LOCALLY-ADAPTIVE IMAGE CODING BASED ON A PERCEPTUAL TARGET DISTORTION

Ingo Höntsch
Telecommunications Research Center
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-5706
hontsch@asu.edu

Lina J. Karam
Telecommunications Research Center
Department of Electrical Engineering
Arizona State University
Tempe, AZ 85287-5706
karam@asu.edu

ABSTRACT

This paper presents a perceptual-based image coder, which discriminates between image components based on their perceptual relevance for achieving increased performance in terms of quality and bit-rate. The new coder uses a locally-adaptive perceptual quantization scheme based on a tractable perceptual distortion metric. Our strategy is to exploit human visual masking properties by deriving visual masking thresholds in a locally-adaptive fashion. The derived masking thresholds are used in controlling the quantization stage by adapting the quantizer reconstruction levels in order to meet a desired target perceptual distortion. The proposed coding scheme is flexible in that it works with any subband-based decomposition and with block-based transform methods. Compared to the existing perceptual transform-based and block-based methods, the proposed perceptual coding method exhibits superior performance in terms of bit rate and distortion control. Coding results are presented to illustrate the performance of the presented coding scheme.

1. INTRODUCTION

Driven by a growing demand for transmission and storage of visual data over media with limited capacity, increasing efforts have been made to improve compression techniques for visual information. One promising path is to integrate models of the *human visual system* (HVS) into the design of coding algorithms. This has been motivated by the fact that, given the diversity of image types and sources, reliable engineering models for image sources currently do not exist. With the absence of reliable image source models, image coding algorithms must rely upon generalized receiver models to optimize their efficiency and performance. For an image the ultimate receiver is the human visual system, and image perception is affected by its sensitivity and masking properties.

Perceptual-based algorithms attempt to discriminate between signal components which are and are not detected by the human receiver. They exploit the visual masking properties of the human visual system and establish thresholds of *just-noticeable distortion* (JND) and *minimally-noticeable distortion* (MND) based on psychophysical masking phenomena. Since images are usually stored and transmitted in a compressed form due to their large information content, the interest has been especially in developing reliable and efficient image coding algorithms. The central ideas in perceptual coding are: 1) to "hide" coding distortion beneath spatial and temporal JND thresholds, and 2) to augment the classical coding paradigm of redundancy removal with elimination of irrelevant signal information, i.e., discarding those signal components which are imperceptible to the human receiver.

This paper presents a locally-adaptive perceptual-based image coder with the objective to minimize the bit rate for a desired perceptual target distortion. The proposed coder is flexible in that it works with any subband-based decomposition as well as with block-based transform methods. Our strategy is to exploit human visual masking properties which we derive in a locally-adaptive fashion for the desired subband decomposition or block-based transform. The specified subband decomposition or frequency transform decomposes the visual scene into elemental components (channels) with varying frequency and and orientations. We then *adaptively* compute local distortion sensitivity profiles in the form of detection thresholds that adapt to the varying local frequency, orientation, and spatial characteristics of the considered image data. The derived thresholds are used to adaptively control the quantization and dequantization stages of the coding system in order to meet a target perceptual distortion.

The paper is organized as follows. Section 2 discusses previous related work. Section 3 describes the proposed coding algorithm. Coding results and comparison with existing perceptual-based coding schemes are presented in Section 4.

2. EXISTING PERCEPTUAL CODING SCHEMES

True perceptual quantization requires computing and making use of image-dependent, locally-varying, masking thresholds. However, the main problem in using a locally-adaptive perceptual quantization strategy is that these locally-varying masking thresholds are needed both for encoding and decoding. This, in turn, would require sending or storing a large amount of side information and would result in a significant increase in bit rate.

The existing and recently developed "perceptual-based" compression methods attempt to avoid this problem by giving up or significantly restricting the local adaptation. One method called DCTune [1] fits within the framework of JPEG. Based on a model of human perception that considers frequency sensitivity and contrast masking, it designs a *fixed* DCT quantization matrix (3 quantization matrices in the case of color images) for each image. The fixed quantization matrix is selected to minimize the overall perceptual distortion. The Perceptual Image Coder (PIC) proposed by Safranek and Johnston [2] works in a subband decomposition setting. Each subband is quantized using a uniform quantizer with a *fixed* step size. The step size is determined by the JND threshold for uniform noise at the most sensitive coefficient in the subband. A scalar multiplier in the range of 2 to 2.5 is then applied to uniformly scale all step sizes in order to compensate for the conservative step

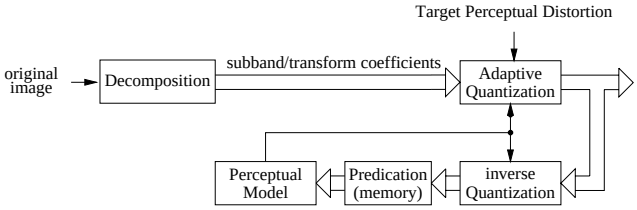

Figure 1. Block diagram of the proposed coding scheme

size selection and to achieve good compression ratio. In block-based methods [3], a scalar value can be used for each block or macro block to uniformly scale a *fixed* quantization matrix in order to account for the variation in available masking (and as a means to control the bit rate). The quantization matrix and the scalar value for each block need to be transmitted, resulting in additional side information.

All of these methods choose a fixed quantization matrix for the whole image, select one fixed step size for a whole subband, or scale all values in a fixed quantization matrix uniformly. They do not take into account the locally-varying masking thresholds which differ based on the local image content. Because of this limitation, they fail to exploit the large dynamic range of the available masking and tend to allocate too much bits to less sensitive coefficients, resulting in over-coding of some image components or in unnecessary visible artifacts. Furthermore, it would be desirable to have control over the resulting perceptual distortion.

3. PROPOSED PERCEPTUAL-BASED CODING SCHEME

The main blocks of the proposed coding scheme are shown in Fig. 1. The goal is to adapt the quantizer parameters to achieve reduction in bit-rate for a given target perceptual distortion.

3.1. Multi-channel Decomposition

The decomposition stage decomposes the input image into components with varying frequency and orientation (channels). These channels differ in terms of their sensitivity and masking properties. Each channel b consists of a matrix of coefficients $i(b,r,c)$, where r and c represent the row and column numbers, respectively, indicating the location of the coefficient within the considered channel. The local amount of masking differ for each coefficient in function of its channel location (frequency and orientation), local luminance (given by lowest frequency channel coefficients), local contrast in each channel (channel contrast masking). For subband-based decompositions, the channels correspond to the resulting subband images. For block-based decompositions, the bth channel would consist of the collection of all bth transform coefficients taken one at a time from subsequent blocks.

3.2. Perceptual Model

A tractable perceptual distortion metric is needed in order to adapt the quantizer parameters to the local masking characteristics of the visual data and meet the desired level of perceptual quality.

Our perceptual distortion metric is based on the "probability summation" model presented in [4]. Our detection model can be described as consisting of a set of detectors, one at each location (b,r,c); $P_{(b,r,c)}$ is then the probability that detector (b,r,c) will signal the occurrence of a distortion or, equivalently, it is the probability that the perceptual error introduced at (b,r,c) is visible (above threshold).

We define the probability of detection over R, P_R, to be the probability that a distortion will be detected over the region R. Then, based on the above, P_R will be

$$P_R = 1 - \prod_{(b,r,c)\in R} (1 - P_{(b,r,c)}). \tag{1}$$

Since human observers can mainly scrutinize the image region that falls on the fovea (two degrees of visual angle), we propose pooling the error over subband image regions, which we will denote by foveal regions $R = \mathcal{F}_{(b,r,c)}$. $\mathcal{F}_{(b,r,c)}$ is the region centered at location (r,c) in visual channel b and whose size is computed to cover two degrees of visual angle. In (1), $P_{(b,r,c)}$ is the probability of detecting a distortion at coefficient $i(b,r,c)$. It is determined by the psychometric function which can be modeled as an exponential of the form

$$P_{(b,r,c)} = 1 - \exp\left(-\left|\frac{e(b,r,c)}{t(b,r,c)}\right|^{\beta_b}\right) \tag{2}$$

where $e(b,r,c)$ is the quantization error, $t(b,r,c)$ denotes the detection (masking) threshold at location (b,r,c), and β_b is a parameter whose value is usually chosen to maximize the correspondence of (2) with an experimentally computed psychometric function for a given type of distortion. In psychophysical experiments that examine summation over space, a β_b of about 4 has been observed [4]. Note that, in this case, the threshold is defined to be the error value $e(b,r,c)$ which results in a detection probability $P_{(b,r,c)} = 0.63$.

Substituting (2) in (1), and setting $R = \mathcal{F}_{(b,r,c)}$, results in

$$P_{\mathcal{F}_{(b,r,c)}} = 1 - \exp\left(-(D_{(b,r,c)})^{\beta_b}\right) \tag{3}$$

where

$$D_{(b,r,c)} = \left(\sum_{\mathcal{F}_{(b,r,c)}} \left|\frac{e(b,r,c)}{t(b,r,c)}\right|^{\beta_b}\right)^{1/\beta_b} \tag{4}$$

which takes the form of a Minkowski metric with exponent β_b. Hence, minimizing the probability of detecting a difference in the foveal region $\mathcal{F}_{(b,r,c)}$ is equivalent to minimizing the metric $D_{(b,r,c)}$. Note that the threshold probability $P_{\mathcal{F}_{(b,r,c)}} = 0.63$ occurs when $D_{(b,r,c)} = 1$. So, lossless perceptual coding would correspond to $D_{(b,r,c)} \leq 1$.

Finally, our perceptual distortion measure D is based on the maximum probability of detection for all foveal regions,

$$D = \max_{(b,r,c)} \left\{D_{(b,r,c)}\right\} \tag{5}$$

The detection thresholds $t(b,r,c)$ used for determining the perceptual distortion are computed as

$$t(b,r,c) = t_{CS}(b,r,c) \cdot a_{CM}(b,r,c) \tag{6}$$

where $t_{CS}(b,r,c)$ is the contrast sensitivity (base detection) threshold and $a_{CM}(b,r,c)$ is the contrast masking adjustment.

```
for channel b
    for (r, c)
        compute t_(b,r,c)
    initialize q_b
    do {
        for (r, c)
            compute m̂_(b,r,c)
            compute t̂_(b,r,c)
            s(b, r, c) = 2 t̂(b, r, c) q_b/α
            e(b, r, c) = i(b, r, c) − s(b, r, c) ⌊i(b,r,c)/s(b,r,c) + 1/2⌋
        compute D_b = max_(r,c) {D(b, r, c)}
        adjust q_b using bisection if D_b > D_T
    } while (q_b has changed)
quantize image using {q_b} as step-size weights
(step-sizes s(b, r, c) generated at decoder and
do not need to be transmitted)
```

Figure 2. Pseudo-code for distortion-constrained locally-adaptive perceptual quantization

$t_{CS}(b,r,c)$ is a measure, for each channel b, of the smallest contrast that yields a visible signal over a background of uniform intensity. With the background luminance set to 127 (neutral gray), the detection threshold $t_{127}(b)$ is established for channel b by psycho-visual detection tests for uniform noise injected in the considered channel b. The test is repeated for all channels to obtain the complete set of *base sensitivity* thresholds $t_{127}(b)$. The base sensitivity thresholds are essentially a measure of the HVS Contrast Sensitivity Function (CSF) for the specified decomposition and are a global characteristic independent of the input image. The obtained base sensitivity thresholds $t_{127}(b)$ measure the contrast sensitivity in function of frequency while fixing the background intensity level. In general, the detection threshold varies also with the background intensity. This phenomenon is known as luminance masking or light adaptation [1]. In order to account for luminance masking, detection thresholds $t_m(b)$ are measured with different uniform background intensities m resulting in a "brightness correction" adjustment $a_D(m) = t_m(b)/t_{127}(b)$. In our case, the uniform background corresponds to the local mean $m(r,c) = mean + i(0,r,c)$, where $mean$ is the global mean of the input image (which is removed before the decomposition and coded separately) and $i(0,r,c)$ is the transform coefficient in subband 0 at the considered location (r,c). The "brightness correction" adjustment $a_D\,(m(r,c))$ accounts for the variations in sensitivity depending on the local mean $m(r,c)$. It follows that the base detection threshold $t_D(b,r,c)$ is given by

$$t_D(b,r,c) \;=\; t_{127}(b) \;\cdot\; a_D(m(r,c))\,. \qquad (7)$$

$a_{CM}(b,r,c)$ refers to the reduction in the visibility of one image component (the target) by the presence of another one (the masker). In our case, the target is the quantization noise and the masker is given by the input image channel components. Our contrast masking model is derived from a non-linear transducer model for masking of sinusoidal gratings [5]. The non-linear transducer model was adapted for the considered band-limited channel components. $a_{CM}(b,r,c)$ is thus given by:

$$a_{CM}(b,r,c) \;=\; \begin{cases} \max\left\{1, \left|\dfrac{m_{(b,r,c)}}{t_{CS}(b,r,c)}\right|^{0.6}\right\} & b \neq 0 \\ 1 & b = 0 \end{cases} \qquad (8)$$

In (8), $m_{(b,r,c)}$ is the weighted average magnitude over the foveal region $\mathcal{F}_{(b,r,c)}$.

3.3. Distortion-Constrained Perceptual Quantization

The computed local masking thresholds are used to adaptively control the step-size $s(b,r,c)$ of a uniform quantizer while meeting a desired perceptual distortion D_T. The proposed adaptive quantization scheme is shown in the form of pseudo-code in Fig. 2.

$\{q_b\}$ are channel weights that need to be determined such that a target perceptual distortion $D = D_T$ is met. In order to stay compliant with the bitstream syntax of the DCT-based standards, when a DCT-based decomposition is used, q_b is taken to be an 8-bit integer and can, thus, be transmitted as the entries of the "quantization matrix" Q. However, in our case, the entries q_b of Q are not directly used as step sizes. Instead, they are interpreted as weights for the local adaptive step-sizes $s(b,r,c)$, which are used to quantize the channel coefficients $i(b,r,c)$. A factor α is used to limit the step-size multiplier q_b/α to a desired maximum value. In our case, we use $\alpha = 32$, which gives $q_b/\alpha < 8$. After estimating the masking thresholds $t(b,r,c)$, the goal is to optimize the weights $\{q_b\}$ in such a way, that the quantization is as coarse as possible while the target perceptual distortion $D = D_T$ is met. q_b can be optimized separately for each channel b since $D = D_T$ is met if $D_b = \max_{(r,c)}\{D(b,r,c)\} \leq D_T$ for all b.

As shown in Fig. 2, the masking thresholds $t(b,r,c)$ are computed and q_b is initialized based on the computed $t(b,r,c)$ and the desired target distortion D_T. Then, the optimal value of q_b is determined by means of an iterative process. In every iteration, D_b is computed. If $D_b \leq D_T$ the value of q_b is increased using a bisection method, otherwise is is decreased using the same method. The process is terminated when q_b has not changed with respect to the previous iteration. Since q_b is an 8 bit integer, the bisection process terminates after at most 9 iterations.

One main issue in developing the proposed perceptual-based coding approach is that the image-dependent, locally-varying, masking thresholds $t(b,r,c)$ are needed both at the encoder and at the decoder in order to be able to reconstruct the coded image. This, in turn, would require sending a large amount of side information, and the associated increase in bit-rate conflicts with the original objective making very low bit-rate coding virtually impossible. Our proposed locally-adaptive coding scheme eliminates the need for transmitting side information for each step-size by estimating the available masking $\hat{t}(b,r,c)$ both at the encoder and decoder from the already received data and a prediction of the transform coefficient to be quantized. In this way, the local masking characteristics of the visual data can be exploited without having to transmit additional side information. The predictor used to estimate $t(b,r,c)$ is a linear, four point, first-order predictor. It uses the causal closest four neighbors. A set of predictor coefficients is computed for each channel based on the correlations of the transform coefficients in this particular channel. $\hat{t}$ and $\hat{m}$ in Fig. 2 denote the estimated values.

4. RESULTS

We illustrate the performance of the proposed coding scheme using 512×512 grayscale images and two different decompositions: 1) a DCT-based decomposition, in which case we compare with Watson's DCTune method [1]; 2) a *generalized quadrature mirror filter-bank* (GQMF) decomposition, in which case we compare

(a) Original Indian image, 8 bpp (b) PIC [2]: 0.640 bpp (b) New peceptual codec: 0.375 bpp

Figure 3. GQMF-based coding results with $D_T = 3.24$

image	first-order entropy		
	DCTune [1]	Proposed Coder	
baboon	1.639	1.364	(-17%)
indian	1.966	1.271	(-35%)
leena	0.956	0.757	(-21%)
lighthouse	1.248	0.825	(-34%)

Table 1. First-order entropies for DCTune [1] and the proposed coding scheme with a DCT-based decomposition

image	D_T	first-order entropy		
		PIC [2]	Proposed Coder	
baboon	3.76	0.902	0.485	(-46%)
indian	3.24	0.640	0.375	(-41%)
leena	3.24	0.488	0.341	(-30%)
lighthouse	3.28	0.588	0.403	(-31%)

Table 2. First-order entropies for PIC [2] and the proposed coding scheme with a GQMF-based decomposition

with the Safranek-Johnston PIC [2]. The perceptual thresholds were optimized for a viewing distance of 6 times the image height.

Tables 1 and 2 compares the proposed coding scheme with Watson's DCTune (DCT-based) and the Safranek-Johnston PIC (GQMF-based), respectively, in terms of first-order entropies. The target distortion has been set to $D_T = 1.0$ (perceptually-lossless) for the DCT-based comparison. For the GQMF-based decomposition, the Safranek-Johnston PIC does not allow a target distortion to be specified. So, in this case, the comparisons were achieved by setting the PIC step-size scalar multiplier to 2.5 (which results in almost transparent, high quality, images [2]) and explicitly computing the resulting coding distortion produced by the PIC algorithm for each coded image using the metric (5). Then, in our coding scheme, D_T is set to be equal to the distortions produced by the PIC. These distortions are listed in Table 2. Fig. 3 presents the obtained coding result for the Indian image using the GQMF-based coders and $D_T = 3.24$ (corresponding to a step-size multiplier of 2.5 in the Safranek-Johnston PIC).

REFERENCES

[1] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," *Society for Information Display Digest of Technical Papers XXIV*, pp. 946–949, 1993.

[2] R. J. Safranek and J. D. Johnston, "A perceptually tuned subband image coder with image dependent quantization and post-quantization," in *IEEE ICASSP*, 1989, pp. 1945–1948.

[3] Bo Tao, "On adaptive quantization and rate control for MPEG video coding environments," Tech. Rep., David Sarnoff, Aug. 1996.

[4] J.G. Robson and N. Graham, "Probability summation and regional variation in contrast sensitivity across the visual field," *Vision Research*, vol. 21, pp. 409–418, 1981.

[5] J. M Foley and G. M. Boynton, "A new model of humane luminance pattern vision mechanisms: Analysis of the effects of pattern orientation, spatial phase and temporal frequency," in *SPIE Proceedings*, T. A. Lawton, Ed., 1994, vol. 2054, pp. 32–42, Computational Vision Based on Neurobiology.