are many ways of performing entropy coding. Generic compressors like the *gzip* program[10] are appropriate when little is known in advance about the statistics of the message being compressed. Here, we know much in advance: that the symbols in the message are quantized pixel values from grayscale document images, that they are spatially interdependent, and that vectors made up of pixels before quantization in a local spatial neighborhood can be usefully characterized by a smooth multidimensional probability density function. In principle, a generic compressor could learn all of the relevant statistical properties from the message as the compression proceeds,[11] but the limited learning rate would likely make this effective only for extremely long messages (i.e., extremely large images). It is more appropriate in the present setting to exploit the available prior knowledge directly via an explicit statistical modeling unit. Arithmetic coding[12,13] is a form of entropy coding that allows this in a very natural way.

## 3.1. Arithmetic Coding

The three main requirements on the entropy coder are (1) bit-efficiency, in the sense that the excess number of expected code bits over $-\log_2 P(y|\text{ context })$ when encoding a quantized pixel $y$ should be minimal; (2) that the coder allow the conditional probability mass function (PMF) to be specified explicitly; and (3) that the conditional PMF be allowed to change freely over time. Because arithmetic coding meets all of these requirements, we use it here. In particular, we use an arithmetic coder described in previous work[8] because it is both available and familiar to us, but other implementations of arithmetic coding would do as well. Arithmetic coding results in a code bit stream whose size, on average, closely matches that predicted on the basis of the $-\log_2 P(y|\text{ context })$ formula. Therefore, the compression that will be achieved depends almost entirely on the quality of this estimated conditional PMF for each quantized pixel. Estimating this conditional PMF is the purpose of the statistical modeling unit, which is discussed next.

## 3.2. Statistical Modeling Unit

The purpose of the statistical modeling unit is to supply the arithmetic coder with an accurate estimate of the conditional PMF for the next quantized pixel to be encoded, where accuracy is measured in relative entropy[11] between the empirical distribution of the data being compressed and that predicted by the model. The more accurate the statistical model, the greater the compression. Complicating this picture is the fact that the distributions we are concerned with are different for different pixels, and they depend on (i.e., are conditioned on) a set of quantized versions of previously processed nearby pixels.

The importance of accurate statistical modeling in the proposed system becomes particularly apparent in a comparison with transform coding. In systems of the latter type, such as JPEG, the transformed signal is quite predictable using a simple model, as the transformation has removed much of the correlation among signal elements. In the proposed system, there is no such decorrelating transformation, so a more sophisticated statistical model must be used to achieve a comparable degree of predictability (and hence compression). Spatially neighboring pixels typically have ample statistical interdependence that may be exploited for compression, provided that the statistical model is capable of representing it.

One example of a model structure capable of usefully capturing nonlinear statistical dependence in the absence of a transform is the finite Gaussian mixture.[14] Here, such a model is applied in the following way. First, the unquantized pixels in a local semicausal neighborhood such as the two shown in Figure 3 are modeled semiparametrically. Specifically, a finite mixture of separable (diagonal-covariance) Gaussian densities is assumed, and parameter values are estimated on the basis of training data. Next, to encode each quantized pixel in raster order in a test image, specific values of previously encoded pixels are substituted into the mixture estimate for the conditioning pixels (the

(a) (b)

**Figure 3.** Semicausal pixel neighborhoods of a type suitable for use in the proposed system.

positions of which are indicated by filled circles in the figure). In boundary cases where one or more conditioning pixels would fall outside the image, the missing values are integrated out of the density. Because of the the diagonal covariance structure of each component, the integration can be accomplished by simply omitting the corresponding dimensions in the expression for the density. The resulting expression is then normalized to yield a conditional density for the current pixel (whose position is indicated by the open circle). Finally, the conditional density is integrated over each quantization region to yield the desired conditional PMF of the quantized pixel.

The use of quantized rather than original conditioning pixels reduces the accuracy of the estimate, but appears to be unavoidable as the conditioning information must be available at the decoder. If the joint pixel density is sufficiently smooth and the quantization sufficiently fine, then the resulting loss in accuracy will be small.

The structural parameters of the model are the number $N$ and locations of the conditioning pixels and the number $M$ of mixture components. For simplicity, the locations of the conditioning pixels are always taken to be the $N$ preceding pixels nearest the current pixel. Suitable values for both $M$ and $N$ are selected among a set of plausible values on the basis of model fit to hold-out data. For the non-structural parameters, i.e., the mixing weights and the means and covariance matrices of the individual Gaussians, maximum-likelihood values are estimated via the EM algorithm[15,16] on a separate set of training images.

Note that maximizing the likelihood of the joint distribution of pixels in the neighborhood, which is what we do here, is *not* the same as maximizing the average likelihood of the conditional distribution for the pixel being quantized, which is what we would ideally like to do. It is easy to construct examples for which the two criteria result in very different parameter estimates. There is some discussion of this issue in previous work[3] and in the text by Bishop,[17] and it is clear that maximizing joint likelihood can be significantly improved upon in many instances. Nevertheless, we adopt the joint likelihood criterion for the sake of simplicity, recognizing that while not ideal, it is also not always at odds with the objective of maximizing average conditional likelihood. In fact it is consistent with it in the case of sufficiently large $M$ and abundant training data.

## 4. DISCUSSION

A preliminary indication of the potential of the proposed approach is provided by the example in Figure 4. A $256 \times 256$ region of a scanned 256-level grayscale text document is compressed in three different ways: by JPEG, by fixed-offset quantization with entropy coding, and by adaptive-offset quantization with entropy coding. For both the fixed- and adaptive-offset results, Gaussian mixtures of $M = 512$ components were used, and the neighborhood was as shown in Figure 3b. Training was performed on a set of ten $512 \times 512$ text regions; this training set did not contain the test image. While JPEG results in good mean-square error (MSE) for the bitrate it produces, it introduces objectionable artifacts near edges. This is an example of MSE failing to capture important desiderata of subjective evaluation, in this case smoothness of background and smoothness and sharpness of edges. Adaptive-offset quantization compares favorably with JPEG in terms of rate-versus-MSE (see caption for details), while avoiding the subjectively objectionable artifacts. It also avoids the contouring artifacts introduced by fixed-offset quantization.

Although it is difficult to draw meaningful conclusions from a single experiment, the results are encouraging. A current difficulty in running larger-scale experiments is the extreme computational burden involved both in evaluating

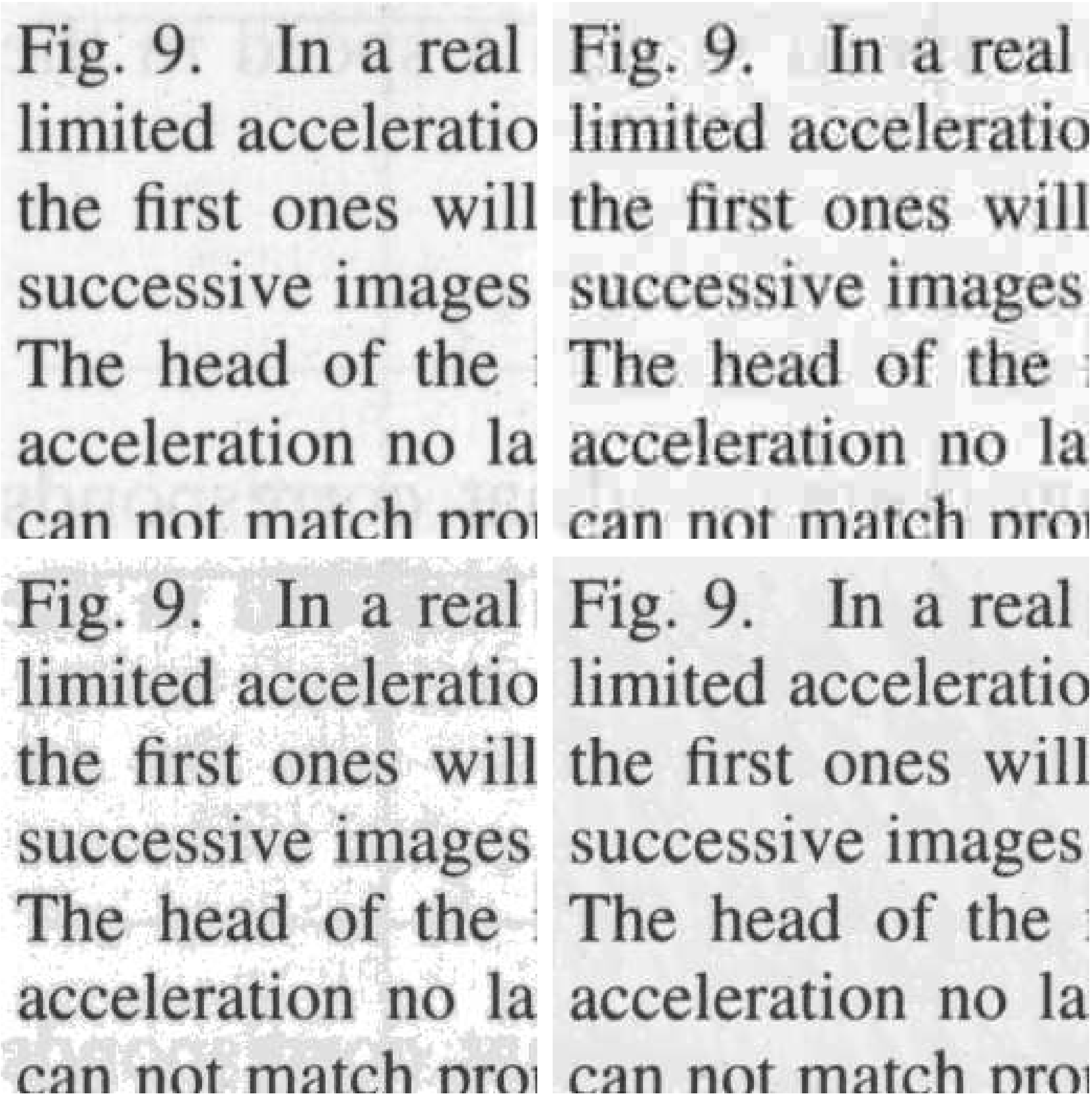


**Figure 4.** Results on a sample image. **Top left**: A $256 \times 256$ patch from a scanned original grayscale text image. **Top right**: result of JPEG compression at 0.54 bpp (MSE = 12.78). **Bottom left**: result of fixed-stepsize ($\Delta = 32$), fixed-offset arithmetic coded quantization using a finite Gaussian mixture model having $M = 512$ components, and using the $N = 10$ predictor pixels of Figure 3b. The rate is 1.13 bpp and the MSE is 10.74. Note the rough contouring artifacts. **Bottom right**: result of using the same parameters but allowing the offset to adapt in the manner described in the text. Here the rate is 0.58 bpp and the MSE is 8.32. Besides the objective rate-versus-MSE improvement over both JPEG and fixed-offset quantization, there is a subjective improvement in the bottom right as well.

the mixture density and in searching for a good offset value, both of which operations must be performed at every pixel location. Implementation of the foregoing mixture model is computationally intensive, even after care is taken to use approximations and table-lookup operations where appropriate. The use of a finite Gaussian mixture has been motivated here by its conceptual simplicity, generality, and sufficiency for the purpose of demonstrating the potential of the overall approach to lossy compression of grayscale document images. To make the approach more practical and to set the stage for larger-scale experiments, a computationally simpler model will be required. For instance, the use of a hybrid decision-tree/mixture model may significantly expedite the computation of the predictive distribution, while at the same time improving its accuracy.[3] Also, an improved search strategy or even a table lookup operation might greatly speed the adaptation of the offset. Investigation along these lines is left to future work.

Besides the potential for good performance in both the rate-distortion and subjective senses suggested by the results of Figure 4, the proposed system has two other characteristics that are worth noting. The first is its ability to reduce contour artifacts usually associated with coarse quantization in the pixel domain, without introducing the ringing, blocking, or other objectionable spatial artifacts that can accompany transform-based approaches. The reason for this is that in adaptive-offset quantization, the output value can assume a continuum of values, even when the quantization is extremely coarse. In grayscale document images, coarse fixed-offset quantization can result in contouring when the background is shaded or when the scanning illumination is not strictly uniform. Adaptive-offset quantization can reduce the severity of such artifacts. Whether or not the elimination of contouring happens automatically or to the degree desired depends, like much else in the system, on the accuracy of the statistical model. If the model remains sufficiently accurate within graded regions, then the entropy-minimizing $\phi$ will likewise vary smoothly in those regions, by continually seeking to center the mode of the conditional distribution within its quantization region. This in turn will induce the output of the quantizer to be a smoothly varying estimate of that mode, thereby ameliorating contouring automatically. Alternatively, if the model is not sufficiently accurate to ensure this behavior, a penalty term involving values of $\phi$ used at previously processed pixel locations can be added to the conditional entropy when choosing $\phi$ to ensure that it does not undergo abrupt changes as a function of spatial position.

Another characteristic of the proposed system is its ability to integrate high-level, externally supplied segmentation information in a seamless way. Document images typically consist of several types of regions such as text, halftone images, and line drawings. Each type of region has its own fidelity requirements and statistical properties. Although one could always apply the proposed compression technique separately to regions of a pre-segmented image, the requisite adaptation can instead be accomplished more smoothly by switching the quantization and statistical model parameters on the basis of estimated region type. The estimated region-type label could be externally supplied, or else the statistical modeling unit could take on the task of estimating it as an additional responsibility. In either case, the entire image is encoded seamlessly in a single raster scan.

## 5. CONCLUSION

We have considered an approach to grayscale document image compression in which the adaptation of the parameters of the lossy element — a scalar quantizer — is informed by downstream considerations, specifically by the entropy of the estimate of the conditional PMF that is used in encoding the quantized pixel value by the arithmetic coder. The technique operates directly in the grayscale pixel domain, rather than by first applying an energy-compacting or other spatial transformation. Preliminary results suggest that the scheme can be competitive with JPEG in terms of rate-distortion performance, and may offer clear advantages in subjective quality for document images. Specifically, blocking artifacts are avoided and edges are reproduced faithfully. Compared with fixed-offset quantization, it achieves better compression and avoids contouring artifacts.

Comparison with other lossy techniques more suited than JPEG to the sharp edges in document images, as well as large scale experiments on a variety of documents, are needed before definite conclusions can be drawn about the relative merits of the approach. The running of large-scale experiments will be simplified when the high computational load of the technique in its present form is reduced. Reducing computational complexity through the use of better models and through other means is an area of ongoing investigation.

## ACKNOWLEDGMENTS

This work has benefitted from discussions with Dan Bloomberg, also at the Xerox Palo Alto Research Center.

## REFERENCES

1. Z. Fan, "JPEG decompression with reduced artifacts," in *SPIE Proceedings: Image and Video Compression*, M. Rabbani and R. J. Safranek, eds., vol. 2186, pp. 50–55, The International Society for Optical Engineering, February 1994.
2. Joint Bi-Level Image Experts Group (JBIG) Committee, "Information technology – coded representation of picture and audio information – lossy/lossless coding of bi-level images," Tech. Rep. 14492 FDC, ISO/IEC, July 1999.
3. A. C. Popat, *Conjoint Probabilistic Subband Modeling*. PhD thesis, Massachusetts Institute of Technology, 1997.
4. K. Popat and D. S. Bloomberg, "Two-stage lossy/lossless compression of grayscale document images," in *Mathematical Morphology and its applications to image and signal processing: Proceedings of the Fifth International Symposium on Mathematical Morphology*, J. Goutsias, L. Vincent, and D. S. Bloomberg, eds., pp. 361–370, Kluwer, (Palo Alto, California), 2000.
5. N. S. Jayant and P. Noll, *Digital Coding of Waveforms : Principles and Applications to Speech and Video*, Prentice-Hall, Englewood Cliffs, NJ, 1984.
6. A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer academic publishers, 1992.
7. H. Gish and J. Pierce, "Asymptotically efficient quantization," *IEEE Transactions on Information Theory* **IT-14**, pp. 676–683, September 1968.
8. A. C. Popat, "Scalar quantization with arithmetic coding," Master's thesis, Dept. of Elec. Eng. and Comp. Science, M.I.T., Cambridge, Mass., 1990.
9. N. Farvardin and J. W. Modestino, "Adaptive buffer- instrumented entropy-coded quantizer performance for memoryless sources," *IEEE Transactions on Information Theory* **IT-32**, pp. 9–22, 1986.
10. J. Gailly and M. Adler, "The gzip homepage," December 1999. http://www.gzip.org/algorthm.txt.
11. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.
12. J. J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM J. Res. Develop.* **23**, pp. 149–162, March 1979.
13. I. Witten, R. Neal, and J. Cleary, "Arithmetic coding for data compression," *Communications of the ACM* **30**, pp. 520–540, June 1987.
14. G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*, Marcel Dekker, New York, 1988.
15. A. P. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc.* **39**, pp. 1–38, 1977.
16. R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood, and the EM algorithm," *SIAM Review* **26**, pp. 195–239, April 1984.
17. C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, New York, 1995.