Improved Mixed Document Compression by Using the DCT Coefficient Distributions

Edmund Y. Lam
Department of Electrical and Electronic Engineering,
University of Hong Kong,
Hong Kong.

Abstract

Data compression is an important area in electronic document processing. A document can consist of text and image, with different statistical behavior. When we use JPEG compression, it is advantageous to tune the parameters for individual blocks to enhance the decompression quality. We demonstrate a method that can be incorporated in the JPEG decoder, which utilizes the amplitude distribution of the DCT coefficients for texts and images to achieve better image quality. This involves both a discriminator function to differentiate between text and image, and an adjustment to the decompression methodology that shifts the decoding value to the minimum mean-square error location in the codeblock.

1. Introduction

Digital documents are ubiquitous in today's information-driven society. For transmission and archival, it is very important to be able to compress them efficiently. Many of these documents contain a mixture of data types, such as natural images, text, line art, and background. It is known that these data types have different characteristics. However, often we will only use a single compression method, such as JPEG [1], for the entire image. Fortunately, we usually still have some control on the parameters of the algorithm to adapt it for different image types. Compression of mixed document by varying these parameters has received significant attention in recent years [2, 3].

The discrete cosine transform (DCT) is at the core of the JPEG algorithm, together with scalar quantization and entropy coding. It is known that the DCT coefficients for natural images can be modeled with a Laplacian distribution [4]. This knowledge can be employed to improve compression efficiency, by shifting the decoding value from the mid-point of the codeblock to the centroid, which gives the minimum mean-square error [5]. This principle can be applied to other image types, provided we have *a priori* knowledge about the distribution characteristics. In this paper, we examine how we can model the DCT coefficients for the documents with both text and images. We use a doubly stochastic model, where the distribution of the variance of the 8×8 blocks becomes the key to this analysis. This model is explained in details in section 2. We will then examine how to incorporate this knowledge into our design of a decompression scheme for mixed document in section 3, as an extension to the standard JPEG decoder. This is achieved by varying the quantization matrices for different data types. We also examine how much gain in signal-to-noise ratio this method can attain in section 4.

2. Image Model

A doubly stochastic model for the coefficient statistics has been shown to be very effective [4]. In the 8×8 blocks used for DCT, assuming that the pixels are identically distributed, the DCT coefficient is approximately Gaussian. Note that we do not need the pixels to be independent. Let I denote the coefficient, and σ^2 denote the variance of the block, we have

$$\mathcal{P}(I|\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{I^2}{2\sigma^2}}. \qquad (1)$$

In the doubly stochastic model, the block variance is itself a stochastic quantity. The actual DCT coefficient distribution is given by

$$\mathcal{P}(I) = \int_0^\infty \mathcal{P}(I|\sigma^2)\mathcal{P}(\sigma^2)\, d(\sigma^2). \qquad (2)$$

From this equation, we can see that $\mathcal{P}(\sigma^2)$ is a determining factor for the distribution of the transform coefficients.

For natural images, the distribution of the block variance resembles an exponential distribution, *i.e.*,

$$\mathcal{P}(\sigma^2) = \lambda e^{-\lambda(\sigma^2)}. \qquad (3)$$

Putting equation (1) and (3) in equation (2), after some manipulation, we have

$$\mathcal{P}(I) = \frac{\sqrt{2\lambda}}{2} e^{-\sqrt{2\lambda}|I|}. \qquad (4)$$

Therefore, the distribution of the DCT coefficients for natural images is Laplacian.

On the other hand, this is not the case for text documents [5]. The distribution of the block variance generally consists of two components:

1. A large concentration of the variance at or near zero.

2. A near uniform distribution of the variance otherwise.

The first component corresponds to a flat region, which represents the background in the document. The contribution from the second component is more important for our study here, because only the lowest DCT coefficient will be affected if a region has zero variance. This coefficient is not modeled with any known distribution. For a uniform distribution for the variance, we have $\mathcal{P}(\sigma^2) = (q-p)^{-1}$ for $p \leq \sigma^2 \leq q$. We can put this in equation (2), but that does not lead to a closed-form solution. We can, however, perform the integration numerically. The result is shown to resemble a Gaussian distribution [5].

Laplacian and Gaussian distributions can both be considered special cases of the generalized Gaussian distribution. The generalized Gaussian distribution, with zero mean, has the probability density function

$$\mathcal{P}(I) = \frac{\nu}{2\beta\Gamma\left(\frac{1}{\nu}\right)} e^{-\left(\frac{|I|}{\beta}\right)^{\nu}}, \tag{5}$$

where $\nu > 0$ controls the shape of the distribution and β the spread. $\Gamma(\cdot)$ is the Gamma function defined as

$$\Gamma(x) = \int_0^{\infty} \zeta^{x-1} e^{-\zeta} d\zeta, \tag{6}$$

with $x > 0$. When $\nu = 2$ and $\beta = \sqrt{2}\sigma$, it becomes a standard Gaussian distribution. When $\nu = 1$ and $\beta = 1/\lambda$, it becomes a Laplacian distribution with parameter λ. Text and natural images therefore can be considered to produce coefficient distributions with generalized Gaussian distribution having different shape parameters.

3. Decompression Scheme

We need to distinguish between text and image regions in a mixed document. Unlike many traditional image segmentation algorithms, we want the classification to be done on a block basis rather than at the pixel level. This has the added advantage of minimizing both memory and processing requirements.

Let $D(j)$ be a discriminator function on the jth block, which indicates whether the block should be classified as text or as image. We simply the method proposed in [2] so that each $D(j)$ is independent of its neighbors. We compute $D(j)$ as a function of the 63 AC coefficients with

$$D(j) = \sum_{k=1}^{63} g\left(I_{q,k}(j)\right), \tag{7}$$

where

$$g(x) = \begin{cases} \log_2(|x|) + 4 & \text{if } |x| > 1 \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

$I_{q,k}$ denotes the quantized DCT coefficient at the kth subband. A higher value in $D(j)$ indicates that this block is more likely to be text. In fact, we use the value $D(j)$ to decide on the nature of the block j as follows:

$$\begin{cases} D(j) \approx 0 & \Rightarrow \text{block is background} \\ D(j) < T & \Rightarrow \text{block is image} \\ D(j) \geq T & \Rightarrow \text{block is text.} \end{cases} \tag{9}$$

If the block is background, there is no need for any adjustment to the DCT coefficients. When the block is classified as image, we use a Laplacian probability density function to model the AC coefficients. Let the codeblock range from a to b. In the case of JPEG, if the quantization table at that frequency is Q_j and the code value for that block is k (assuming $k > 0$ for the calculation below, without loss of generality), then $a = (k-0.5)Q_j$ and $b = (k+0.5)Q_j$. The centroid of this range is

$$\begin{aligned} \hat{x} &= \frac{\int_a^b x \frac{\lambda}{2} e^{-\lambda x} dx}{\int_a^b \frac{\lambda}{2} e^{-\lambda x} dx} \\ &= \frac{\frac{1}{2}\left[-xe^{-\lambda x} - \frac{1}{\lambda} e^{-\lambda x}\right]_a^b}{\frac{1}{2}e^{-\lambda a} - \frac{1}{2}e^{-\lambda b}} \\ &= \frac{ae^{-\lambda a} + \frac{1}{\lambda}e^{-\lambda a} - be^{-\lambda b} - \frac{1}{\lambda}e^{-\lambda b}}{e^{-\lambda a} - e^{-\lambda b}} \\ &= \frac{ae^{-\lambda a} - be^{-\lambda b}}{e^{-\lambda a} - e^{-\lambda b}} + \frac{1}{\lambda}. \end{aligned} \tag{10}$$

When the block is classified as text, we use a Gaussian probability density function to model the AC coefficients. The centroid of this range is

$$\begin{aligned} \hat{x} &= \frac{\int_a^b x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx}{\int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} dx} \\ &= \frac{\frac{1}{2\sqrt{2\pi}\sigma} \int_{a^2}^{b^2} e^{-\frac{y}{2\sigma^2}} dy}{Q\left(\frac{a}{\sigma}\right) - Q\left(\frac{b}{\sigma}\right)} \\ &= \frac{\frac{\sigma}{\sqrt{2\pi}} \left(e^{-\frac{a^2}{2\sigma^2}} - e^{-\frac{b^2}{2\sigma^2}}\right)}{Q\left(\frac{a}{\sigma}\right) - Q\left(\frac{b}{\sigma}\right)}, \end{aligned} \tag{11}$$

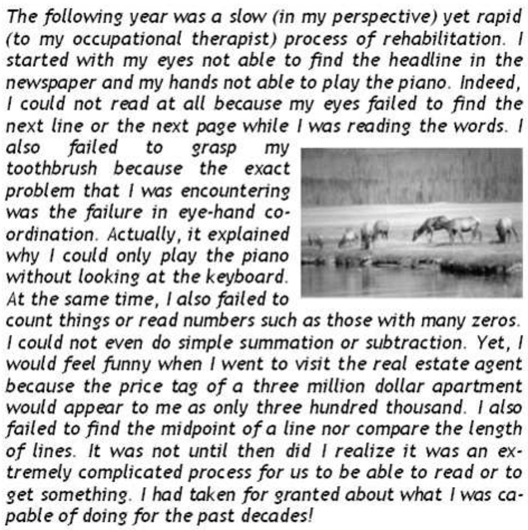
where the function $Q(x)$ is the Q-function defined as [6]

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-t^2} dt. \qquad (12)$$

In short, we modify the JPEG decompression with the following scheme: compute $D(j)$ for each block to decide the nature of the block. If it is classified as background, we simply use the original decompression scheme. If it is classified as image, we shift the decoding value according to equation (10). If it is classified as text, we use equation (11) instead.

4. Simulation

To test the ideas proposed in this paper, we evaluate the performance of the algorithm on a couple of test images. Figure 1 shows one such image, with predominantly text and an embedded image. The figure is of size 512×512 pixels. We also test with documents such as Figure 2, which is mostly image but with embedded text. This latter one is of size 256×256 pixels.



The following year was a slow (in my perspective) yet rapid (to my occupational therapist) process of rehabilitation. I started with my eyes not able to find the headline in the newspaper and my hands not able to play the piano. Indeed, I could not read at all because my eyes failed to find the next line or the next page while I was reading the words. I also failed to grasp my toothbrush because the exact problem that I was encountering was the failure in eye-hand coordination. Actually, it explained why I could only play the piano without looking at the keyboard. At the same time, I also failed to count things or read numbers such as those with many zeros. I could not even do simple summation or subtraction. Yet, I would feel funny when I went to visit the real estate agent because the price tag of a three million dollar apartment would appear to me as only three hundred thousand. I also failed to find the midpoint of a line nor compare the length of lines. It was not until then did I realize it was an extremely complicated process for us to be able to read or to get something. I had taken for granted about what I was capable of doing for the past decades!

Figure 1: *An example of a mixed document.*

We test the images in a couple of ways. First, we use the decompression mechanism in the baseline JPEG, which does not assume any distribution in the coefficients. We record the signal-to-noise ratio (SNR) of the resultant image as compared with the original. Second, we test with our algorithm without the discriminator function, and assumes that the DCT coefficients for all the blocks have a Laplacian distribution. Third, we examine the case where the DCT coefficients for all the blocks are assumed to have a Gaussian distribution. Finally, we test with a mixed model, using the discriminator function described above. We set the threshold $T = 180$. Therefore, each block is classified and we use the two prior models in decompression.

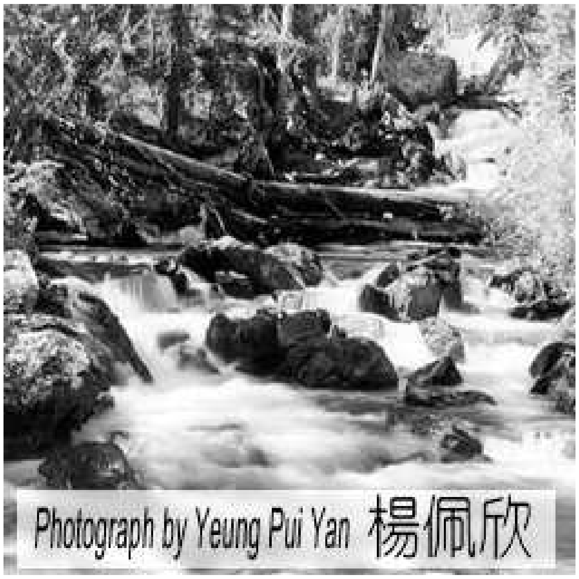


Figure 2: *An example of a text inside an image.*

The results for the two images are summarized in table 1. In both cases, we observe that using a biased reconstruction will always produce an image with better quality. For Figure 1, we see that assuming all the blocks have a Gaussian distribution produces better SNR than a Laplacian distribution. This is in line with the discussion above that for a document with text, the transform coefficient distribution resembles Gaussian. Using a mixed model will further increase the quality by only a small margin. On the other hand, for Figure 2, which is predominantly image, using a Laplacian model produces better quality output than using a Gaussian model. Again, this is in accordance with the theoretical discussion in section 2. Once again, a mixed model produces little improvement over a single model. These results indicate that if a document has predominantly text or image, a single model will suffice. However, if both have significant proportion, it is better to use the discriminator function to classify the document, and then apply the appropriate model for the best decompression performance.

5. Conclusions

In this paper, we have proposed a mechanism of improving the decompression quality of mixed documents by taking advantage of the DCT coefficient distributions. This method is seen to produce documents with better quality than using the baseline JPEG. However, this is achieved

	SNR	
	Fig. 1	Fig. 2
Normal dequantization	20.40dB	17.75dB
Laplacian model	20.47dB	18.05dB
Gaussian model	20.73dB	17.93dB
Mixed model	20.74dB	18.05dB

Table 1: *Simulation Results.*

at the expense of more computation. Further improvement can be made by using the generalized Gaussian distribution as a model for both text and natural images, and by reducing the computational workload in this algorithm.

6. Acknowledgement

This work was supported in part by the University Research Committee of the University of Hong Kong under Grant Number URC-10204526.

7. References

1. William Pennebaker and Joan Mitchell, *JPEG Still Image Data Compression Standard*, Van Nostrand Reinhold, New York, 1992.
2. Konstantinos Konstantinides and Daniel Tretter, "A JPEG variable quantization method for compound documents," *IEEE Transactions on Image Processing*, vol. 9, no. 7, pp. 1282–1287, July 2000.
3. Marcia G. Ramos and Ricardo L. de Queiroz, "Classified JPEG coding of mixed document images for printing," *IEEE Transactions on Image Processing*, vol. 9, no. 4, pp. 716–720, April 2000.
4. Edmund Y. Lam and Joseph W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1661–1666, October 2000.
5. Edmund Y. Lam, "Improved mixed document compression by using the DCT coefficient distributions," accepted for publication in *IEEE Signal Processing Letters*.
6. Alberto Leon-Garcia, *Probability and Random Processess for Electrical Engineering*, Addison Wesley, Reading, Massachusetts, second edition, 1994.