Context-Based Lossless Interband Compression - Extending CALIC

Xiaolin Wu
Dept. of Computer Science,
The Univ. of Western Ontario,
London, Ontario, Canada N6A 5B7.

Nasir Memon*
Computer Science Department
Polytechnic University
Six MetroTech Center
Brooklyn, NY 11201

Abstract

This paper proposes an inter-band version of CALIC (Context-based, Adaptive, Lossless Image Codec) which represents one of the best performing, practical and general purpose lossless image coding techniques known today. Inter-band coding techniques are needed for effective compression of multi-spectral images like color images and remotely sensed images. It is demonstrated that CALIC’s techniques of context modeling of DPCM errors lend themselves easily to modeling of higher-order inter-band correlations that cannot be exploited by simple inter-band linear predictors alone. The proposed inter-band CALIC exploits both inter-band and intra-band statistical redundancies, and obtains significant compression gains over its intra-band counterpart. On some types of multispectral images, inter-band CALIC can lead to a reduction in bit rate of more than 20% percent as compared to intra-band CALIC. Inter-band CALIC only incurs a modest increase in computational cost as compared to intra-band CALIC.

EDICS Category: 1.1 - Still Image Coding

*Corresponding Author. Email: memon@poly.edu, Phone: 718-260-3970, Fax: 718-260-3609
1 Introduction

Most images in modern visual communications and computing are multispectral in nature. Color images, that consist of three (eg. RGB, YIQ, YUV, LAB, HSV, etc) or four (eg. CMYK) separate color bands, are the simplest form of multispectral images. A satellite image can have hundreds of bands. For example the High-Resolution Imaging Spectrometer (HIRIS) to be placed on the Earth Observing System (EOS) scheduled to be launched by NASA within this decade, is designed to acquire images with 192 spectral bands at 30m spatial resolution with resulting bit-rates of more than 200 Megabits per second.

Much of the work done towards developing algorithms for compressing image data has focused on lossy compression where “perceptually transparent” distortions are introduced in the reconstructed image. Perceptual coding is indeed an important activity and the future will see further developments in this area. However, perceptual coding techniques only apply when the end user is an “average” human viewing the reconstructed image. In many applications, the end user is not a human viewer. Instead, the reconstructed image is subjected to some processing based on which “meaningful” information is extracted. This is especially true for remotely sensed images that are often subject to processing in order to extract ground parameters of interest. For example, remotely sensed microwave emission images of the Arctic are used to obtain a number of different parameters including ice concentration and ice type. In such a case, an earth scientist is concerned about the effects of distortions introduced by the compression algorithm on the resulting ice concentration estimates. Hence, in remote sensing applications, scientists typically prefer not to use lossy compression and use lossless compression instead. Indeed, remote sensing is one of the most important application areas of lossless image compression. Other applications that often require lossless compression include medical imaging, pre-press imaging and archiving.

Although lossless image compression has been studied for decades, the past couple of years have witnessed a renewed flurry of activity in this area due to the standardization activities being carried out by the JPEG committee [9]. One result of this activity has been
the development of CALIC, a Context-based Adaptive Lossless Image Codec designed by the authors [18], that to the best of our knowledge, gives superior compression performance, on the average, over currently known practical and general purpose lossless compression techniques \(^1\). However, CALIC is an intra-band compression technique and has no mechanism for exploiting inter-band correlations found in multispectral data. Lossless inter-band image coding is, in some sense, more challenging as it demands a more thorough statistical modeling of the data. Inter-band coding techniques that exploit statistical correlations between and within image bands are needed to obtain significantly higher compression than intra-band compression alone.

In this paper we incorporate new inter-band prediction and context modeling techniques into the framework of CALIC [18], and develop an inter-band lossless image coder called inter-band CALIC. On many types of multispectral images, inter-band CALIC can reduce the bit rate of intra-band CALIC by more than 20% percent, with a modest increase in computational cost. Although we focus on lossless compression, any improvement in lossless coding efficiency could also benefit lossy inter-band image compression, since the latter can be derived from the former with a suitable quantization scheme.

The paper is structured as follows. In Section 2, we first prepare the groundwork for the proposed inter-band CALIC technique, by presenting a brief sketch of intra-band CALIC. In Sections 3, 4, 5, and 6 we show how various elements in CALIC can be generalized to lossless inter-band image coding. In Section 7 we present encouraging compression results obtained by inter-band CALIC and compare with previous results published in the literature.

2 A Brief Overview of CALIC

A major difficulty with statistical modeling of continuous-tone images is the large alphabet size, typically 256 or larger. Context modeling of alphabet symbols (pixel values) leads to an

\(^1\)It should be noted that the key design goal behind CALIC was high compression at reasonable complexity. There are other techniques in the literature, especially JPEG-LS, that aim for low complexity and yet achieve compression performance within a few percentagey points of CALIC
intractably large number of possible model states (contexts). This is more than a problem of high time and space complexity for modeling. If the number of model states is too large with respect to the size of the image, one may not have enough samples to reach good estimates of conditional probabilities within each model state, leading to poor coding efficiency. This is known as “the sparse context” [4] or “context dilution” problem [15]. The problem was theoretically formulated by Rissanen in the framework of stochastic complexity as “model cost” [11]. Intuitively speaking, model cost refers to either the amount of side information necessary to describe a source model if the model is explicitly sent to the decoder, or to the loss in coding efficiency before an accurate model is “learnt” on the fly from previously coded symbols. In the latter case, both encoder and decoder fit an adaptive model to an unknown source via an on-line learning process, relying on no side information nor prior knowledge. In information theory literature this principle is called universal source modeling [15, 11] and it is this framework on which CALIC is based.

A pivotal issue in lossless image coding is how to reduce the model cost, or how to achieve fast model convergence to optimal compression. The reader will be frequently brought to this issue in the following developments.

2.1 Overview of Intra-frame CALIC

Intra-frame CALIC encodes and decodes an image in raster scan order with a single pass through the image. For the purposes of context modeling and prediction, the coding process uses a neighborhood of pixel values taken only from the previous two rows of the image. Consequently, the encoding and decoding algorithms require a buffer that holds only two rows of pixels that immediately precede the current pixel. In Fig. 1 we give a schematic description of CALIC (the figure depicts only the encoding process; decoding is achieved by the reverse process).

As shown in Fig. 1, CALIC operates in two modes: binary mode and continuous-tone mode. This allows CALIC to distinguish between binary and continuous-tone types of im-
ages on a local, rather than a global, basis. This distinction between the two modes is important due to the vastly different compression methodologies employed within each mode. In continuous-tone mode predictive coding is used, whereas in binary mode, CALIC codes pixel values directly. Selection between the two modes is done depending on whether or not the local neighborhood of the current pixel has more than two distinct pixel values. This two-mode design contributes to the universality and robustness of CALIC over a wide range of images.

In continuous-tone mode, the system has four major integrated components: prediction, context selection and quantization, context-based bias cancellation of prediction errors, and conditional entropy coding of prediction errors. In the prediction step, a gradient-adjusted prediction (GAP) \( \hat{y} \) of the current pixel \( y \) is made. A detailed description of GAP is presented in Section 3 where we present new prediction techniques for inter-band prediction.

The predicted value \( \hat{y} \) is further adjusted via bias cancellation procedure that involves an error feedback loop of one-step delay. The feedback value is the sample mean of prediction errors \( \bar{e} \) conditioned on the current context. This results in an adaptive, context-based, non-linear predictor \( \tilde{y} = \hat{y} + \bar{e} \). In Fig. 1, these operations correspond to the block “bias cancellation”. In section 4 we describe these steps in more detail and show how they can be extended to take into account inter-band correlations.

The bias corrected prediction error \( \tilde{y} \) is finally entropy coded based on few estimated conditional probabilities in different conditioning states or coding contexts. A small number of coding contexts are generated by context quantization. The context quantizer partitions prediction error terms into few classes by the expected error magnitude. The details of this context quantization scheme in association with conditional entropy coding are given in Section 5.

Finally, Section 6 describes the context-based ternary arithmetic coding scheme used in the binary mode of CALIC and its inter-band extension. The arithmetic coder is used to code three symbols, including an escape symbol.
3 Correlation based switched inter/intra-band prediction

When designing a compression technique that exploits inter-band correlations, the question of band selection arises. That is, which band(s) is (are) the best to use as reference band(s) for predicting and modeling intensity values in a given band. In [13], it was shown that significant compression benefits can be obtained by re-ordering the bands of multi-band images, prior to prediction. The problem of computing an optimal ordering was formulated in a graph theoretic setting, admitting an $O(N^2)$ solution for an $N$-band image. For clarity of exposition we assume that such an ordering has been determined and the bands have been appropriately re-ordered such that the previous band is used as a reference band for encoding the current band.

There is one more potential complication. The different bands in a multi-band image may be represented one pixel at a time (pixel interleaved), one row at a time (line interleaved), or an entire band at a time (band sequential). Since the coder needs to utilize at least one band (called reference band in this paper) in order to make compression gains on other bands, buffering strategies and requirements would vary with the different representations. For example, the third option would require the entire reference band to be buffered, which may be excessive for very large images. In practice, the techniques we propose in this paper for inter-band modeling and coding use only a small three-dimensional causal neighborhood around the current pixel and need to buffer only a few rows of the reference band at any given time. Therefore, for ease of exposition, in the rest of this paper we assume that the required neighboring pixels from the reference band, used for inter-band modeling and coding, are available as needed. We do this, knowing that it is not difficult to buffer modeling contexts and handle file I/O no matter whether the input multispectral image is pixel, row, or band interleaved. Furthermore, by restricting our prediction and context modeling techniques to a causal neighborhood we ensure that proposed techniques can be utilized with any of the three interleaving orders described above.
We denote by $X[\cdot, \cdot]$ and $Y[\cdot, \cdot]$ the reference and current band respectively, with $y = Y[i, j]$ being the pixel to be coded next. In context modeling of $y$ we cannot escape from the fact that the number of modeling states, even with context quantization, grows exponentially in the size of modeling contexts, while the number of samples only grows linearly from intra-band to inter-band coding. Clearly, context-based inter-band lossless image compression is more susceptible to context dilution problem than its intra-band counterpart. Furthermore, we would like not to significantly increase time and space complexities from intra-band to inter-band coding. Given these two considerations, we resort to predictive inter-band coding in which context modeling of prediction errors rather than pixels of multispectral images is performed.

### 3.1 Interband prediction

It is easy to generalize a DPCM-like predictor from two-dimensional to three-dimensional sources. Namely, we predict $y = Y[i, j]$ to be

$$
\hat{y} = \sum_{(a, b) \in N_1} \theta_{a,b} Y[i - a, j - b] + \sum_{(a', b') \in N_2} \theta'_{a', b'} X[i - a', j - b'],
$$

where $N_1$ and $N_2$ are appropriately chosen neighborhoods that are causal with respect to the scan and the band interleaving being employed. The coefficients $\theta_{a,b}$ and $\theta'_{a',b'}$ can be optimized by standard linear regression techniques to minimize $\|Y - \hat{Y}\|$ over a given multispectral image. However, since multispectral images are seldom stationary, optimizing $\theta$ and $\theta'$ over the entire image is ineffective. Computing an optimal least-square multispectral predictor for different image segments does not necessarily improve coding efficiency despite the high computational costs involved. This is because frequently changing prediction coefficients incur too much side information (high model cost).

Alternatively, we consider a correlation-driven adaptive predictor whose parameters are uniquely determined by the previously coded pixels without using any side information. Let $X = (x_1, x_2, \ldots, x_8)$ and $Y = (y_1, y_2, \ldots, y_8)$ be two random vectors whose elements are causal neighbors of the current pixel $y$ in the reference and current bands respectively as
illustrated by Fig. 2. The pixels $x_i$ and $y_i$ belong to the same spatial location but in different bands. Now, if the correlation coefficient

$$
\varphi(X, Y) = \frac{8 \sum_{i=1}^{8} x_i y_i - \sum_{i=1}^{8} x_i \sum_{i=1}^{8} y_i}{\sqrt{8 \sum_{i=1}^{8} x_i^2 - (\sum_{i=1}^{8} x_i)^2 \left[8 \sum_{i=1}^{8} y_i^2 - (\sum_{i=1}^{8} y_i)^2\right]}}
$$

is high, then we have strong similarity of the two-dimensional waveforms between the two bands in the locality of interest. In this case, the local waveform of the reference band can be projected to the current band in order to predict the current pixel $y$. Specifically, we could compute

$$
\hat{Y} = \alpha X + \beta
$$

such that $\|\hat{Y} - Y\|_2$ is minimized. It is easy to show that

$$
\alpha = \frac{8 \sum_{i=1}^{8} x_i y_i - \sum_{i=1}^{8} x_i \sum_{i=1}^{8} y_i}{8 \sum_{i=1}^{8} x_i^2 - (\sum_{i=1}^{8} x_i)^2} \quad \text{and} \quad \beta = \frac{8 \sum_{i=1}^{8} y_i - \alpha \sum_{i=1}^{8} x_i}{8}.
$$

Clearly, if $\varphi(X, Y)$ is sufficiently high, then

$$
\hat{y} = \alpha x + \beta
$$

is a good inter-band predictor of $y$. However, if there is a sharp edge at $y$, $\hat{y}$ may not be a good prediction of $y$. Hence we adapt the prediction $\hat{y}$ to the local gradients around $y$. To this end we use the approximations

$$
y - y_1 \approx \hat{y} - \hat{y}_1 = \alpha (x - x_1) \quad \text{and} \quad y - y_2 \approx \hat{y} - \hat{y}_2 = \alpha (x - x_2)
$$

(6)

to get the predictors

$$
\hat{y}_h = y_1 + \alpha (x - x_1)
$$

and

$$
\hat{y}_v = y_2 + \alpha (x - x_2),
$$

(7)

with $\hat{y}_h$ and $\hat{y}_v$ being extrapolations based on horizontal and vertical gradients computed in the reference band. It is intuitively expected and was verified by our experiments that $\hat{y}_h$ is
a better predictor than \( \hat{y}_v \) if \( |x - x_1| \ll |x - x_2| \) and vice versa. This is due to the fact that when an edge is detected at the current location in the reference band, it is highly likely to also occur in the current band. Therefore, we weigh \( \hat{y}_h \) and \( \hat{y}_v \) to get a gradient-adjusted inter-band predictor:

\[
\hat{y} = \frac{|x - x_2|\hat{y}_h + |x - x_1|\hat{y}_v}{|x - x_2| + |x - x_1|}. \tag{8}
\]

In practice we save on some computation by simplifying (8) to the following procedure with negligible loss in compression.

\[
\text{IF } (|x - x_2| - |x - x_1| > T) \hat{y} = \hat{y}_h; \{\text{sharp horizontal edge}\}
\]

\[
\text{ELSE IF } (|x - x_2| - |x - x_1| < -T) \hat{y} = \hat{y}_v; \{\text{sharp vertical edge}\}
\]

\[
\text{ELSE } \hat{y} = (\hat{y}_h + \hat{y}_v)/2,
\]

where \( T \) is a threshold.

It was observed in our experiments that the inter-band predictor introduced above out-performs comparable intra-band predictors if \( \varphi(X,Y) \) is high. This is because the gradient at the pixel being currently coded is known in the reference band but unknown in the current band. A high inter-band correlation \( \varphi(X,Y) \) indicates a high confidence level in the approximations \( y - y_1 \approx \alpha(x - x_1) \) and \( y - y_2 \approx \alpha(x - x_2) \).

### 3.2 Switching to intra-band prediction

The degree of correlation between bands can vary drastically from region to region in a multispectral image. For instance, for color images the inter-band correlation decreases as the color saturation increases. For remotely sensed images, correlation between a given pair of bands will vary across the spatial extent of the image according to the ground conditions in the scene and the properties of the objects being imaged. Given this fact, it is counterproductive to use inter-band prediction in cases of weak inter-band correlation. Instead, we can improve coding efficiency by switching between inter-band and intra-band predictors according to the value of inter-band correlation \( \varphi(X,Y) \) in a local window. Indeed, in our
experiments this strategy significantly outperforms unconditional inter-band predictors designed by linear regression [14] and by KL-transform [2]. Specifically, if $\varphi(X, Y) > 0.5$ then we use the proposed inter-band predictor $\hat{y}$; otherwise we use an intra-band predictor $\tilde{y}$. We call this a correlation-based switched inter/intra-band predictor.

For intra-band prediction we use the Gradient Adjusted Predictor (GAP), used in intra-band CALIC [18]. GAP first computes estimates $d_h$ and $d_v$ of the horizontal and vertical gradient magnitudes near the current pixel $y$ given by (refer to Figure 2),

$$d_h = |y_5 - y_1| + |y_3 - y_2| + |y_2 - y_4|, \quad d_v = |y_6 - y_2| + |y_3 - y_1| + |y_7 - y_3|, \quad (9)$$

Having computed $d_h$ and $d_v$, the coder estimates the strength and orientation of edges in the current band, and then predicts the current pixel $y$ by weighting neighboring pixels according to the estimated gradients. This aims to alleviate the adverse effect of edges on DPCM-type predictors. Specifically, the intra-band predictor $\hat{y}$ is computed according to the following procedure.

\[
\begin{align*}
\text{IF} & \ (d_v - d_h > T_1) \ \{ \text{sharp horizontal edge} \} \ \hat{y} = y_1 \\
\text{ELSE IF} & \ (d_v - d_h < -T_1) \ \{ \text{sharp vertical edge} \} \ \hat{y} = y_2 \\
\text{ELSE} & \ \{ \hat{y} = (y_1 + y_2)/2 + (y_4 - y_3)/4; \\
& \ \text{IF} \ (d_v - d_h > T_2) \ \{ \text{horizontal edge} \} \ \hat{y} = (\hat{y} + y_1)/2 \\
& \ \text{ELSE IF} \ (d_v - d_h > T_3) \ \{ \text{weak horizontal edge} \} \ \hat{y} = (3\hat{y} + y_1)/4 \\
& \ \text{ELSE IF} \ (d_v - d_h < -T_2) \ \{ \text{vertical edge} \} \ \hat{y} = (\hat{y} + y_2)/2 \\
& \ \text{ELSE IF} \ (d_v - d_h < -T_3) \ \{ \text{weak vertical edge} \} \ \hat{y} = (3\hat{y} + y_2)/4 \}
\end{align*}
\]

where $T_1, T_2, T_3$ are fixed thresholds. In practice $T_1 = 80, T_2 = 32$ and $T_3 = 8$ seem to work well for 8 bit images.

### 3.3 Time complexity

The predictor coefficients and thresholds given above were empirically chosen. A major criterion in choosing these coefficients was ease of computation. Also note that efficient
incremental and/or parallel schemes for evaluating \(d_h\) and \(d_v\) in a row-by-row scan of images are straightforward. The most expensive part of the proposed inter/intra-band prediction scheme is the computation of \(\varphi(X, Y)\). But we can eliminate the multiplications for terms \(x_iy_i\), \(x_i^2\), and \(y_i^2\) in (2) by using a \(256 \times 256\) multiplication table for all possible products \(uv\), \(0 \leq u, v < 256\), and a \(256\)-entry square table for all possible \(u^2\), where \(u\) and \(v\) are pixel values. We can reduce the sizes of the tables to \(2^{2m}\) and \(2^m\) by using only the \(m\) most significant bits of \(u\) and \(v\). In practice we found that using \(m = 6\) does not affect compression performance. Furthermore, since the coder only uses \(\varphi(X, Y)\) to decide if it should use inter-band or intra-band predictor, we change the condition \(\varphi(X, Y) > 0.5\) to \(\varphi^2(X, Y) > 0.25\) to avoid the square root operation in (2). We also choose a window of eight pixels to compute \(\varphi^2(X, Y)\) to turn a few more multiplications to bit shifting. If \(\varphi^2(X, Y) > 0.25\) the coder needs to compute \(\alpha\). By comparing (2) and (4) we see that \(\alpha\) can be computed, if necessary, by only one extra division since both the numerator and denominator of (4) also appear in (2). For modern computers, the time complexity of the proposed inter/intra-band prediction scheme is reasonable. However, we would again like to emphasize that inter-band CALIC, like its intra-band counterpart has been designed with the goal of achieving high compression performance and is necessarily complex in comparison to techniques that focus on simplicity.

4 Context-based Bias Cancellation

Local gradients alone cannot adequately characterize some of the more complex relationships between the predicted pixel \(y\) and its surrounding. We can improve coding efficiency by context modeling of the prediction error \(e = y - \hat{y}\) prior to entropy coding. Conditioning of the prediction error to its context can exploit higher-order structures such as texture patterns and local activity in the image for further compression gains. However, the large number of possible contexts can lead to the ‘sparse context’ or ‘high model cost’ problem mentioned before. CALIC employs an effective solution for this problem. Instead of estimating the pdf of prediction errors, \(p(e|C)\), within each context \(C\), only its conditional expectation \(E\{e|C\}\)
is estimated using the corresponding sample means $\bar{\epsilon}(C)$. These estimates are then used to further refine the prediction prior to entropy coding, by an error feedback mechanism that cancels prediction biases in different contexts. We call this process bias cancellation. Bias-cancellation leads to an improved predictor for $y$: $\hat{y} = \hat{y} + \bar{\epsilon}(C)$, where $\bar{\epsilon}(C)$ is the sample mean of $\epsilon$ conditioned on the context $C$.

Bias cancellation is really not new to CALIC but has appeared in various forms in prior literature, for example in [7, 5] and [15]. Conceptually, bias cancellation can also be viewed as a two-stage adaptive prediction scheme via conditioning of prediction errors to contexts and the subsequent error feedback. Hence contexts used for bias cancellation have are also called prediction contexts.

In interband CALIC we adopt this same bias cancellation technique. The new and important issue when extending from intra-band to interband coding is the identification of modeling event(s) involving the reference band that reveal additional prediction error structure, that would otherwise be hidden in intra-band coding, and use them in a way of low model cost. Note, context error modeling should be done separately for the interband predictor $\hat{y}$ and intra-band predictor $\hat{y}$, because the two predictors rely on different contexts, and the prediction errors in each case behave very differently.

### 4.1 Context formation for interband predictor

If $\varphi(X,Y) > 0.5$, then the interband predictor $\hat{y}$ is used. In this case the robustness of $\hat{y}$ increases as $\varphi(X,Y)$. Clearly, the prediction error $e = y - \hat{y}$ is statistically related to $\varphi(X,Y)$; smaller $|e|$’s are more likely to be associated with higher $\varphi(X,Y)$. Thus $\varphi(X,Y)$ is an important observable feature that can reveal much about $e$. It is intuitive and has also been verified by our experiments that there is a quite a strong correlation between $|e|$ and

$$\hat{d} = \alpha(|x - x_1| + |x - x_2|),$$

with $\varphi(|e|, \hat{d}) \approx 0.5$, where $\hat{d}$ is a least-square estimator of $|y - y_1| + |y - y_2|$. This is because $\hat{d}$ can be viewed as a measure of image activity level around the current pixel and a larger $\hat{d}$
leads to a larger expectation of $|e|$. In addition, the correlations $\varphi(|e|, |e_1|)$ and $\varphi(|e|, |e_2|)$ are also significant, being around 0.4, where $e_1$ and $e_2$ are the prediction errors at coded pixels $y_1$ and $y_2$ respectively. Therefore, we include $\hat{d}$, $|e_1|$, and $|e_2|$ in context modeling of $e$ as well. Furthermore, since $e$ tends to behave differently in different texture patterns $T$ (waveforms) surrounding $y$, we choose $T = \{y_1, y_2, \ldots, y_6\}$, the six closest immediate neighbors of $y$ (see Figure 2) as another informative feature of $e$.

By combining all the above potentially useful features in context modeling, we would like to drive an entropy coder for $e$ with conditional probability $P(e \mid \varphi, \hat{d}, |e_1|, |e_2|, T)$. But $P(e \mid \varphi, \hat{d}, |e_1|, |e_2|, T)$ is generally unknown. Learning $P(e \mid \varphi, \hat{d}, |e_1|, |e_2|, T)$ on the fly incurs too high a model cost to benefit compression. Even a high-resolution multispectral image does not provide enough samples for a reliable estimate of $P(e \mid \varphi, \hat{d}, |e_1|, |e_2|, T)$ in a 5-dimensional modeling space. Hence we quantize the modeling space in order to drastically reduce the number of modeling contexts.

First we quantize the random variable $\varphi(X, Y)$ into two levels at threshold $q$ such that the conditional entropy

$$P(\varphi \leq q)H(e \mid \varphi \leq q) + P(\varphi > q)H(e \mid \varphi > q)$$  \hspace{1cm} (11)

is minimized. This can be done via a linear search for a given image or set of test images. In our experiments we chose a fixed $q$ based only on empirical observations for a few test images.

Next we use a linear combination of $\hat{d}$, $|e_1|$, and $|e_2|$

$$\Delta = a\hat{d} + b|e_1| + c|e_2|$$  \hspace{1cm} (12)

to reduce the dimensionality of the modeling space. We let $\Delta$ be the least-square estimate of $|e|$ in terms of $\hat{d}$, $|e_1|$, and $|e_2|$. In other words, the coefficients $a$, $b$, and $c$ were chosen via linear regression to minimize $\|\Delta - |e|\|_2$. By combining the three features $\hat{d}$, $|e_1|$, and $|e_2|$ into a single one, namely $\Delta$, we reduce vector quantization of $\hat{d}$, $|e_1|$, and $|e_2|$ into scalar
quantization of $\Delta$. The scalar quantizer can be efficiently optimized under an entropy criterion via dynamic programming [17], whereas a similar optimization becomes NP-complete in the Vector Quantization (VQ) case. Specifically, denoting by $Q_\Delta = (\delta_1, \delta_2, \ldots, \delta_{K-1})$ a $K$-level scalar quantizer of $\Delta$, we can optimize $Q_\Delta$ to minimize the conditional entropy

$$
\sum_{0 \leq k < K} P(\delta_k < \Delta \leq \delta_{k+1}) H(e \mid \delta_k < \Delta \leq \delta_{k+1}),
$$

where $\delta_0 = 0$ and $\delta_K = \infty$. In practice we chose $K = 4$. Larger $K$ only makes negligible improvement on coding efficiency.

Finally, we quantize the texture pattern $\mathcal{T}$ into a bit pattern $B = b_6 b_5 \cdots b_1$, where

$$
b_k = \begin{cases} 
0 & \text{if } y_k \geq \hat{y} \\
1 & \text{if } y_k < \hat{y}
\end{cases}, \quad 1 \leq k \leq 6,
$$

and $y_i$ are as shown in Figure 2. For convenient description, we denote the quantizer outputs of (11), (13), and (14) by $\rho$, $\delta$, and $\beta$, i.e., $\rho = Q_\varphi(\varphi)$, $\delta = Q_\Delta(\Delta)$, and $\beta = Q_\mathcal{T}(\mathcal{T})$. By context quantization of (11), (13), and (14), we now have 512 modeling contexts $(\rho, \delta, \beta)$. Bias cancellation is done by estimating the conditional means $\mu(e \mid \rho, \delta, \beta)$ and computing $\hat{y} = \hat{y} - \mu(e \mid \rho, \delta, \beta)$.

### 4.2 Context formation for intra-band predictor

If the local interband correlation is weak, then we switch to CALIC’s intra-band predictor. In this case intra-band context modeling of prediction errors is performed. But unlike in interband context modeling $\varphi$ is no longer considered for context modeling of prediction errors. This is because the intra-band predictor $\hat{y}$ does not relate to $\varphi$. Including $\varphi$ as a modeling event simply increases model cost without gaining additional knowledge about $e$. The texture pattern $B$ and error magnitude estimator $\Delta$ are the remaining modeling events, but some minor changes to $\Delta$ and $B$ are required from interband to intra-band context modeling. In intra-band coding we have greater uncertainty about the image waveform $\mathcal{T}$ than in interband coding due to the lack of interband correlation. In order to reduce this uncertainty, we make the quantization of $\mathcal{T}$ finer by extending quantized texture pattern $B$
from 6 bits to 8 bits, setting additional two bits \( b_7 \) and \( b_8 \) to represent whether \( \hat{y} > 2y_2 - y_6 \) and \( \hat{y} > 2y_1 - y_5 \). The two bits \( b_7 \) and \( b_8 \) characterize the convexity of the image waveform in the current window as explained in [18]. Since \( \hat{y} \) is based on \( d_h \) and \( d_v \) defined in (9) not on \( \hat{d} \), we change the least-square estimator \( \Delta \) of \( |e| \) from (13) to

\[
\Delta = a(d_h + d_v) + b|e_1| + c|e_2|,
\]

where coefficients \( a \), \( b \), and \( c \) are determined via a linear regression process in the same way as in the interband context modeling. Of course, optimal \( \Delta \) quantizer needs to be designed separately for the intra-band case under the criterion of minimum conditional entropy as we did in (13).

With the proposed intra-band context modeling techniques, the coder can then use conditional error mean \( \mu(e \mid Q(\Delta), B) \) to cancel prediction bias via error feedback, namely,

\[
\bar{y} = \hat{y} + \mu(e \mid \delta, \beta).
\]

The total number of intra-band modeling contexts is roughly the same as in interband modeling contexts, being 576 as shown in [18]. We use two extra bits in \( B \) but eliminate one bit in \( Q(\varphi) \). Furthermore, the values of \( b_7 \) and \( b_8 \) are not completely independent of \( b_1, \cdots, b_6 \), reducing the number of possible combinations of \( b_1, \cdots, b_7, b_8 \) (see [18]).

### 4.3 Optimizing model parameters

There are a few model parameters that can be optimized for specific images or specific application domains. These are coefficients \( a, b, c \) in (10) and those in (15), optimal quantizer parameters of \( Q_\Delta \) and \( Q_\varphi \). The side information required to encode these optimized model parameters is negligible compared with the savings in bit rate obtained with an optimized model. Using optimized model parameters the coder can achieve two to seven percent lower bit rates, depending on image types, than using fixed model parameters that were empirically chosen for a set of test images of various types.
For some applications the additional seven percent gained with an optimized model may not be deemed worth the added computational complexity required for optimizing model parameters. Specifically, computing \( a, b, c \) involves a linear regression process, and designing optimal quantizers \( Q_\Delta \) and \( Q_\rho \) requires dynamic programming [17]. But in at least two common situations, the extra cost of image-dependent optimization can be easily justified by the improved coding efficiency. If a multispectral image is to be encoded only once, such as in a image database, we can afford a relatively expensive off line encoder, as long as decoding is simple and fast. The optimized inter/intra-band predictor perfectly suits such asymmetrical coding systems because the decoder can simply use the optimized model parameters as side information without computing them. The other situation is domain specific multispectral image compression. In some applications, the coder has prior knowledge of the input image type. For instance, the header of a medical image sequence often tells the coder of the imaging modality, the human anatomy involved, the projection orientation, etc. We can then classify image sequences that are acquired by the same imaging device, of the same human anatomy, and in the same orientation into a single class, and customize model parameters for this class in order to maximize compression. If model parameters can be optimized for specific image classes in an off-line manner, then the optimization cost becomes irrelevant.

5 Formation of Entropy Coding Contexts

The last step of predictive coding is entropy coding of prediction errors. It seems natural to use the same modeling contexts in entropy coding as in error modeling. That is, we estimate conditional probabilities \( P(\hat{e} \mid \rho, \delta, \beta) \) and \( P(\hat{e} \mid \delta, \beta) \), depending on whether interband predictor \( \hat{y} \) or intra-band predictor \( \hat{y} \) is used, and use them to drive an arithmetic coder. But the high model costs associated with \( P(e \mid \rho, \delta, \beta) \) and \( P(e \mid \delta, \beta) \) prevent on-line count statistics from reaching reliable estimates of the underlying conditional probabilities. We have to further reduce the number of coding contexts. To benefit from the computations done in the earlier error modeling stages, we simply merge the 512 interband modeling
contexts \((\rho, \delta, \beta)\) and 576 intra-band modeling contexts \((\delta, \beta)\) into a much smaller number \(K\) of coding contexts, \(C_i, 1 \leq i \leq K\) for conditional coding.

Ideally, the context merging should be optimized to minimize the conditional entropy \(\sum_{i=1}^{K} P(C_i) H(e|C_i)\). An exact solution of this optimization problem is computationally intractable even for a specific image. Heuristically speaking, we should merge conditional probabilities that are close to one the other. The bias cancellation process tends to make the distribution of prediction errors within each context to be zero mean. However, we could use the conditional variance of \(e\) within each modeling context to classify different conditional probability functions. Fortunately, on-line count statistics can satisfactorily estimate the variance of \(P(e \mid \rho, \delta, \beta)\) and \(P(e \mid \delta, \beta)\) even with a large number of modeling contexts, just as it can estimate \(E\{e|\rho, \delta, \beta\}\) and \(E\{e|\delta, \beta\}\) in the case of error modeling. Once again estimating some parameters of \(P(e \mid \rho, \delta, \beta)\) and \(P(e \mid \delta, \beta)\) rather than the conditional error probabilities themselves greatly alleviates the problems of context dilution and excessive memory use that would otherwise limit high-order context modeling.

For ease of implementation, we replace the conditional variance by \(E\{|e| \mid \rho, \delta, \beta\}\) and \(E\{|e| \mid \delta, \beta\}\), depending on whether \(\hat{y}\) or \(\tilde{y}\) is in question. In practice, both encoder and decoder compute the sample averages of past observed \(|e|\) in different model contexts, denoted by \(\sigma(e|\rho, \delta, \beta)\) and \(\sigma(e|\delta, \beta)\), and determine the current coding context \(C_i\) by quantizing \(\sigma(\rho, \delta, \beta)\) and \(\sigma(\delta, \beta)\). Specifically,

\[
C_i = \{(\rho, \delta, \beta) : q_{i-1} < \sigma(\rho, \delta, \beta) \leq q_i \} \cup \{\delta, \beta) : q_{i-1} < \sigma(\delta, \beta) \leq q_i \}, \quad 1 \leq i \leq K, \tag{17}
\]

where \(q_0 = 0, q_K = \infty\), and the vector \(Q_\sigma = (q_1, q_2, \ldots, q_{K-1})\) defines the context quantizer of \(\sigma\). In practice we recommend \(K = 8\). That is, \(Q_\sigma\) merges 512 interband modeling contexts \((\rho, \delta, \beta)\) and 576 intra-band modeling contexts \((\delta, \beta)\) into eight coding contexts. Using more than eight coding contexts brings less than 0.3 percent improvement in compression. When \(K\) increases beyond 16 we start to observe gradually increasing code length, indicating the negative impact of model cost. As in the cases of optimizing \(Q_\rho\) and \(Q_\Delta\), we can optimize \(Q_\sigma\) via dynamic programming to minimize the conditional entropy \(\sum_{1 \leq i \leq 8} P(C_i) H(e|C_i)\).
If $P(e|C_i)$ is Laplacian, then we have an efficient way of determining near-optimal coding contexts based on $\sigma(e|\rho, \delta, \beta)$ and $\sigma(e|\delta, \beta)$. The method is due to Weinberger, Seroussi and Sapiro [16] who used it to encode symbols from a Laplacian symbol with a Golomb-Rice coder [3]. The Golomb-Rice coder has an integer parameter $m$ that can be viewed as a variance-based coding context; the larger the value of $m$, the larger the variance of coded symbols in the coding context. The Golomb-Rice coder encodes a positive integer value $z$ in two parts: 1) a binary representation of $z \mod m$, and 2) a unary representation of $\lfloor z/m \rfloor$. A signed integer error value $e$ can be mapped to a positive integer $z(e)$ by

$$z(e) = \begin{cases} 
2e & e \geq 0 \\
2|e| - 1 & e < 0
\end{cases}.$$  

(18)

For $m = 2^k$ the Golomb-Rice coder becomes very simple with an implementation involving only bit manipulations. By requiring $m$ to be some integer power $k$ of 2, we effect an exponential context quantization of $m$ with $k$ being the quantizer output. The key question in Golomb-Rice coder is how to determine the optimal quantized context $k$ that yields the minimum code length. Weinberger et. al showed that optimal $k$ can be satisfactorily approximated by

$$k = \lceil \log_2 \sigma(e|\cdot, \cdot) \rceil.$$  

(19)

For 8-bit gray scale images, we have $0 \leq k \leq 7$. In practice the simple scheme of (19) worked very well in choosing one of eight coding contexts even when an arithmetic coder rather than the Golomb-Rice coder is used. It obtained bit rates less than three percent more than those obtained with an optimal $Q(\sigma)$ that minimizes $\sum_{1 \leq i \leq 8} P(C_i)|H(e|C_i)$, unless the input image is highly discontinuous so that $P(e|C_i)$ significantly deviates from Laplacian.

6 Binary Mode

CALIC operates in two modes: binary and continuous-tone modes. This allows the CALIC system to distinguish between binary and continuous-tone types of images on a local, rather than a global, basis. This distinction between the two modes is important as the compression
methodology employed in each is very different. The latter uses predictive coding, whereas the former codes pixel values directly. CALIC selects one of the two modes depending on whether or not a local neighborhood has more than two distinct pixel values, using no side information whatsoever. The two-mode design contributes to the universality and robustness of CALIC over a wide range of images, particularly for multimedia images that mix text, graphics, line art, and photographs. In the binary mode, a context-based adaptive ternary arithmetic coder is used to code three symbols, including an escape symbol.

First let us see how CALIC’s binary mode works. Before \( y = Y[i, j] \) is to be coded, the algorithm first checks the six neighboring pixels: \( y_1, y_2, \ldots, y_6 \) (see Figure 2). If these six pixels have no more than two different values, then binary mode is triggered; otherwise the system enters continuous-tone mode. In binary mode, let \( s_1 = w \) and let the other value, if any, be \( s_2 \). \( y \) is coded as one of the three symbols:

\[
T = \begin{cases} 
0 & \text{if } y = s_1 \\
1 & \text{if } y = s_2 \\
2 & \text{otherwise}
\end{cases}
\] (20)

This mapping between \( y \) and \( T \) is called symbolization. In the escape case of \( T = 2 \), the encoder switches from the binary mode to the continuous-tone mode. This happens when the actual \( y \) is found to violate the condition for binary mode.

Similar to continuous-tone mode, upon entering binary mode, CALIC considers a context of six events

\[
C = \{y_1, y_2, \ldots, y_6\}
\] (21)

and quantizes the context \( C = \{y_1, y_2, \ldots, y_6\} \) to a 6-bit binary number \( B = b_6b_5\cdots b_1 \):

\[
b_k = \begin{cases} 
0 & \text{if } y_k = s_1 \\
1 & \text{if } y_k = s_2
\end{cases}, \quad 1 \leq k \leq 6.
\] (22)

The binary number \( B \) represents a texture pattern around \( y \). Note that because \( s_1 = w \), \( b_1 \equiv 0 \) can be ignored from the texture pattern \( B \). Thus there are only \( 2^5 = 32 \) different prediction contexts in the binary mode. Consequently, CALIC uses 32 conditional probabilities \( P(T|B) \) to drive an adaptive ternary arithmetic coder.
In interband CALIC, we use the binary pattern $B'$ of the reference band $X$ by the same
definition of Eq(22) as well as $B$ of band $Y$ for interband context modeling. Let $T'$ and $T$
be the symbolizations of $x$ and $y$ respectively in intra-band CALIC binary mode by Eq(20).
If $B' \neq B$, we simply code $y$ as in intra-band CALIC binary mode. But if $B' = B$, we use a
binary arithmetic coder to tell whether $T' = T$. Clearly, if $T' = T$ and $B' = B$ then $y$ can
be uniquely determined. If $T' \neq T$ but $B' = B$ we simply switch to continuous-tone mode
to code $y$. Because of the extremely high correlation between the two events $B' = B$ and
$T' = T$, we get very high compression in the case of $B' = B$. Intra-band CALIC obtains
significantly lower lossless bit rates for the compound images in the ISO test set (refer to
Table 1), largely because of the above interband binary coding scheme.

7 Experimental Results

In Table 1 we list bit rates of some multispectral test images that were obtained by LOCO-
I [16], intra-band CALIC, and the proposed interband CALIC. Note that LOCO-I is the
technique on which JPEG-LS, the new standard for lossless image compression is based,
and its performance is slightly superior to JPEG-LS due to simplifications made during the
standardization process.

The image “band” is a satellite image with four bands, “aerial” is an aerial photo with
three bands, “cats” and “water” are scanned RGB color images, “cmpnd1”, “cmpnd2”, and
“chart” are so-called compound RGB color images that have graphics, texts, and natural
photographs mixed in a single page. “ridgely” is a LANDSAT image with seven bands
commonly used as a test image for multispectral compression. Other than image “chart”,
interband CALIC outperforms intra-band CALIC by an appreciable margin. The improve-
ment is 27 per cent on “cats”, and 25 per cent on “cmpnd2”. The image “chart” presents the
worst case in terms of compression performance of interband CALIC relative to intra-band
CALIC. This is because “chart” is a synthesized image with a majority of its colors being
highly saturated. The high color saturation makes the interband correlation very weak. That
explains the very small compression gain made by interband CALIC over intra-band CALIC on this image.

To the best of our knowledge, these results are better than those proposed in the literature, based on comparison of the common images in the test set. For example in [12], bit rates of 2.04 and 1.59 were reported on “cats” and “water” using an inter-band compression technique. Furthermore, the performance of this technique on compound images is significantly poorer. This is due to the lack of a binary mode as in CALIC. In our previous work involving an inter-band extension of JPEG-LS [6] we had achieved bit rates of 3.93, 1.93, 1.55, 1.16, 1.06 and 2.83 on “aerial”, “cats”, “water”, “cmpnd1”, “cmpnd2” and “chart” respectively. Similarly, in [1] a bit rate of 2.83 was given for the “ridgely” image with a two-pass technique and reported to be better than that obtained in [13] and [14].

8 Acknowledgements

The authors would like to thank W. Choi for providing assistance in implementing the proposed technique. N. Memon was partially supported by NSF CAREER award NCR 964044.

References


Figure 1: Schematic description of the proposed image coding system.
Figure 2: The labeling of neighboring pixels used in prediction and modeling.
<table>
<thead>
<tr>
<th>Image</th>
<th>LOCO-I</th>
<th>Intra-band CALIC</th>
<th>Interband CALIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>band</td>
<td>3.36</td>
<td>3.20</td>
<td>2.72</td>
</tr>
<tr>
<td>aerial</td>
<td>4.01</td>
<td>3.78</td>
<td>3.47</td>
</tr>
<tr>
<td>cats</td>
<td>2.59</td>
<td>2.49</td>
<td>1.81</td>
</tr>
<tr>
<td>water</td>
<td>1.79</td>
<td>1.74</td>
<td>1.51</td>
</tr>
<tr>
<td>cmpnd1</td>
<td>1.30</td>
<td>1.21</td>
<td>1.02</td>
</tr>
<tr>
<td>cmpnd2</td>
<td>1.35</td>
<td>1.22</td>
<td>0.92</td>
</tr>
<tr>
<td>chart</td>
<td>2.74</td>
<td>2.62</td>
<td>2.58</td>
</tr>
<tr>
<td>ridgely</td>
<td>3.03</td>
<td>2.91</td>
<td>2.72</td>
</tr>
</tbody>
</table>

Table 1: Lossless bit rates of different algorithms.
Figure and Table Captions

Figure 1: Schematic description of the proposed image coding system.

Figure 2: The labeling of neighboring pixels used in prediction and modeling.

Table 1: Lossless bit rates of different algorithms.
Biosketches

Nasir Memon

Nasir Memon received his B.E. in Chemical Engineering and M.Sc. in Mathematics from the Birla Institute of Technology, Pilani, India in 1981 and 1982 respectively. He received his M.S. and Ph.D. degrees from the University of Nebraska, both in Computer Science, in 1989 and 1992 respectively. He has held an assistant professor position at Arkansas State University and at Northern Illinois University. During 1997-98, he was a visiting faculty at the Imaging Technology Department of Hewlett Packard Research Laboratories, Palo Alto, California. He is currently an Associate Professor in the computer science department at Polytechnic University, New York. His research interests include Data Compression, Multimedia Computing and Security.

Xiaolin Wu

Xiaolin Wu received his B.Sc. degree in computer science from Wuhan Cehui Technical University, China in 1982, and his Ph.D. degree in computer science from the University of Calgary, Canada in 1988. He is currently a professor at the Department of Computer Science, the University of Western Ontario, London, Ontario, Canada. His research interests include algorithms for visual computing and communications, data compression, image processing, and computer graphics. He published over seventy research papers in those areas. Dr. Wu is an active JPEG member and the 1998 UWO distinguished research professor.