Local Entropy Estimation for Low-Rate Wavelet Image Coding

Matthew Gaubatz, Student Member, IEEE, and Sheila Hemami, Senior Member, IEEE

Abstract—Wavelet-based image coding involves discarding subband information in an organized fashion to compress an image to a given rate, i.e. some kind of rate control. In a practical coding system, this task requires knowledge of the relationship between subband quantization step-size and compressed rate. At high rates, this behavior can be characterized by modeling each subband as a Gaussian random process. At low rates, such methods break down since 1) quantized coefficients no longer resemble Gaussian distributions, and 2) most of the coefficients are zero, but the positions of non-zero coefficients are spatially dependent. These dependencies can be handled by extending models to characterize more localized behavior. Based on this observation, a model-based rate-control technique is presented. Rate estimates are generated by considering each subband a collection of quantized random processes. It is demonstrated that at low rates (0.5 bpp and below), these methods accurately predict the rate of a compressed image within five percent of the rate achieved using a state-of-the-art coder.

Index Terms—rate control, quantization, Laplacian modeling, subband entropy.

I. INTRODUCTION

Low rate wavelet image coding under a rate constraint involves controlled subband quantization. For some applications, the problem of rate control has been circumvented with embedded coders, such as SPIHT [1]. For low rate coding, however, a truncated embedded representation might not be optimal, especially in a perceptual setting such as in [2], where step-sizes cannot be described as a single function of rate. Given an optimization criterion (such as mean-squared-error or a perceptual distortion measure) that dictates some relationship between subband step-sizes, a bisection search can be used meet the rate constraint. A set of step-sizes generates quantized subbands, where the sum of the compressed subband sizes equals the target file size. A method of mapping step-sizes to rate, however, is needed to run any such rate-control algorithm.

Traditional methods of mapping step-size to rate include compressing the quantized data and subband modeling. Simply re-compressing an image until the rate constraint is met is neither elegant nor efficient. Gaussian subband modeling can be used to map step-size to rate with much less computation, but is not effective at low rates because heavily quantized subband data is not fit well with Gaussian distributions. In addition, characterizing a subband with only a distribution function does not address coefficient dependencies that are important when most coefficients in a quantized subband are zero.

More sophisticated methods have been proposed for modeling subband rate-quantization-step-size (R-Q) relationships. Gormish and Gill, for instance, presented a model whereby each subband is treated as a quantized Laplacian process [3]. This technique performs well, but overestimates the entropy of subbands at low rates. The estimation-quantization (EQ) coder uses local generalized Gaussian subband models for rate estimation; the task of fitting and adapting the data to determine step-sizes is pushed into lookup tables [4]. Generalized Gaussian modeling has also been proposed as the basis for a stripe-based rate control procedure in wavelet coders [5], which is applied on-the-fly to successive rows in a given image.

Mallat and Falzon introduced an analytical framework for coarsely quantized subbands [6] that models the entropy of the location and magnitude of non-zero quantized wavelet coefficients separately. The model is based on run-length encoding the locations of non-zero wavelet coefficients and entropy coding the remaining quantized coefficients. Current state-of-the-art coders that address the dependency between non-zero coefficients (“significant coefficients”) in a more sophisticated manner achieve superior compression performance. Percentage of significant coefficients has also been suggested as a parameter to describe subband R-Q behavior [7], [8]. As a method of handling significant coefficient dependencies, the relationship between rate and percentage of significant coefficients is coupled with a method to map this percentage to step-size. Nevertheless, this kind of model can require training data (such as the rates associated with several step-sizes) to be fit correctly [7].

Collectively, these methods involve rate-estimation techniques that illustrate trade-offs between simplicity and accuracy. For example, re-compressing an image
determines exact rate estimates at the expense of time and computational resources, while Gaussian modeling quickly and simply generates rate estimates that are higher than the target. Similarly, the quantized Laplacian estimate is simple to generate, but is not as accurate as the significant coefficient percentage based methods, which can involve a data-gathering step. In [5], for instance, the algorithm accurately estimates image rate-distortion characteristics, but is recommended for use in “off-line” applications. An ideal rate-estimation solution should require little overhead computation, especially for repeated use as in a bisection search, while yielding highly accurate rate estimates.

This paper investigates a simple yet effective model-based method for accurately predicting the entropy of a quantized wavelet image. By breaking each subband into blocks of random processes, more accurate rate estimates are achieved. Probabilities of quantized coefficients are derived based on a generalized Laplacian model. Furthermore, the trade-off between localization and accuracy is analyzed as a function of rate, giving insight into the interdependence of wavelet coefficients. This type of modeling estimates the compressed image size to within five percent of the size generated by a state-of-the-art coder, roughly 5 times more accurate than subband-wide Gaussian modeling.

This paper is organized as follows. Section II outlines rate control algorithms that require minimal manipulation of subband data. Section III gives derivations of expressions for probabilities of quantized coefficients. Results are illustrated in section IV, and section V concludes the paper.

II. “SINGLE-PASS” RATE CONTROL

Bisection search is a standard method of performing rate control. This type of algorithm is outlined below. Rate is expressed in terms of bit-per-pixel (bpp). Let \( s \) index the wavelet subbands, \( Q(s) \) represent a step-size associated with subband \( s \), and \( R_{\text{target}} \) represent the target rate.

**Subband-based bisection rate control**

1. set \( \text{tol} \ll 1 \), \( Q_{\text{low}} = 1 \), \( Q_{\text{high}} \gg 1 \) such that \( Q_{\text{high}} \) quantizes all wavelet coefficients to zero
2. set \( Q_0 = \frac{Q_{\text{low}} + Q_{\text{high}}}{2} \), \( Q(s) = Q_0 \) \( \forall s \)
3. compute \( R_{\text{estimate}} \) from the set \( \{Q(s)\} \)
4. if \( R_{\text{estimate}} > R_{\text{target}} \) set \( Q_{\text{low}} = Q_0 \)
5. else set \( Q_{\text{high}}(s) = Q_0 \)
6. if \( \frac{R_{\text{estimate}} - R_{\text{target}}}{R_{\text{target}}} < \text{tol} \) end
7. else go to step ii)

Step iii) includes data analysis required to estimate the rate, and as such a single-pass algorithm which only analyzes data once prior to compression is desirable from a computational standpoint. This quality is attractive for practical compression systems due to the associated speed and simplicity of implementation. If the coder itself is used to compute \( R_{\text{estimate}} \), the third step is given by:

1. iiiia) quantize subband \( s \) with \( Q(s) \) \( \forall s \)
2. iiib) set \( R_{\text{estimate}} = \frac{\text{compressed file size (in bits)}}{\text{number of pixels in image}} \)

This implementation is not single-pass since during each iteration, all wavelet data is quantized and compressed. A single-pass algorithm might be based, for example, on a Gaussian subband model. Let \( n \) index quantization bins, \( p_n(s) \) represent the probability of finding a quantized coefficient in bin \( n \), and \( N(s) \) denote the number of coefficients in each subband:

1. iiiia) compute discrete distribution \( \{p_n(s)\} \) from \( Q(s) \) and subband variance \( \sigma_s \) \( \forall s \) (statistics are only calculated during the first iteration)
2. iiib) compute subband rate \( R(s) \) from \( \{p_n(s)\} \) \( \forall s \) and set \( R_{\text{estimate}} = \frac{\sum N(s) R(s)}{\sum N(s)} \)

Only one pass over the data is required to compute the variance of each subband; subsequent iterations of step iii) can be evaluated instantaneously.

One important observation is that the computational cost of calculating a statistic for an entire subband is roughly the same as the cost of calculating that statistic for every block in a group that collectively makes up the band (as long as the number of coefficients in each block dominates the number of blocks in the band). In either case, every coefficient is involved in an accumulation process, but the second method involves more division operations. Though there is no increase in accuracy without an increase in computation and storage, the amount of additional resources required to keep track of the local statistics (as opposed to subband-wide statistics) is negligible in comparison to the cost of computing the wavelet transform. Formalizing this idea, step iii) would be expanded as follows. Let \( k \) index the blocks in a subband, \( M(s) \) denote the number of coefficients in a block in subband \( s \):

1. iiiia) compute discrete distribution \( \{p_n(s,k)\} \) from \( Q(s) \) \( \forall k, s \)
2. iiib) compute block rate \( R(s,k) \) from \( \{p_n(s,k)\} \)
\( \forall k \) and set \( R(s) = \frac{M(s) \sum_{k} R(s,k)}{N(s)} \) \( \forall s \) 

\( \text{iii} \text{c) compute } R_{\text{estimate}} = \frac{\sum_{s} N(s) \cdot R(s)}{\sum_{s} N(s)} \) 

Two types of methods can be used for step \( \text{iii} \text{a)\), empirical and model-based methods. Each subband-block distribution may be determined with an empirical histogram after quantization by \( Q(s) \), but this technique is not single-pass. (Note that the cost of computing a histogram for each block in the subband is almost the same as computing one histogram for the whole subband.) A more computationally amenable alternative is to generate one set of high-resolution histograms, and approximate the histograms associated with each step-size by “quantizing” the high-resolution histogram. In this kind of implementation, however, accuracy is determined by the resolution of the initial histograms.

This paper examines model-based methods. Probabilistic subband model-based methods are inherently single pass. Since such models describe the data in terms of block-level statistics, fitting the model only requires computing these statistics once. From a model, it is possible to derive probabilities of quantized coefficients, and therefore first order entropy estimates as well. Ease of implementation is a key advantage of model-based methods, since keeping track of several scalar block parameters is simpler than storing block-level histograms.

Simoncelli proposed a generalized Laplacian subband model for noise removal [9]. The same type of model is considered here for rate estimation. Each subband of quantized coefficients is considered to be composed of several blocks of quantized generalized Laplacian processes. Once the block statistics required to fit this model have been computed, rate estimates based on these statistics can be used in a bisection search to determine the set of quantization step-sizes that will generate a compressed image at the target rate.

### III. Quantized Generalized Laplacian Modeling

Gormish and Gill outline a quantized Laplacian model in [3]. This model is based on a “regular” Laplacian distribution of the form:

\[
p(x) = \frac{1}{C} e^{-|x|/\sigma} \quad (1)
\]

A generalized Laplacian process is characterized by a distribution function of the following form:

\[
p(x) = \frac{1}{C} e^{-|x|^p/\sigma} \quad (2)
\]

Let \( \kappa \) be kurtosis. The parameter \( p \) determines the peakedness of the distribution. Since it can be shown for such a process that \( \kappa = \frac{\Gamma(\frac{1}{p})\Gamma(\frac{3}{p})}{\Gamma(\frac{2}{p})^2} \), where \( \Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt \), \( p \) can be determined numerically. \( S \) is directly computed with \( S = \sigma\sqrt{\frac{\Gamma(\frac{3}{p})}{\Gamma(\frac{2}{p})}} \), where \( \sigma^2 \) denotes variance. To normalize the distribution, \( C = \frac{2\Gamma(\frac{1}{p})}{p} \).

Wavelet subband data quantized with a high resolution quantizer often can be modeled with a continuous distribution, since the empirical distribution is approximately smooth. A low rates, however, this property does not hold. Therefore, a quantized version of a continuous model is used to capture subband characteristics. Quantizing data of distribution \( p(x) \) with step-size \( Q \) yields a discrete distribution \( \{p_n\} \), where each \( p_n \) is given by

\[
p_n = \int_{Q(n-\frac{1}{2})}^{Q(n+\frac{1}{2})} p(x)dx. \quad (3)
\]

Clearly, \( p_n \) may be calculated with a direct sum. If the distribution \( p(x) \) is a generalized Laplacian, however, this equation can also be expressed in the form

\[
p_n = \frac{S}{CP} \left( \Gamma(\frac{Q(n-\frac{1}{2})}{S}, \frac{1}{p}) - \Gamma(\frac{Q(n+\frac{1}{2})}{S}, \frac{1}{p}) \right), \quad (4)
\]

where \( \Gamma(x,a) = \int_x^\infty t^{a-1}e^{-t}dt \) denotes the upper in-complete Gamma function, implemented in many computational packages. Furthermore, the speed of the rate-control procedure can be optimized with a fast implementation of \( \Gamma(x,a) \).

### IV. Results

A Tarp-filter-based [10] coder is used as a baseline to evaluate the proposed rate-estimation/control technique. This coder was chosen for simplicity and performance; it yields compression ratios that are competitive with JPEG-2000. In addition, it can be easily used to evaluate per subband performance, since bands are compressed independently. Compression experiments use an ensemble of five 512x512 8 bit greyscale images with a 5-level 9/7 wavelet transform. The images are compressed using a mean-squared-error distortion measure to rates in the range 0.01-0.5 bpp. The resulting compressed rates are compared with rate estimates predicted (from quantization step-sizes) by local quantized Laplacian models, on a per image and per subband basis.

Given a set of step-sizes, a discrete probability distribution is generated for every block in each subband using (4). The rate of a block is thus estimated as the entropy of the corresponding distribution. The estimated rate of each subband is computed as the average estimated rate of all the blocks in the subband. In turn, the estimated rate of the image is computed as a weighted average of estimated subband rates, where the weights are
Fig. 1. Accuracy of Laplacian (left) and generalized Laplacian (right) rate estimates v. step-size and block-size.

Fig. 2. Error fraction of model-based per image rate estimates (using $16 \times 16$ blocks) v. step-size (left) and rate (right).

proportional to subband size. Since each model breaks subbands up into blocks of $M$ coefficients, this same experiment is conducted using a range of block sizes. Rate estimate accuracy is measured with error fraction:

\[
\text{error fraction} = \frac{|R_{\text{estimate}} - R_{\text{target}}|}{R_{\text{target}}} \tag{5}
\]

All of the rate estimate accuracy results presented represent an ensemble average.

**A. Laplacian v. Generalized Laplacian**

Figure 1 compares the performance of Laplacian modeling with that of generalized Laplacian modeling. As expected, the generalized version gives more accurate estimates, especially at low rates (i.e. high quantization step-sizes). This behavior is observed even for large subband block sizes. Note that a block size of $256 \times 256$ corresponds to a model where each subband is modeled by one random process. (If the size of a subband is smaller than a block, one process models the band.) When the size of a block is smaller than $16 \times 16$, the model-based rate estimation process over-models the subbands, producing inaccurate rate estimates. This trend is illustrated by the abrupt upward curves at the edges of these surfaces.

Figure 2 illustrates these estimates for a $16 \times 16$ block size on one plot, along with the estimates generated by a subband-wide Gaussian model, as well as the same comparison as a function of rate. The Gaussian model is less accurate. In particular, the local generalized Laplacian estimate yields an error fraction that is 4-7 times smaller than the Gaussian counterpart. This behavior is expected, since the localized Laplacian models more accurately characterize quantized coefficients, in terms of their spatial characteristics and the distributions they form. Observe that for rates lower than 0.5 bpp, the generalized Laplacian estimate is within five percent of the true compressed size.
B. Subband and Image Rate Estimates

The accuracy of subband rate estimates and overall image rate estimates are compared in Figure 3. These plots illustrate model performance using $64 \times 64$ blocks, as well as several important trends. First, at low rates, the per subband accuracy of the regular Laplacian estimates is considerably worse than that of the generalized Laplacian estimates. Second, these plots show that local empirical rate estimates are not always more accurate than local model-based estimates. Finally, at low rates, the per image Laplacian estimates are much more accurate than the per subband estimates. This observation indicates a type error cancellation, more formally expressed in the following way.

Using the notation from section II, let $R_{actual}$ equal the rate of a compressed image, where each subband has been quantized with step-size $Q(s)$ and compressed to a rate $R^s(s)$. Using the notation from section II,

$$R_{actual} = \frac{\sum_s N(s) \cdot R^s(s)}{\sum_s N(s)}.$$

Recall $R(s)$ denotes the estimated rate of subband $s$ quantized with $Q(s)$. Let $\epsilon(s) = R^s(s) - R(s)$ denote the difference between the actual rate and estimated rate of subband $s$. Then $R_{actual}$ can be expressed as

$$R_{actual} = R_{estimate} + \frac{\sum_s N(s) \cdot \epsilon(s)}{\sum_s N(s)}.$$

Though this type of model overestimates and underestimates the entropy of quantized subbands, on average, the overall image rate estimates are quite accurate. Because the $\epsilon(s)$ error terms vary in sign, cancellation occurs in the summation above, and $R_{estimate}$ is close to $R_{actual}$, despite subband-level inaccuracies.

C. Application: Distortion-Contrast-Based Rate Control

Distortion-contrast quantization (DCQ) is a perceptual lossy compression scheme, where wavelet subbands are quantized to induce a desired level of contrast in the distortion image [2]. The step-sizes associated with each subband vary as an image/subband-specific function of visual quality (and thus rate). Quantized generalized Laplacian rate-estimation can be used for rate-control in a perceptual compression scheme based on this kind of quantization. A modified version of the rate control algorithm outlined in section 2 is used to determine the perceptual step-sizes that will yield a compressed image at the desired rate. In particular, a bisection search is performed on a quality parameter that is used to derive a vector of step-sizes (one for each subband), instead of on actual step-sizes.

Ideal step-sizes associated with a target rate are determined with a bisection search using a Tarp-based coder for rate-estimation. Next, the local generalized quantized Laplacian model-based approach is used (with $16 \times 16$ blocks) to determine step-sizes based on the same target rate. Figure 4 and depicts one such example, where 

harbour

is coded at 0.25 bpp. The model-based approach generates an image at 0.2529 bpp, yielding an error fraction of 0.0116. Clearly, ideal rate-control and the model-based algorithm produce visually similar images.

The performance of this algorithm is averaged over ten $512x512$ 8 bit greyscale images, five of which were used to generate the error fraction surfaces in Figure 1. The results are illustrated in Figure 5. This experiment reveals that the choice of block size, derived from the error fraction surface in Figure 1, is appropriate for a wider class of images. Furthermore, it confirms that the algorithm still generates accurate rate estimates when each subband is quantized with a different step-size.
Fig. 4. *harbour*, coded with distortion-contrast step-sizes derived from ideal (left) and model-based (right) rate control.

Fig. 5. Error fraction of generalized-Laplacian-based rate estimates generated from distortion-contrast step-sizes v. rate.

V. CONCLUSION

A single-pass method for entropy estimation based on localized quantized generalized Laplacian modeling is presented. This approach treats wavelet subbands as small blocks of quantized random processes, and generates block-level entropy estimates based on quantization step-sizes. The method estimates compressed image size to within five percent of the true compressed size (resulting from state-of-the-art wavelet coding) at low rates, for a variety of quantization schemes. Coupled with bit plane truncation, this kind of entropy estimation can be used for rate-control. Due to its computational simplicity, this method is suitable for rate-constrained real-time imaging applications.

REFERENCES


