Analysis of the DCT Coefficient Distributions for Document Coding

Edmund Y. Lam, *Member, IEEE*

***Abstract*—It is known that the distribution of the discrete cosine transform (DCT) coefficients of most natural images follow a Laplacian distribution, and this knowledge has been employed to improve decoder design. However, such is not the case for text documents. In this letter, we present an analysis of their DCT coefficient distributions, and show that a Gaussian distribution can be a realistic model. Furthermore, we can use a generalized Gaussian model to incorporate the Laplacian distribution found for natural images.**

***Index Terms*—Discrete cosine transform (DCT), document processing, image analysis, image coding, probability statistics.**

I. Introduction and Motivation

ELECTRONIC document processing is getting more important these days with the widespread practice of "paperless" office. In particular, to save on the memory required to archive the documents, a good compression mechanism is needed. Specialized compression schemes have been developed for texts, such as JBIG and JBIG2 [1]. However, there is also a tendency to use methods designed for natural images in document compression as well, such as when the application does not support these text compression techniques. This is also needed for mixed document compression, where it may contain text, line art, images, and background together [2].

Text and natural images have many different characteristics, and compression schemes such as JPEG [3] and JPEG 2000 [4] that are designed for the latter may not work well for the former. For example, it is known that using JPEG on text, even with moderate compression ratio, there are noticeable artifacts along the edges that degrade the visual quality. In this letter, we focus our attention on analyzing the distribution of the discrete cosine transform (DCT) coefficients for text documents. It has been justified that such distribution resembles a Laplacian distribution for natural images [5], and this knowledge has been used to improve the decoding in JPEG [6], [7]. Fig. 1(a) shows a typical grayscale image, and Fig. 1(b) shows the distribution of one of the DCT coefficients. The dotted line represents the maximum-likelihood Laplacian fit of the empirical data. As seen in the Figure, the data match the distribution rather well. On the other hand, consider Fig. 2. Fig. 2(a) shows a typical text image,

(a)

(b)

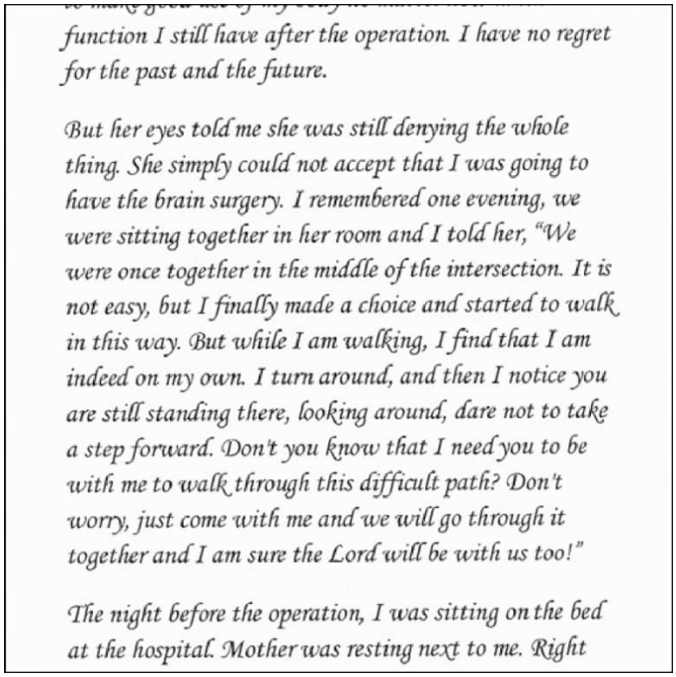
Fig. 1. Grayscale image with one DCT coefficient distribution.

and Fig. 2(b) shows the distribution of the corresponding DCT coefficient. Again we show the maximum-likelihood Laplacian fit of the empirical data. This is not a good fit. Therefore, if we design a decoder that shifts the decoding value because the coefficient behaves like a Laplacian distribution [6], it may not be

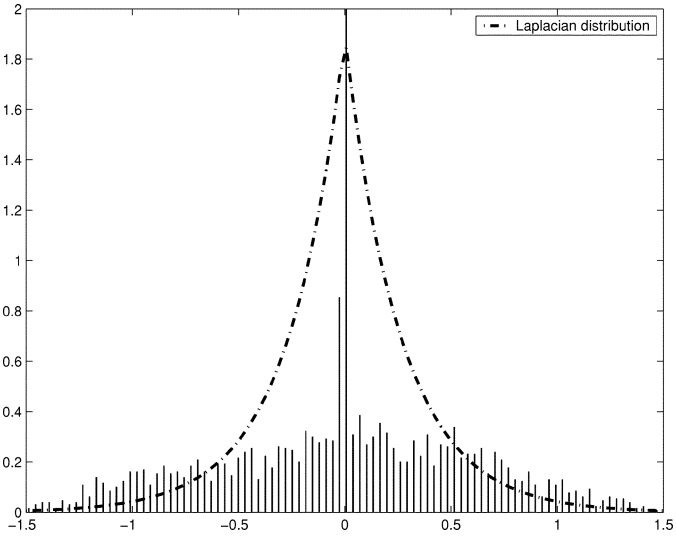
Manuscript received November 8, 2002; revised February 7, 2003. This work was supported by the University Research Committee in the University of Hong Kong under Grant Number URC-10204526. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Scott. T. Acton.

The author is with the Department of Electrical and Electronic Engineering, University of Hong Kong, Hong Kong (e-mail: elam@eee.hku.hk).

Digital Object Identifier 10.1109/LSP.2003.821789



(a)



(b)

Fig. 2. Text image with one DCT coefficient distribution.

applicable for text images. In Section II, we formulate a mathematical framework to examine what would be a better model for the DCT coefficient distributions for text.

II. DCT Coefficient Distribution Modeling

In [5], it was shown that a doubly stochastic model is very helpful to provide us an insight into the distribution. In this model, we argue that within an 8×8 block used for DCT, assuming that the pixels are identically distributed, the DCT coefficient is approximately Gaussian. Let I be the coefficient, and σ^2 be the variance of the block, we have

$$\mathcal{P}(I|\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{I^2}{2\sigma^2}\right\}. \tag{1}$$

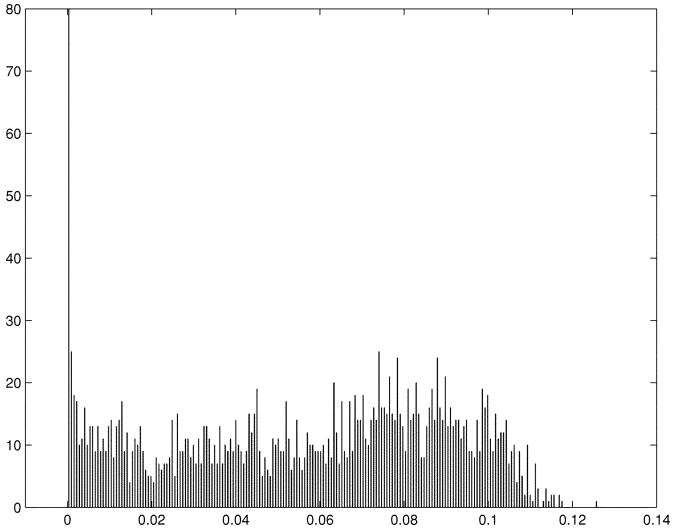


Fig. 3. Distribution of block variance for a text image.

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

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

It has also been shown that for natural images, the distribution of the block variance is close to exponential, in which case we can put in $\mathcal{P}(\sigma^2) = \lambda\exp\left\{-\lambda\sigma^2\right\}$ and (1) in (2) to get

$$\mathcal{P}(I) = \frac{\sqrt{2\lambda}}{2}\exp\left\{-\sqrt{2\lambda}|I|\right\} \tag{3}$$

which largely explains why the empirical distribution for DCT coefficients for natural images are close to Laplacian [5]. On the other hand, this is not the case for text documents. Fig. 3 shows the distribution of the block variance for Fig. 2(a). As seen in the diagram, there are two components in the distribution.

1) A large concentration of the variance at or near zero.
2) A nearly uniform distribution of the variance otherwise.

The first component corresponds to a flat region, which represents the background in the document. We focus on the contribution from the second component, because only the lowest DCT coefficient will be affected if a region has zero variance, and this coefficient is not modeled with any known distribution anyway. We will analyze the case where the block variance is uniform, and the case that it is Gamma-distributed. The second case will help us tie into the exponential distribution for natural images.

A. Uniform Distribution of Block Variance

We put $\mathcal{P}(\sigma^2) = (b-a)^{-1}$ for $a \le \sigma^2 \le b$ in (2), and obtain

$$\begin{aligned}\mathcal{P}(I) &= \int_a^b \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{I^2}{2\sigma^2}\right\}\cdot\left(\frac{1}{b-a}\right)d(\sigma^2) \\ &= \frac{1}{\sqrt{2\pi}(b-a)}\int_a^b \frac{1}{\sqrt{s}}\exp\left\{-\frac{I^2}{2s}\right\}ds\end{aligned} \tag{4}$$

when we substitute s for σ^2. We cannot find a closed-form solution for (4), but we can perform the integration numerically, and

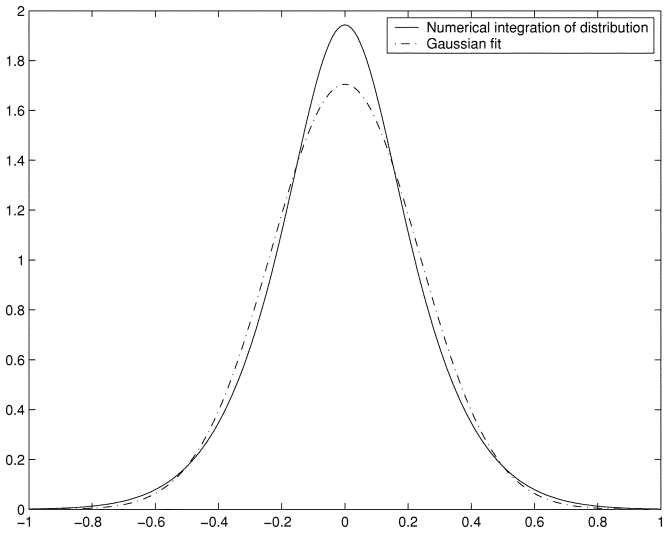


Fig. 4. DCT coefficient distribution with uniform block variance.

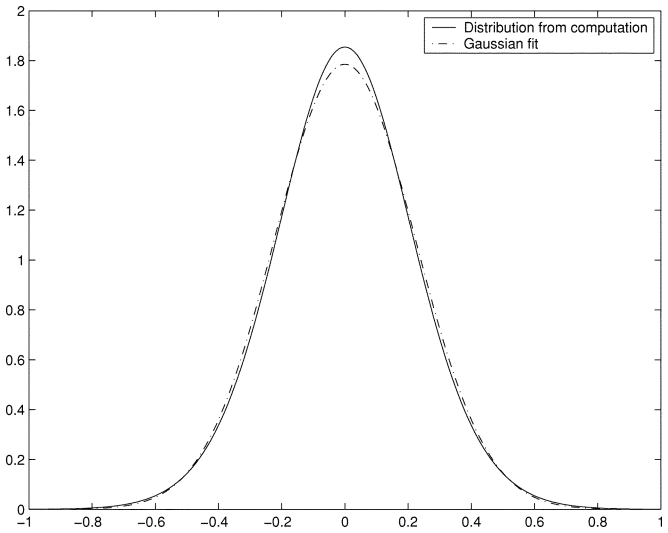


Fig. 5. DCT coefficient distribution with gamma block variance.

the result is plotted in Fig. 4. We observe that its shape resembles a normal distribution, so we also plot the closest Gaussian curve in the figure.

B. Gamma Distribution of Block Variance

Gamma distribution is defined as

$$\mathcal{P}(x) = \begin{cases} 0, & x < 0 \\ \frac{\lambda(\lambda x)^{\alpha-1}\exp\{-\lambda x\}}{\Gamma(\alpha)}, & x \geq 0 \end{cases} \tag{5}$$

where $\Gamma(\cdot)$ is the gamma function. α is called the shape parameter, while λ is called the scale parameter [8]. The Gamma distribution becomes the exponential distribution for $\alpha = 1$. As α increases, the mean of the distribution will also shift up. Although the Gamma distribution is more complicated than the uniform distribution, when put it in (2) there is actually a closed-form solution

$$\mathcal{P}(I) = \frac{\sqrt{\lambda}\exp\left\{-\sqrt{2\lambda}|I|\right\}}{(\alpha-1)!(2^{2\alpha-3/2})} \sum_{v=0}^{\alpha-1} \frac{(2\alpha-v-2)!(2\sqrt{2\lambda}|I|)^v}{v!(\alpha-v-1)!} \tag{6}$$

for integer value of α. The detailed derivation can be found in [9]. Fig. 5 plots (6) for $\alpha = 10$. Again we also plot the closest Gaussian curve in the figure.

Since $\mathcal{P}(I)$ is a linear function of $\mathcal{P}(\sigma^2)$, we can infer from the above two special cases that a block variance distribution close to a linear combination of them will lead to a DCT coefficient distribution resembling Gaussian. The data as plotted in Fig. 3 is seen to be one such case. As a verification to the above analyses, we use the data in Fig. 2(b) but ignore the coefficients near zero, which would mainly come from the background of the text. We plot the maximum-likelihood estimate of the Gaussian distribution to the data in Fig. 6. It confirms our analysis that a Gaussian fit to the coefficient is better for text document.

Note that the Laplacian and Gaussian distributions can be considered special cases for the generalized Gaussian distribution. Its probability density function, with zero mean, is

$$\mathcal{P}(I) = \frac{\nu}{2\beta\Gamma\left(\frac{1}{\nu}\right)}\exp\left\{-\left(\frac{|I|}{\beta}\right)^{\nu}\right\} \tag{7}$$

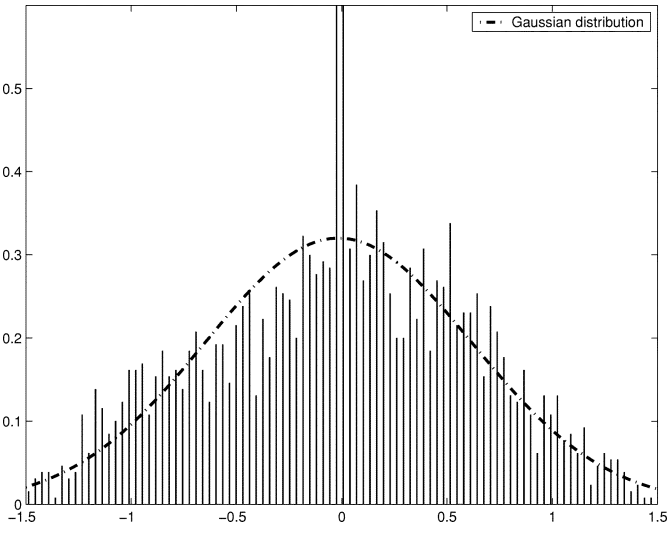


Fig. 6. Gaussian fit to empirical coefficient distribution of text document.

where $\nu > 0$ controls the shape of the distribution and β the spread. 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. Combining this study with the study of DCT coefficient distributions for natural images, we argue that the generalized Gaussian distribution is a very versatile model for such a purpose. The shape parameter is around 1 for natural images, and near 2 for text. This will be especially helpful for mixed document compression, where both images and text are present [10]. We compute the shape parameters for the image in Fig. 2(a), and plot the results for the 63 ac coefficients in Fig. 7. In performing this fitting, we use the method described in [11], and ignore the blocks with near zero variance. We can see that using $\nu = 2$ is a good model for text documents.

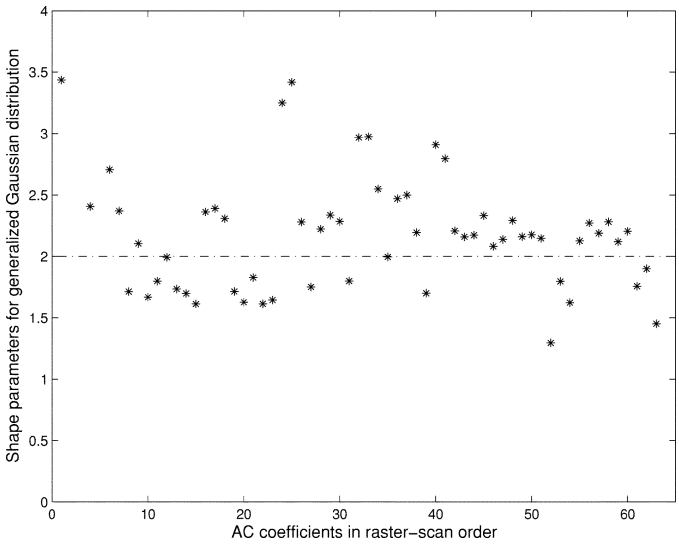


Fig. 7. AC coefficients for generalized Gaussian fit to text document.

III. Conclusion

It has been shown that for natural images, the coefficient distribution resembles a Laplacian distribution. This letter extends the study into text documents. Although the model would call for rather complicated equations such as (4) and (6) to compute the distribution, we find that a Gaussian distribution is a much simpler model that is still rather accurate. This can further be considered a special case of the generalized Gaussian distribution, and can combine with the Laplacian distribution for natural images under the same model. Since the distribution of the coefficients has been widely studied and used in decoder design, the study presented here can serve as an aid to improve text document compression, while the actual coding gain still needs to be explored.

Acknowledgment

The author wishes to thank P. Y. Yeung for the document in Fig. 2(a) and the University Research Committee in the University of Hong Kong for financial support.

References

[1] I. Witten, A. Moffat, and T. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images*, 2nd ed. New York: Morgan Kaufmann, 1999.

[2] J Li and R. Gray, "Context-based multiscale classification of document images using wavelet coefficient distributions," *IEEE Trans. Image Processing*, vol. 9, pp. 1604–1616, Sept. 2000.

[3] W. Pennebaker and J. Mitchell, *JPEG Still Image Data Compression Standard*. New York: Van Nostrand Reinhold, 1992.

[4] D. Taubman and M. Marcellin, *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Boston, MA: Kluwer, 2001.

[5] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Processing*, vol. 9, pp. 1661–1666, Oct. 2000.

[6] G. Lakhani, "Distribution-based restoration of DCT coefficients," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 10, pp. 819–823, Aug. 2000.

[7] J. Price and M. Rabbani, "Biased reconstruction for JPEG decoding," *IEEE Signal Processing Lett.*, vol. 6, pp. 297–299, Dec. 1999.

[8] J. Rice, *Mathematical Statistics and Data Analysis*, 2nd ed. Belmont, CA: Duxbury, 1995.

[9] D. Teichroew, "The mixture of normal distributions with different variances," *Ann. Math. Stat.*, vol. 28, no. 2, pp. 510–512, June 1957.

[10] E. Y. Lam, "Analysis of the DCT coefficient distributions for document coding," in *Proc. Digital Photography Conf.: PICS 2003*, May 2003, pp. 518–521.

[11] K. Sharifi and A. Leon-Garcia, "Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video," *IEEE Trans. Circuit Syst. Video Technol.*, vol. 5, pp. 52–56, Feb. 1995.