SCALABLE PERCEPTUAL IMAGE CODING FOR REMOTE SENSING SYSTEMS

Graduate Students: Han Oh and Hariharan G. Lalgudi
Advisors: Michael W. Marcellin and Ali Bilgin
Department of Electrical and Computer Engineering, The University of Arizona

ABSTRACT

In this work, a scalable perceptual JPEG2000 encoder that exploits properties of the human visual system (HVS) is presented. The algorithm modifies the final three stages of a conventional JPEG2000 encoder. In the first stage, the quantization step size for each subband is chosen to be the inverse of the contrast sensitivity function (CSF). In bit-plane coding, two masking effects are considered during distortion calculation. In the final bitstream formation step, quality layers are formed corresponding to desired perceptual distortion thresholds. This modified encoder exhibits superior visual performance for remote sensing images compared to conventional JPEG2000 encoders. Additionally, it is completely JPEG2000 Part-1 compliant, and therefore can be decoded by any JPEG2000 decoder.

Keywords: JPEG2000, remote sensing images, perceptual image coding, scalable compression

1. INTRODUCTION

With advances in sensor and imaging technology, the size of remote sensing images has increased dramatically over the last few decades. Efficient and scalable image compression has become an essential part of many remote sensing systems [1]. Scalability allows end users to selectively obtain different image products (e.g., resolution, quality, component and region of interest) without the need for transmitting the entire compressed data. In this regard, JPEG2000 [2] offers these functionalities and is well fitted to remote sensing systems [3].

Typically, a JPEG2000 encoder adopts mean squared error (MSE) as the distortion metric and
compression is performed by removing statistical redundancies. The reduction of statistical redundancies does not always mean the reduction of psychovisual redundancies. In particular, visually important information, such as boundaries of buildings or small pathways in remote sensing images, is often lost at low bitrates when MSE is used as the distortion metric. Thus, incorporating properties of the human visual system (HVS) in a JPEG2000 encoder provides room for visual improvement at such low bitrates and leads to better compression performance.

One property of the HVS is the contrast sensitivity function (CSF) which characterizes variation of contrast sensitivity with spatial frequency. While the CSF affects the entire image, visual masking influences more localized areas. Using visual masking effects, more precise control of distortion is possible. In this paper, we introduce a scalable perceptual encoder for remote sensing images. Our encoder is based on a recent perceptual encoder [4] which exploits the CSF and locally calculated visual masking effects. Quality scalability is added in our encoder. Also, our encoder allows more precise control of the visual quality in the near visually lossless regime.

2. PERCEPTUAL DISTORTION MODEL

In perceptual image coding, distortion is quantified based on what we perceive, rather than on statistical values. The HVS is non-linear and quite complex. However, this system can be approximated with a few well-known components such as the CSF and visual masking effects. In JPEG2000, the discrete wavelet transform (DWT) decorrelates images and its output is several bands with different spatial frequencies. Human eyes have different sensitivity for each spatial frequency. Watson et al. [5] have measured the visibility of quantization error for 9/7 DWT coefficients on each band. At uniform background intensity, the minimum quantization step sizes at which quantization errors become visible were modeled by gradually increasing the quantization step sizes. This is called the base just-noticeable difference (JND) threshold and this set of quantization step sizes ensures visually lossless coding. The JND threshold is a function of visual resolution [5] \( r \). The JND values at a visual resolution of 33.0 pixels/degree are listed in Table 1.

<table>
<thead>
<tr>
<th>Orientation</th>
<th>DWT level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>LL</td>
<td>7.3165</td>
<td>5.7340</td>
<td>5.8168</td>
<td>7.3599</td>
<td>11.4241</td>
<td></td>
</tr>
</tbody>
</table>

Table 1. Base JND threshold (\( r = 33.0 \) pixels/degree)
Since the base JND threshold was modeled at uniform background, the actual JND threshold should take into account visual masking effects of the image background. In [4], two visual masking effects are considered. One is luminance masking based on Weber’s law. Luminance masking is sometimes called luminance adaptation and adjusts the base JND threshold depending on the background intensity. The adjustment \( l_{b}[j] \) for the \( j \)-th DWT coefficient of subband \( b \) is modeled by

\[
\begin{align*}
   l_{b}[j] &= \left( \frac{y_{LL}[j']}{\bar{y}_{LL}} \right)^\alpha \\
\end{align*}
\]

where \( y_{LL}[j'] \) denotes the value of the \( j' \)-th DWT coefficient of subband \( LL_s \) which corresponds to the \( j \)-th DWT coefficient of subband \( b \). The corresponding position \( j' \) is calculated as \( \left\lfloor j / 2^{b_{D}} \right\rfloor \) where \( D_{b} \) is the DWT level index of subband \( b \). \( \bar{y}_{LL} \) denotes the average value of DWT coefficients in \( LL_s \). For unsigned 8-bit images, \( \bar{y}_{LL} \) is 128. The parameter \( \alpha \) represents the degree of luminance masking effect and is set to 0.649 [6].

Another masking that affects the JND threshold is contrast masking. A high contrast background reduces the visibility of distortion. It is assumed that points with large magnitude of DWT coefficients are high contrast points. Using this property, the perceived distortion is adjusted based on the magnitude of DWT coefficients. The contrast effect is defined as

\[
\begin{align*}
   c_{b}[j] &= \max \left\{ 1, \left( \frac{|y_{b}[j]|}{JND_{b} \cdot l_{b}[j]} \right)^{\epsilon} \right\} \\
\end{align*}
\]

where \( JND_{b} \) is the base JND threshold and \( \epsilon \) experimentally takes a value of 0.25. For \( LL_s \), the contrast masking is ignored (\( c_{LL_s}[j] = 1 \) for \( \forall j \)).

The final JND threshold that considers the visual masking effects is given by

\[
JND_{b}^*[j] = JND_{b}[j] \cdot l_{b}[j] \cdot c_{b}[j] 
\]

This threshold states whether the quantization error of the \( j \)-th DWT coefficient in subband \( b \) is visible or not. Though the JND threshold is designed for visually lossless coding, the visibility of quantization errors at suprathreshold levels is assumed to be inversely proportional to the threshold [7]. The probability of detection of the quantization error is given by

\[
p_{b}[j] = 1 - \exp \left( - \frac{\hat{y}_{b}[j] - y_{b}[j]}{JND_{b}^*[j]} \right) 
\]

where \( \hat{y}_{b}[j] \) is the reconstructed DWT coefficient and \( \hat{y}_{b}[j] - y_{b}[j] \) is the quantization error.
The parameter $\beta$ is an experimental value of 4 [7].

It is not easy to control quantization errors for individual DWT coefficients. The foveal region $F$ which covers the visual angle of 2° in the spatial domain is defined as the basic size in which distortion is calculated, as shown in Figure 1. The summation of the numbers of DWT coefficients $\mathcal{N}(F_b)$ that constitute a foveal region is the same as the number of pixels for a foveal region $\mathcal{N}(F)$ and is defined as

$$\mathcal{N}(F) = \sum_b \mathcal{N}(F_b) = (\lfloor 2r \rfloor)^2$$

(5)

Distortion for the foveal region, centered at $(n_1, n_2)$, is calculated by considering the quantization errors of all DWT coefficients that belong to that region. Specifically,

$$D_{F(n_1, n_2)} = \left( \sum_b \sum_{j \in F} \left| e_b[j] \right|^\beta \right)^{1/\beta}$$

(6)

There can be several foveal regions depending on the size of the image. Distortion for the entire image takes the maximum value among distortion values of foveal regions.

$$D = \max_{(n_1, n_2)} D_{F(n_1, n_2)}$$

(7)

Figure 1. Foveal region $F$, and DWT coefficients belonging to the region $F_b$, $b = HH_1, LH_1, \ldots, LL_3$

3. SCALABLE PERCEPTUAL IMAGE CODING WITH JPEG2000

Figure 2 shows the JPEG2000 encoding block diagram. In order to incorporate the properties of the HVS into JPEG2000, the algorithm modifies the final three stages of a conventional JPEG2000 encoder. In the first stage, the quantization step size for each subband is chosen to be
the base JND threshold.

$$\Delta_b = k \cdot JND_b \cdot 2^{-b_b}, \quad 0 < k \leq 1$$  \hspace{1cm} (8)

where $2^{-b_b}$ accounts for the difference between the 9/7 DWT of JPEG2000 and that of [5]. The JND threshold is a statistical value obtained based on human observers. In our work, we allow the option to reduce the base JND threshold by half ($k = 0.5$), to ensure strict visually lossless coding.

In EBCOT [8] tier-1, the block coder does independent bit-plane coding of quantized indices in codeblock $B$. Each bit-plane has three coding passes except the most significant bit-plane (MSB), which has 1 coding pass. Thus, quantized indices with $M$ bit-planes have $3M - 2$ coding passes. For each coding pass $z$, a distortion value is calculated. In a conventional JPEG2000 encoder, the MSE is used as the distortion metric and is given by

$$D^{(z)} = G_b \sum_{j \in B} (\hat{y}^{(z)}[j] - y[j])^2$$  \hspace{1cm} (9)

where $G_b$ is the synthesis gain for subband $b$. $\hat{y}^{(z)}[j]$ denotes the reconstructed value from the quantization index $q^{(z)}[j]$, which is encoded up to truncation point $z$. $y[j]$ is the reconstructed value from the original quantization index $q[j]$.

EBCOT has modularity and many distortion metrics can be used. The scalable perceptual image coding scheme uses the following distortion metric that takes into account the visual masking effects over the foveal region $\mathcal{F}_b$.

$$D_{\mathcal{F}_b}^{(z)} = \left( \sum_{j \in \mathcal{F}_b \cap \mathcal{J}_b} \left( \frac{q^{(z)}[j] - q[j]}{l_b[j] \cdot c_b[j]} \right)^\beta \right)^{\frac{1}{\beta}}$$  \hspace{1cm} (10)

Here, the luminance masking $l_b[j]$ is the same as Equation (1), but the contrast masking $c_b[j]$ is modified as

$$c_b[j] = \max \left\{ 1, \left( \frac{k \cdot y_b[j]}{\Delta_b \cdot l_b[j]} \right)^\epsilon \right\}$$  \hspace{1cm} (11)

Figure 3 shows the visual masking effects. The luminance masking values are simply replicated.
values of the LL band and the contrast masking values are proportional to the magnitude of the DWT coefficients. More distortion can be placed on the bright areas which are assumed to be high-contrast areas.

While Equation (9) is calculated over a code-block, Equation (10) is calculated over a foveal region $\mathcal{F}_b$. The reason for using $\mathcal{F}_b$ instead of $\mathcal{F}$ as in Equation (6) is due to JPEG2000’s independent encoding characteristic of code-blocks. The size of $\mathcal{F}_b$ is usually smaller than that of a code-block $B$. Since there can be several foveal regions in a code-block, the maximum distortion value is chosen as the distortion value over code-block $B$.

$$D(z) = \max_{\mathcal{F}_b \subset B} D(z)$$

(12)

Finally, once distortions and bit lengths associated with all coding passes are calculated, they are transferred to EBCOT tier-2.

In this work, quality layers are formed for a given set of target distortions. Distortion needs to be a monotonic decreasing function of $z$. Hence, a set of feasible truncation points $\mathcal{H}$ is defined as follows

$$\mathcal{H} \triangleq \left\{ z \mid D(z) \leq D(z-1) \text{ for } z = 0, 1, \ldots, Z \right\}, \quad D^{(-1)} = \infty$$

(13)

For $z \in \mathcal{H}$, a bitstream for the $i$-th quality layer includes all coding passes up to $z^*$ such that

$$D^{(z)} \leq \left( N(\mathcal{F}_b), D^\theta \right) \left( \frac{1}{\beta} \right)$$

(14)

where $D^\theta_i$ indicates the target distortion for $i$-the quality layer. Bitstreams produced by this modified encoder are completely JPEG2000 Part-1 compliant and therefore decodable by any JPEG2000 decoder.

![Figure 3](image-url) Luminance masking (left) and contrast masking (right)
4. EXPERIMENTAL RESULTS

The scalable perceptual image coding algorithm was implemented in Kakadu v6.0 [9]. We present results for a visual resolution of 33.0 pixels/degree (viewing distance of 60 cm). The step sizes were set to half of the base JND ($k = 0.5$). For simplicity, the size of foveal region is $64 \times 64$ ($N(F) = 4096$). Test images used here are **aerial1** and **aerial2**, remote sensing images in the ISO JPEG2000 test image set. Because of the huge size of the original **aerial1** image (14680×14565), it has been cropped and down-scaled by a factor of 4. The sizes of test images are 2048×2048. In both cases, 10 quality layers are generated.

Figure 4 shows the bitrate and perceptual distortion for each quality layer. From the figure, we can see that the achieved distortions are upper-bounded by the target distortions and become closer to the target distortions as the target distortions decrease. Reconstructed images corresponding to different layers of the **aerial1** image are shown in Figure 5. All images in this figure were reconstructed from a single compressed file. At a perceptual distortion of 1.0, the reconstructed images are visually lossless.

Figure 6(a) and 6(b) show reconstructed **aerial1** images using the perceptual JPEG2000 encoder and a conventional JPEG2000 encoder respectively. Though a conventional JPEG2000 encoder yields the optimal bitstreams in terms of MSE, we can see that images encoded with the perceptual image coder exhibit less blurring and superior visual quality than those encoded with a conventional JPEG2000 encoder. In particular, at low bitrates, the perceptual image coder better preserves details compared to a conventional JPEG2000 encoder. Figure 7 compares the bitrate of the two methods to achieve the same visual quality on the **aerial2** image. A 27% reduction in the bitrate is achieved using the perceptual JPEG2000 encoder when compared to a conventional JPEG2000 encoder.

5. CONCLUSIONS

We have presented a JPEG2000 Part-I compliant scalable perceptual image encoder. Remote sensing images often contain significant high resolution detail. At low bitrates, this information is easily lost via blurring when encoded with a conventional JPEG2000 encoder. In the perceptual encoder, distortions are locally hidden in high contrast areas in which they are less visible. Thus, the encoder yields superior visual performance and visually important details are better preserved.
REFERENCES


Figure 4. Perceptual distortion and bitrate, *aerial1* (left) and *aerial2* (right)

Figure 5. Reconstructed images encoded with the perceptual JPEG2000 encoder

(a) First layer \((R = 0.343, \; D = 3.317)\)
(b) Fourth layer \((R = 0.426, \; D = 2.227)\)
(c) Sixth layer \((R = 0.842, \; D = 1.334)\)
(d) Tenth layer \((R = 1.262, \; D = 0.894)\)
Figure 6. Visual quality comparison at the same bitrate $R = 0.493$ (fourth quality layer): boundaries of the tiles on the roof are more obvious in the left image.

Figure 7. Visually equivalent images: (a) $R = 0.857$ bpp (b) $R = 1.176$ bpp