

Adaptive Image Coding With Perceptual Distortion Control

Ingo Höntsch, *Member, IEEE*, and Lina J. Karam, *Member, IEEE*

Abstract—This paper presents a discrete cosine transform (DCT)-based locally adaptive perceptual 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 the desired target perceptual distortion. The proposed coding scheme is flexible in that it can be easily extended to work with any subband-based decomposition in addition to block-based transform methods. Compared to existing perceptual coding 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.

Index Terms—Image compression, locally adaptive quantization, perceptual distortion.

I. INTRODUCTION

THE ability of humans to analyze, capture, and recall visual information significantly outperforms their ability to analyze and recall other types of sensory information. This fact makes humans rely heavily on their visual sense for extracting information and learning about their surroundings, and for planning and executing purposeful acts. As a result, the need to reliably process and transmit visual data has become central to many applications. In particular, with the rapid development and migration of end-user applications and services toward transmission media which place high constraints on bandwidth, such as the Internet and wireless media, there is a growing need for the development of new efficient image and video compression techniques, which offer reduction in bit rates and improvement in quality at low bit rates.

Although the importance of exploiting human perception has long been recognized within the signal and image processing community, the previous research efforts in image and video compression have concentrated on developing methods to minimize not perceptual but rather mathematically tractable, easy

to measure, distortion criteria, such as the mean squared error (MSE). While non perceptual distortion measures were found to be reasonably reliable for higher bit rates (high-quality applications), they do not correlate well with the perceived quality at lower bit rates and they fail to guarantee preservation of important perceptual qualities in the reconstructed images despite the potential for a good signal-to-noise ratio (SNR).

Perceptual-based coding algorithms attempt to discriminate between signal components which are and are not detected by the human receiver. They attempt to remove redundant as well as the perceptually less significant information. This is typically achieved by exploiting the masking properties of the human visual system and establishing detection thresholds of *just-noticeable distortion* (JND) and *minimally noticeable distortion* (MND) based on psychophysical masking phenomena. The main ideas in perceptual coding are 1) to “hide” coding distortion beneath the detection thresholds and 2) to augment the classical coding paradigm of redundancy removal with elimination of perceptually irrelevant signal information.

Mannos and Sakrison’s work [1] was one of the first image processing works that used vision science concepts. It demonstrated that simple computational distortion measures, such as the MSE, cannot reliably predict the difference of the perceived quality of one image with another. Based on psychovisual experiments, Mannos and Sakrison inferred some properties of the human visual system and developed a closed form expression of the contrast sensitivity function (CSF). This CSF model was incorporated in a distortion measure, which provided a better prediction of perceived quality. After Mannos and Sakrison’s contribution, several other schemes using models of human vision appeared in the engineering community. Nill [2] extended the model by Mannos and Sakrison and proposed a weighted cosine transform for the coding of digital images. By adapting the contrast sensitivity function for use with the discrete cosine transform (DCT), he developed a tractable model that could be incorporated in DCT-based coders. Eggerton and Srinath [3] proposed a perceptually weighted quantization for DCT-based schemes with entropy coding. Saghri *et al.* [4] further refined Nill’s model to account for display device calibration, viewing distance and image resolution. Macq [5] derived, for transform coding schemes, perceptual weighting factors related to the quantization noise introduced on transform coefficients; the derived weighting factors vary as a function of the display and the viewing conditions but are independent of the image content. Additional examples for contributions that incorporated visual factors without explicitly optimizing the compression for individual images include the work of Daly [6], and perceptual weighting applied to trellis coded quantization (TCQ) [7], [8] and embedded zerotree coding [9].

Manuscript received March 15, 2000; revised December 11, 2001. This work was supported by the National Science Foundation under Grant CCR-9733897. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Touradj Ebrahimi.

I. Höntsch is with the Institut für Rundfunktechnik, 80939 Munich, Germany (e-mail: hontsch@irt.de).

L. J. Karam is with the Telecommunications Research Center, Department of Electrical Engineering, Arizona State University, Tempe, AZ 85287-7206 USA (e-mail: karam@asu.edu).

Publisher Item Identifier S 1057-7149(02)01738-4.

Safranek and Johnston [10] introduced and validated a quantization noise detection model in the context of subband-based coding. The perceptual model accounts for contrast sensitivity, light adaptation, and texture masking. The technique is efficient since computations are performed in the linear transform domain and the transform is applied only once. Based on a maximum allowable perceptual quantization error, the coding scheme called the *perceptual image coder* (PIC), computes a different set of quantizer thresholds (one for each subband) for each individual image. However, since the quantization step size is fixed for all coefficients in a subband, this method is not able to adapt to the local changes in the masking characteristics, in addition to allocating too many bits to insensitive regions of the image. Tran and Safranek [11] proposed an improved perceptual model to better account for the local variations in masking. They introduced an image segmentation scheme in order to allow for local adaptation, but adaptation is still limited by the tradeoff between coding gain through localization and the cost to transmit the segmentation information.

The DCT of 8×8 image blocks is at the heart of several image and video compression standards (JPEG, MPEG, H.261) [12], [13]. A substantial part of the compression gained by using DCT-based methods is generated by quantization of the DCT coefficients. This quantization introduces distortion in the decompressed image. To achieve optimal compression, the DCT coefficients should be quantized as coarsely as possible while allowing minimal visible distortion in the decompressed image. Quantization is accomplished by division and rounding, according to the rule

$$u_{ijk} = \lfloor c_{ijk}/q_{ij} + 0.5 \rfloor \quad (1)$$

where c_{ijk} is the (i, j) th coefficient in the k th block (in raster scan order) and the values q_{ij} are the elements of the quantization matrix Q , which is transmitted as part of the coded image.

Ahumada *et al.* [14]–[16] developed a detection model to predict the visibility thresholds for the DCT coefficient quantization error

$$e_{ijk} = c_{ijk} - u_{ijk}q_{ij} \quad (2)$$

based on the luminance, chrominance, and contrast of the error as well as display characteristics and viewing conditions.

Based on this detection model, Watson introduced an iterative technique called DCTune [17]–[19] that, for a particular image, optimizes the quantization matrix Q for a given perceptual quality or a given bit rate. ISO/IEC DIS 10918-3 specifies an extension to the JPEG standard which allows for limited locally adaptive coding within an image by specifying a multiplier m_k that can be used to scale the quantization matrix for each block. Rosenholtz and Watson [20] proposed a technique that optimizes Q as well as the set of multipliers m_k for either a given perceptual image quality or a given bit rate. However, neither of these perceptually optimized methods is able to adapt to the full amount of masking available locally at each transform coefficient. While DCTune derives and uses a single fixed quantization matrix Q for all the blocks of the image, the second method requires additional side information and adaption is limited to uniformly scaling the fixed Q for each DCT block.

This paper presents a DCT-based, locally adaptive, perceptual-based image coder with the objective to minimize the bit rate for a desired perceptual target distortion. Our strategy is to exploit human visual masking properties which are derived in a locally adaptive fashion based on the local image characteristics. 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 the desired target perceptual distortion.

One main issue in developing a locally adaptive perceptual-based coding approach is that the image-dependent, locally varying, masking thresholds are needed both at the encoder and at the decoder in order to be able to reconstruct the coded visual data. 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. We attack this issue by estimating the locally available amount of masking at the decoder [21]. In this way, the local masking characteristics of the visual data can be exploited without having to transmit additional side information. The questions naturally arising from the limitations of the existing methods are 1) how large is the reduction in first order entropy of the quantized DCT coefficients when locally adaptive perceptual quantization [21] is used and 2) whether it is possible to implement such a locally adaptive quantization scheme with distortion control.

This paper is organized as follows. Section II introduces the perceptual distortion measure with respect to which the parameters of the proposed locally adaptive quantization scheme are optimized. Section III describes how the perceptual detection thresholds are computed. Section IV presents the proposed locally adaptive perceptual coding algorithm with perceptual distortion control. Coding results and comparison with the popular DCTune technique of Watson [17]–[19] are presented in Sections V and VI.

II. PERCEPTUAL DISTORTION METRIC

The first step in designing an image coder with a perceptual distortion control is to select a perceptual distortion metric. In the context of perceptual image compression, it is especially interesting to determine the lowest bit rate at which a distortion becomes just perceptible. This perceptual lossless point constitutes a natural target for high quality, or for perceptually lossless image coding, which is important in applications where loss in quality cannot be tolerated, such as medical imaging. The perceptual lossless point can also be used as a benchmark for performance comparison with other coders that use perceptual distortion metrics. A JND-based model, by definition, provides a natural way of determining the perceptually lossless point at the pixel or the transform coefficient level. While the JND thresholds provide a localized measure of the noise threshold for a single subband coefficient, a perceptual distortion metric must also account for the spatial and spectral summation of individual quantization errors.

In our case, the DCT decomposes the image into frequency- and orientation-selective subbands which differ in terms of their

sensitivity and masking properties. The n th subband consists of the collection of all n th DCT transform coefficients taken one at a time from subsequent blocks. Let $c_{i,j,k}$ be the DCT coefficient at location (i, j) in the k th 8×8 block, where k is the number of the DCT block in row-scan order, and $0 \leq i, j \leq 7$. Let the image consists of K_1 8×8 blocks along the vertical dimension, and K_2 8×8 blocks along the horizontal dimension. Then, the coefficients, $c_{i,j,k}$, of the DCT image blocks are mapped to DCT subbands, $c_{(b,n_1,n_2)}$, as follows: the subband number is $b = 8i + j$, and the coefficient locations within the subband are computed as $n_1 = \lfloor k/K_2 \rfloor$ and $n_2 = k - n_1 K_2$, respectively. All DCT transform coefficients that correspond to the same two-dimensional (2-D) DCT basis function are collected into one subband while preserving the spatial ordering of the DCT blocks.

The perceptual distortion metric that is used for the proposed adaptive image coder with perceptual distortion control is based on probability summation [22], [23]. The probability summation model considers a set of independent detectors, one at each subband location (b, n_1, n_2) . The probability, $P_{(b,n_1,n_2)}$, of detecting distortion at the location of a subband coefficient, is then the probability that detector (b, n_1, n_2) will signal the occurrence of a distortion. $P_{(b,n_1,n_2)}$ is determined by the psychometric function, which is commonly modeled as an exponential of the form

$$P_{(b,n_1,n_2)} = 1 - \exp\left(-\left|\frac{e(b,n_1,n_2)}{t_{JND}(b,n_1,n_2)}\right|^{\beta_b}\right) \quad (3)$$

where $e(b, n_1, n_2)$ is the quantization error at location (b, n_1, n_2) , $t_{JND}(b, n_1, n_2)$ denotes the detection (masking) threshold at location (b, n_1, n_2) , and β_b is a parameter whose value is chosen to maximize the correspondence of (3) with the experimentally determined psychometric function for a given type of distortion. In psychophysical experiments that examine summation over space, a value of β_b of about four has been observed to correspond well to probability summation [23]. Note that in (3), a quantization error, $e(b, n_1, n_2)$, that has a magnitude equal to the JND threshold results in a detection probability $P_{(b,n_1,n_2)} = 0.63$. This detection probability will be referred to as the perceptually lossless coding point.

A less localized probability of error detection can be computed by adopting the ‘‘probability summation’’ hypothesis [23] which pools the localized detection probabilities $P_{(b,n_1,n_2)}$ over a region of interest $R = \{(b, n_1, n_2) : (b, n_1, n_2) \text{ in a specified region}\}$. This probability summation scheme is based on two main assumptions.

- 1) *Assumption 1:* A distortion is detected in the region of interest R if and only if at least one detector in R signals the presence of a distortion, i.e., if and only if at least one of the distortions $e(b, n_1, n_2)$ is above threshold and, therefore, considered to be visible.
- 2) *Assumption 2:* The probabilities of detection $P_{(b,n_1,n_2)}$ are independent; i.e., the probability that a particular detector will signal the presence of a distortion is independent of the probability that any other detector will.

In the human visual system, highest visual acuity is limited to the size of the foveal region and covers approximately 2° of

visual angle. Let $\mathcal{F}_{(b,n_1,n_2)}$ denote the area in subband b that is centered at location (n_1, n_2) , and that covers 2° of visual angle. Then, $\mathcal{P}_{\mathcal{F}_{(b,n_1,n_2)}}$, the probability of detecting a distortion in the foveal region centered at (b, n_1, n_2) , can be written as

$$\mathcal{P}_{\mathcal{F}_{(b,n_1,n_2)}} = 1 - \prod_{(b',m_1,m_2) \in \mathcal{F}_{(b,n_1,n_2)}} (1 - P_{(b',m_1,m_2)}). \quad (4)$$

Substituting (3) in (4) results in

$$\mathcal{P}_{\mathcal{F}_{(b,n_1,n_2)}} = 1 - \exp\left(-\left(D_{\mathcal{F}_{(b,n_1,n_2)}}\right)^{\beta_b}\right) \quad (5)$$

where

$$D_{\mathcal{F}_{(b,n_1,n_2)}} = \left(\sum_{(b',m_1,m_2) \in \mathcal{F}_{(b,n_1,n_2)}} \left| \frac{e(b',m_1,m_2)}{t_{JND}(b',m_1,m_2)} \right|^{\beta_b} \right)^{1/\beta_b}. \quad (6)$$

In (6), $D_{\mathcal{F}_{(b,n_1,n_2)}}$ takes the form of a Minkowski metric with exponent β_b . From (5), it is clear that minimizing the probability of detecting a difference in the foveal region $\mathcal{F}_{(b,n_1,n_2)}$ is equivalent to minimizing the metric $D_{\mathcal{F}_{(b,n_1,n_2)}}$.

The distortion measure for a subband b , D_b , is defined as the maximum probability to detect a difference for any foveal region $\mathcal{F}_{(b,n_1,n_2)}$ in subband b . It is obtained by using a Minkowski metric with $\beta = \infty$ (which corresponds to a maximum operation) [17] for pooling the foveal distortions $D_{\mathcal{F}_{(b,n_1,n_2)}}$ in subband b

$$D_b = \max_{n_1, n_2} \left\{ D_{\mathcal{F}_{(b,n_1,n_2)}} \right\}. \quad (7)$$

Finally, the total distortion measure D for the whole image is defined to be the maximum probability of detection over all foveal regions and is obtained by using a Minkowski metric with $\beta = \infty$ for intraband and interband pooling [17] of the foveal distortions $D_{\mathcal{F}_{(b,n_1,n_2)}}$

$$D = \max_b \{D_b\} = \max_{(b,n_1,n_2)} \left\{ D_{\mathcal{F}_{(b,n_1,n_2)}} \right\}. \quad (8)$$

III. JND THRESHOLDS FOR DCT COEFFICIENTS

The perceptual model for the DCT coefficient quantization error provides one JND threshold, $t_{JND}(b, n_1, n_2)$, for each transform coefficient. Two visual phenomena are modeled to compute the thresholds: contrast sensitivity dependent on background luminance, and contrast masking. The JND thresholds $t_{JND}(b, n_1, n_2)$ are thus computed as

$$t_{JND}(b, n_1, n_2) = t_{DCT}(b, n_1, n_2) \cdot a_{CM}(b, n_1, n_2) \quad (9)$$

where $t_{DCT}(b, n_1, n_2)$ is the background luminance-adjusted contrast sensitivity threshold and $a_{CM}(b, n_1, n_2)$ is the contrast masking adjustment.

Background Luminance-Adjusted Contrast Sensitivity Threshold, $t_{DCT}(b, n_1, n_2)$: This is a measure, for each subband b (i.e., for each DCT basis function), of the smallest contrast that yields a visible signal over a background of uniform intensity. The inverse of the measured threshold

defines the sensitivity of the eye in function of the DCT basis functions' frequency and orientation and in terms of the background luminance.

The contrast sensitivity model considers the resolution of the screen, the viewing distance, the minimum display luminance, L_{\min} , and the maximum display luminance, L_{\max} . In order to make the coding results more portable, it is assumed that the display is gamma corrected. This means that prior to viewing, the signal intensity values are mapped such that they translate linearly into luminances.

The contrast sensitivity threshold $t_{DCT}(b, n_1, n_2)$ is computed as [14]

$$t_{DCT}(b(i, j), n_1, n_2) = \frac{MT_{i,j}(n_1, n_2)}{2\alpha_i\alpha_j(L_{\max} - L_{\min})} \quad (10)$$

where $T_{i,j}(n_1, n_2)$ is the background luminance-adjusted contrast sensitivity of the luminance error due to quantization of DCT coefficient $c_{i,j}$ in DCT block (n_1, n_2) , M is the number of gray levels, L_{\min} and L_{\max} are the minimum and the maximum display luminance, and the terms α_i and α_j are the DCT coefficient normalizing factors given by

$$\alpha_p = \frac{1}{\sqrt{N_{DCT}}} \begin{cases} 1, & p = 0 \\ \sqrt{2}, & p \neq 0. \end{cases} \quad (11)$$

N_{DCT} in (11) denotes the block size of the DCT and is equal to eight.

The value of $T_{i,j}(n_1, n_2)$ is based on the parametric model by Ahumada and Peterson [14] that approximates DCT coefficient contrast sensitivity by a parabola in log spatial frequency. $T_{i,j}(n_1, n_2)$ is computed as

$$T_{i,j}(n_1, n_2) = 10^{g_{i,j}(n_1, n_2)} \quad (12)$$

with

$$g_{i,j}(n_1, n_2) = \log_{10} \frac{T_{\min}(n_1, n_2)}{r + (1-r)\cos^2 \Theta_{i,j}} + K(n_1, n_2)(\log_{10} f_{i,j} - \log_{10} f_{\min}(n_1, n_2))^2. \quad (13)$$

The spatial frequency $f_{i,j}$ associated with DCT coefficient $c_{i,j}$, is computed as

$$f_{i,j} = \frac{1}{2N_{DCT}} \sqrt{\frac{i^2}{w_x^2} + \frac{j^2}{w_y^2}} \quad (14)$$

and the angle of its orientation, $\Theta_{i,j}$, is computed as

$$\Theta_{i,j} = \arcsin \frac{2f_{i,0}f_{0,j}}{f_{i,j}^2}. \quad (15)$$

The luminance-dependent parameters of the parabola are computed as

$$T_{\min}(n_1, n_2) = \begin{cases} \left(\frac{L(n_1, n_2)}{L_T}\right)^{\alpha_T} \frac{L_T}{S_0}, & L(n_1, n_2) \leq L_T \\ \frac{L(n_1, n_2)}{S_0}, & L(n_1, n_2) > L_T \end{cases} \quad (16)$$

$$f_{\min}(n_1, n_2) = \begin{cases} f_0 \left(\frac{L(n_1, n_2)}{L_f}\right)^{\alpha_f}, & L(n_1, n_2) \leq L_f \\ f_0, & L(n_1, n_2) > L_f \end{cases} \quad (17)$$

and

$$K(n_1, n_2) = \begin{cases} K_0 \left(\frac{L(n_1, n_2)}{L_K}\right)^{\alpha_K}, & L(n_1, n_2) \leq L_K \\ K_0, & L(n_1, n_2) > L_K. \end{cases} \quad (18)$$

The values of the constants in (13)–(18) are $r = 0.7$, $N_{DCT} = 8$, $L_T = 13.45$ cd/m², $S_0 = 94.7$, $\alpha_T = 0.649$, $f_0 = 6.78$ cycles/deg, $\alpha_f = 0.182$, $L_f = 300$ cd/m², $K_0 = 3.125$, $\alpha_K = 0.0706$, and $L_K = 300$ cd/m². The screen distance and resolution enter via w_x and w_y which are the horizontal and vertical size of a pixel in degree of visual angle, respectively. The local background luminance, $L(n_1, n_2)$, is computed as

$$L(n_1, n_2) = L_{\min} + \frac{L_{\max} - L_{\min}}{M} \cdot \left(\frac{\sum_{(0, m_1, m_2) \in \mathcal{F}(0, n_1, n_2)} c_{(0, m_1, m_2)}}{N_{DCT} \mathcal{N}(\mathcal{F}(0, n_1, n_2))} + m_g \right) \quad (19)$$

where M is the number of gray levels in the image, m_g is the global mean of the input image which is removed prior to the decomposition, and $\mathcal{N}(\mathcal{F}(0, n_1, n_2))$ denotes the number of subband coefficients contained in the foveal region at location (n_1, n_2) in subband zero.

With D denoting the viewing distance (distance between viewer and screen) in inches, R the display resolution in dots per inch, and θ the visual angle sustained by a foveal region (approximately 2°), $\mathcal{N}(\mathcal{F}(b, n_1, n_2))$ can be computed as follows:

$$\mathcal{N}(\mathcal{F}(b, n_1, n_2)) = \left(\left\lfloor \frac{2DR \tan(\theta/2)}{N_{DCT}} \right\rfloor \right)^2 \quad (20)$$

where the symbol $\lfloor \cdot \rfloor$ denotes rounding to the nearest smallest integer. For example, for a viewing distance $D = 24$ inches, a display resolution $R = 80$ dots per inch, $N_{DCT} = 8$, and $\theta = 2$ degrees, we obtain $\mathcal{N}(\mathcal{F}(b, n_1, n_2)) = 16$.

Contrast Masking Adjustment, $a_{CM}(b, n_1, n_2)$: Contrast masking refers to the change in the visibility of one image component (the target) by the presence of another one (the masker). It measures the variation of the detection threshold of a target signal as a function of the contrast of the masker. The resulting masking sensitivity profiles are referred to as target threshold versus masker contrast (TvC) functions [24], [25] and account for the change in the detection threshold as a function of the masker contrast. In our case, the masker signal is represented by the subband coefficients, $c_{(b, n_1, n_2)}$ of the input image to be coded while the target signal is represented by the quantization error.

The contrast masking model for DCT coefficient quantization noise is derived from a nonlinear transducer model for masking of sinusoidal gratings [24]. The nonlinear transducer model was

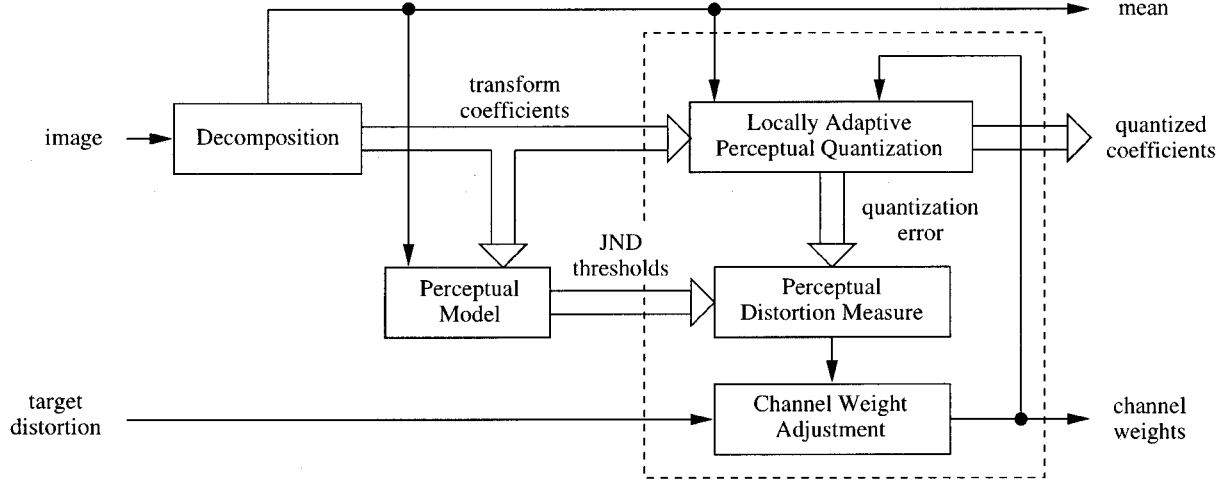


Fig. 1. Block diagram of an adaptive image coder with perceptual distortion control. Using a search method, the eight-bit channel weights are adjusted until the perceptual target distortion is obtained. The dashed rectangle identifies the iterative part of the encoder.

adapted for the considered band-limited subband components. The contrast masking adjustment $a_{CM}(b, n_1, n_2)$ is given by

$$a_{CM}(b, n_1, n_2) = \begin{cases} \max \left\{ 1, \left| \frac{c_{\mathcal{F}(b, n_1, n_2)}}{t_{DCT}(b, n_1, n_2)} \right|^{0.6} \right\}, & b \neq 0 \\ 1, & b = 0 \end{cases} \quad (21)$$

where $c_{\mathcal{F}(b, n_1, n_2)}$ is the average magnitude of the DCT coefficients in the foveal region $\mathcal{F}(b, n_1, n_2)$.

IV. ADAPTIVE IMAGE CODER WITH PERCEPTUAL DISTORTION CONTROL

The computed local masking thresholds $t_{JND}(b, n_1, n_2)$ are used to adaptively control the step-size $s(b, n_1, n_2)$ of a uniform quantizer while meeting a desired perceptual distortion D_T . Note that the step size varies for each coefficient $c(b, n_1, n_2)$ depending on the computed local amount of masking given by $t_{JND}(b, n_1, n_2)$.

The problem that arises when implementing a locally adaptive scheme is that the quantizer step size is intimately related to the data, and some form of side information needs to be transmitted to ensure reconstruction of the coded image. A locally adaptive quantization scheme that allocates bits for step size information will likely not result in better compression in spite of the improved removal of perceptually redundant information because too much of the bit budget is allocated for step size information. Our locally adaptive coding scheme eliminates the need for transmitting side information for the locally varying step size by estimating the available masking from the already received data and a prediction of the transform coefficient to be quantized. The proposed coding strategy exploits the correlation between the transform coefficients in the subbands by using linear prediction. This provides estimates for computing the amount of perceptual masking available at the location of each transform coefficient without requiring the transmission of side information. The support of the perceptual model needs to be sufficiently localized to allow masking threshold computation based on the already quantized data and the predicted value of the coefficient

to be quantized. In our case, the perceptual model uses the local mean (coefficients in DC subband), the already decoded coefficients, as well as the *predicted* value of the transform coefficient at the considered location (b, n_1, n_2) , to calculate an estimate $\hat{t}_{JND}(b, n_1, n_2)$ of the local noise tolerance.

For a uniform quantizer with step size $s(b, n_1, n_2)$, the maximum possible quantization error is $s(b, n_1, n_2)/2$. The JND quantization step size is defined as the step size at which the quantization errors are invisible [17] and is given by

$$s_{JND}(b, n_1, n_2) = 2\hat{t}_{JND}(b, n_1, n_2) \quad (22)$$

where $\hat{t}_{JND}(b, n_1, n_2)$ is the estimated detection threshold at location (b, n_1, n_2) . $\hat{t}_{JND}(b, n_1, n_2)$ is computed as discussed in Section III except that the foveal region $\mathcal{F}(b, n_1, n_2)$ is replaced by a causal foveal region that consists only of the already coded/decoded coefficients and a prediction of the considered coefficient $c(b, n_1, n_2)$. The prediction of is done using a linear, four-point, first-order predictor that uses the causal closest four neighbors to predict the coefficient $c(b, n_1, n_2)$ under consideration.

Distortion control is achieved by introducing a scalar channel weight, w_b , for each subband. Each channel weight w_b is used to scale the JND quantizer step sizes $s_{JND}(b, n_1, n_2)$ in subband b . The channel weights w_b , are adjusted using bisection [17], [26]. The procedure is repeated iteratively until the channel weights converge and the perceptual target distortion $D = D_T$ is just met. If eight-bit channel weights are used (values range from 0 to 255), the algorithm converges after at most nine iterations. Entropy coding is provided by the JPEG Huffman coding stage, with the difference that the quantization matrix Q is replaced by the matrix of the eight-bit channel weights. However, in our case, the entries w_b of Q are not directly used as step sizes. Instead, they are interpreted as weights for the local adaptive step sizes $s_{JND}(b, n_1, n_2)$. Note that the step sizes $s_{JND}(b, n_1, n_2)$ do not need to be transmitted since they are estimated at the decoder from the already decoded coefficients and a prediction of the transform coefficient to be quantized.

Fig. 1 shows a block diagram of the proposed adaptive image coder with perceptual distortion control. First, the image is de-

```

compute and remove global mean,  $m_g$ , of input image
decompose image into DCT subbands
for all channels  $b$  starting with the lowpass channel
{
  for all locations  $(n_1, n_2)$  compute  $t_{JND}(b, n_1, n_2)$  (Eqs. (9), (10), (21) )
  initialize  $w_b$ 
  do {
    for all locations  $(n_1, n_2)$ 
    {
      compute JND threshold estimate,  $\hat{t}_{JND}(b, n_1, n_2)$ 
      compute step size,  $s(b, n_1, n_2) = 2 w_b/G \hat{t}_{JND}(b, n_1, n_2)$  (Eq. (23))
      compute quantization error,  $e(b, n_1, n_2) = c_{(b, n_1, n_2)} - s(b, n_1, n_2) \left\lfloor \frac{c_{(b, n_1, n_2)}}{s(b, n_1, n_2)} + 0.5 \right\rfloor$ 
    }
    compute perceptual distortion,  $D_b = \max_{(n_1, n_2)} \{D_{\mathcal{F}_{(b, n_1, n_2)}}\}$  (Eq. (7))
    adjust  $w_b$  using bisection: increase  $w_b$  if  $D_b < D_T$ ; decrease  $w_b$  if  $D_b \geq D_T$ .
  } while ( $w_b$  has changed)
  quantize channel  $b$  using  $w_b$  as channel weight
}
entropy code all quantizer outputs and channel weights
(step-sizes  $s(b, n_1, n_2)$  generated at decoder and do not need to be transmitted)

```

Fig. 2. Pseudocode of the proposed locally adaptive perceptual coding algorithm with perceptual distortion control.

composed into channels or subbands (in the DCT case, by computing the block-based 8×8 DCT and remapping the DCT coefficients into subbands), the JND thresholds $t_{JND}(b, n_1, n_2)$ are computed, and the channel weights are initialized based on the perceptual distortion target, D_T . Subsequently, the transform coefficients or transform coefficient prediction residuals are quantized using the locally varying step sizes, $s(b, n_1, n_2)$, given by

$$\begin{aligned}
 s(b, n_1, n_2) &= \frac{w_b}{G} s_{JND}(b, n_1, n_2) \\
 &= 2 \frac{w_b}{G} \hat{t}_{JND}(b, n_1, n_2). \quad (23)
 \end{aligned}$$

The eight-bit channel weight, w_b , is divided by the factor G in order to allow step size adjustments to be small enough to obtain perceptual distortions close to target for a large range of perceptual target distortions, D_T . Without G , large target distortions would lead to saturation of the integer, eight-bit, step size multiplier w_b , while small target distortions would result in relative weight adjustments that are too large to provide fine distortion control. For example, if $G = 1$, the adjustment from $w_b = 2$ to $w_b = 1$ results in an effective step size change of 50%. The constant G should be optimized such that w_b is as close as possible to its maximum value ($w_b = 255$) without reaching saturation. The factor G depends on the target distortion D_T and is given by (see the Appendix)

$$G = \frac{192}{D_T} (\mathcal{N}(\mathcal{F}_{(b, n_1, n_2)}))^{1/\beta_b} \quad (24)$$

where β_b is the exponent of the Minkowski metric in (6) and is equal to four as discussed in Section II.

The factor G depends on the target distortion, D_T , and on the viewing conditions (distance and display resolution). Since the viewing conditions are fixed and known by the encoder as well as the decoder, only information about D_T needs to be added to the bit stream in order to allow inverse quantization at the decoder. For this purpose, the target distortion is quantized to eight bits and added to the bit stream.

Fig. 2 shows a pseudocode listing for the proposed DCT-based locally adaptive quantization with a perceptual distortion measure. After removing the mean and decomposing the image, each DCT subband is processed independently. First, the true (nonpredicted) thresholds $t_{JND}(b, n_1, n_2)$ are computed and are used in computing the subband distortions D_b . Quantization begins with the lowpass band since its values determine the estimate of the local luminance which is needed to compute the JND thresholds of all other bands.

Then, for each subband, an iterative bisection procedure is used to adjust the channel weight w_b such that the perceptual distortion target D_T is just met. At every iteration, the perceptual distortion measure of the current subband, D_b , is computed as in (7) using the true JND thresholds $t_{JND}(b, n_1, n_2)$. If $D_b \leq D_T$, the value of w_b is increased using bisection, otherwise it is decreased. The process is terminated when w_b has not changed with respect to the previous iteration. Since w_b is an eight-bit integer, the bisection process terminates after at most nine iterations. After the optimal value of the channel weight w_b has been determined, the corresponding subband b is quantized using the locally varying step sizes $s(b, n_1, n_2) = 2(w_b/G)\hat{t}_{JND}(b, n_1, n_2)$, and the algorithm proceeds to the next subband.

TABLE I
FIRST-ORDER ENTROPIES OF DCTUNE [17], PROPOSED LOCALLY ADAPTIVE QUANTIZATION, AND OPTIMAL QUANTIZATION WITH DISTORTION CONTROL FOR PERCEPTUALLY LOSSLESS QUANTIZATION ($D_T = 1$)

image	quantization method		
	DCTune [17]	locally-adaptive	optimal
Actor	1.473 bpp	1.275 bpp (-13.4%)	1.253 bpp (-14.9%)
Boat	1.208 bpp	0.995 bpp (-17.6%)	0.994 bpp (-17.7%)
Lena	1.008 bpp	0.905 bpp (-10.1%)	0.899 bpp (-10.8%)
Mandrill	1.639 bpp	1.544 bpp (-5.8%)	1.544 bpp (-5.8%)

The image decoder consists of an entropy decoding stage followed by a locally adaptive perceptual inverse quantization stage and an inverse DCT stage. After entropy decoding, inverse quantization is performed, beginning with the lowpass subband and then for each subsequent subband, by estimating the detection thresholds $\hat{t}_{JND}(b, n_1, n_2)$, as done at the encoder. Using (23), the estimated thresholds together with the channel weights w_b are utilized to compute the locally varying quantization step sizes, $s(b, n_1, n_2)$, that are needed for the inverse quantization. Then, an inverse DCT operation is applied, and the decoded image mean is added to the result in order to obtain the final decoded image.

V. CODING RESULTS

The locally adaptive perceptual quantization with distortion control has been applied to code 512×512 gray-scale images. The perceptual thresholds were optimized for a viewing distance of 60 cm on an 80 points-per-inch (ppi) display. A value of $\beta_b = 4$ was used for computing the perceptual distortion D_b for all bands. The above viewing conditions result in $\mathcal{N}(\mathcal{F}(b, n_1, n_2)) = 64$ for all bands.

Table I compares the first-order entropies obtained for $D_T = 1.0$ by using DCTune [17], [18], and the new coding scheme for perceptually lossless compression. In order to evaluate the performance penalty of using JND threshold estimates, the last column contains the first-order entropies that were obtained using the exact local JND thresholds to compute the quantizer step sizes (assuming no side information is needed to transmit the true JND thresholds). This provides an upper bound on the compression gain that can be achieved with the locally adaptive perceptual quantization.

For the same perceptual distortion, significant coding gain is obtained when using the new technique. The performance penalty for using estimates of the JND threshold, $\hat{t}_{JND}(b, n_1, n_2)$, instead of the exact values, $t_{JND}(b, n_1, n_2)$ is small compared to the total compression gain, which indicates that the estimated detection thresholds present good estimates to the available amount of masking.

Figs. 3 and 4 compare the performance of the proposed distortion-controlled locally adaptive perceptual coder with the DCTune coder [17], [18] in terms of PSNR and the perceptual distortion measure D of (8), respectively, for the Actor and the Lena images. Since the locally adaptive coder is optimized to adapt the quantization level to the local JND thresholds, and not to minimize MSE, the MSE-based PSNR does not well indicate the achieved performance gain. However, if the images are compared in terms of the perceptual distortion measure D , given by

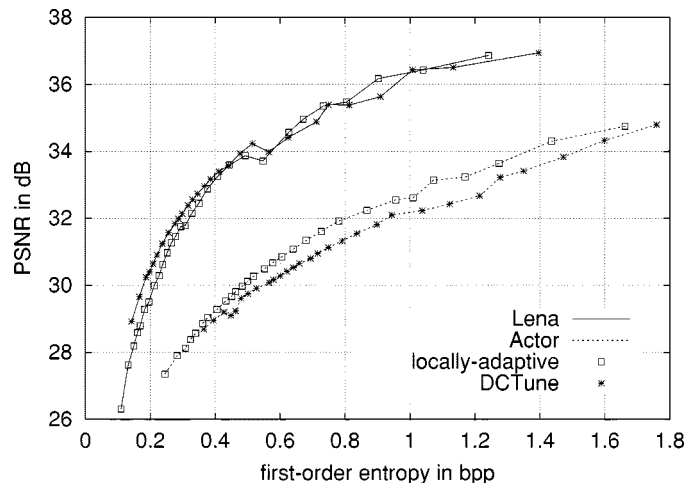


Fig. 3. PSNR as a function of bit rate for the Actor and Lena images coded using the proposed distortion-controlled locally adaptive perceptual image coder and DCTune [17], [18].

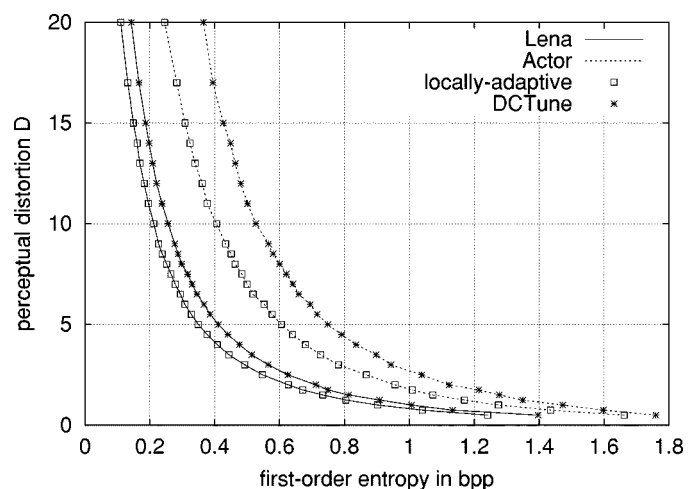


Fig. 4. Perceptual distortion D as a function of bit rate for the Actor and Lena images coded using the proposed distortion-controlled locally adaptive perceptual image coder and DCTune [17], [18].

TABLE II
FIRST-ORDER ENTROPIES FOR DCTUNE AND THE PROPOSED LOCALLY ADAPTIVE PERCEPTUAL CODER WITH DISTORTION CONTROL FOR HIGH QUALITY IMAGES (ALMOST TRANSPARENT, $D_T = 3.0$)

image	quantization method	
	DCTune [17]	locally-adaptive
Actor	0.943 bpp	0.781 bpp (-17.2%)
Boat	0.704 bpp	0.553 bpp (-21.4%)
Lena	0.565 bpp	0.491 bpp (-13.2%)
Mandrill	0.968 bpp	0.888 bpp (-8.2%)

TABLE III
BIT-RATES FOR DCTUNE AND THE PROPOSED LOCALLY ADAPTIVE PERCEPTUAL IMAGE CODER WITH DISTORTION CONTROL FOR HIGH QUALITY IMAGES (ALMOST TRANSPARENT, $D_T = 3.0$)

image	quantization method	
	DCTune [17]	locally-adaptive
Actor	1.073 bpp	0.860 bpp (-19.9%)
Boat	0.727 bpp	0.553 bpp (-23.9%)
Lena	0.591 bpp	0.498 bpp (-15.7%)
Mandrill	1.094 bpp	0.994 bpp (-9.1%)



Fig. 5. Coding example for DCT-based distortion-controlled locally adaptive perceptual coding for a target distortion $D_T = 3.0$, and comparison with DCTune [17]. (a) Original Lena image. (b) Non-locally adaptive coder with perceptual distortion control DCTune [17], $D_T = 3$, rate = 0.591 bpp. (c) Proposed locally adaptive coder with perceptual distortion control, $D_T = 3$, rate = 0.498 bpp

(8), the proposed locally adaptive coder clearly outperforms the nonadaptive coder at all rates.

Tables II and III compare both coders for high-quality or almost transparent perceptual quantization ($D_T = 3.0$). The coding gains in this situation are consistent with the gains for the perceptually lossless compression. The Lena images corresponding to the bit rates in Table III are shown in Fig. 5.

VI. PERCEPTUAL VALIDATION

A set of impairment tests has been conducted using the ASU *Image and Video Evaluation System* (IVES) which is described in more detail in [27] and [28]. The purpose of the tests was to compare the coding performance in terms of perceived quality for the proposed locally adaptive perceptual image coder with

that of DCTune [17], [18]. The subjects were shown a sequence of image combinations consisting of the original image and a decoded image at various bit rates. They were asked to use a five-point scale to evaluate the level of distortion present in the decoded image. The categories on the scale and their numerical values were “imperceptible” 5), “perceptible, not annoying” 4), “slightly annoying” 3), “annoying” 2), and “very annoying” 1).

The 512×512 gray-scale images were displayed on a NEC MultiScan P1150 monitor with a display resolution of 1152×864 pixels. A viewing distance of 60 cm (23.6 inches) was used, and the background intensity of the monitor was set to a gray-scale level of 128. The intensity to luminance relation of the setup used for the perceptual testing has been measured using a Photo Research SpectraColorimeter PR-650. For the measurements, the *Image and Video Evaluation System*

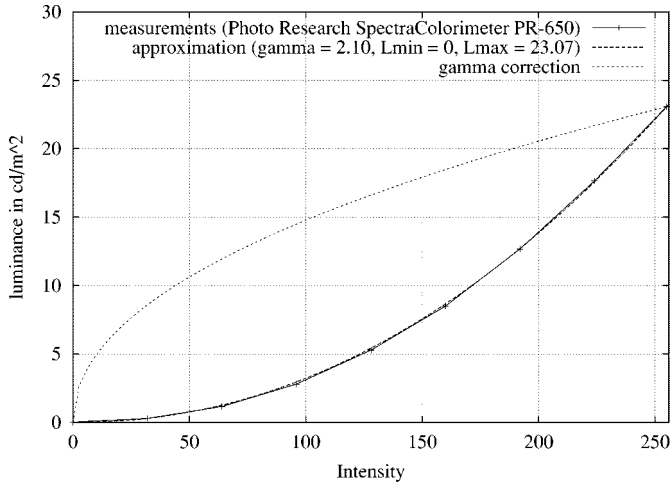


Fig. 6. Intensity to luminance measurements, analytic approximation, and gamma correction curve for the perceptual testing setup. The display used was a NEC MultiSync P1150 computer monitor.

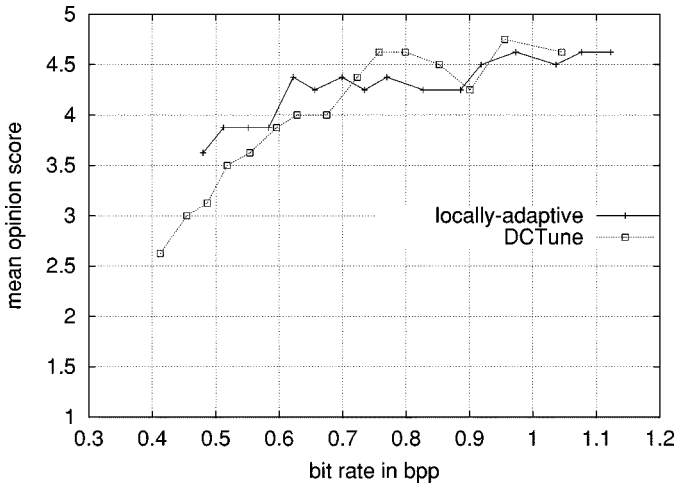


Fig. 7. Mean opinion score as a function of bit rate for the Actor image coded using distortion-controlled adaptive perceptual quantization and nonadaptive DCTune [17], [18].

(IVES) [27], [28] was used to display screen filling constant gray level images with intensities ranging from 0 to 255 (in increments of 32). The monitor settings were adjusted to prevent luminance saturation at high intensity levels. Fig. 6 shows the measurements, the analytic approximation, and the gamma correction curve that linearizes relationship between intensity and luminance.

The images were gamma-corrected prior to viewing since the perceptual model of the JND thresholds is luminance-based and assumes a linear display. The group of viewers consisted of 11 individuals with normal or corrected-to-normal vision. Some of them did have prior experience in image processing, and some did not.

The *mean opinion score* (MOS) results for the impairment test for the Lena image as a function of bit rate are shown in Fig. 7. They indicate that the proposed distortion-controlled adaptive perceptual coder does result in an improvement in perceived image quality compared to DCTune [17], [18], particularly at low bit rates.

APPENDIX DERIVATION OF G (24)

It follows from the definitions of the perceptual distortion measure in (6)–(8) that

$$\begin{aligned}
 D_T \leq D &= \max_{(b, n_1, n_2)} \left(\sum_{(b', m_1, m_2) \in \mathcal{F}_{(b, n_1, n_2)}} \left| \frac{e(b', m_1, m_2)}{t_{JND}(b', m_1, m_2)} \right|^{\beta_b} \right)^{1/\beta_b} \\
 &\leq \left(\sum_{(b', m_1, m_2) \in \mathcal{F}_{(b, n_1, n_2)}} \left(\frac{s(b, m_1, m_2)}{2t_{JND}(b, m_1, m_2)} \right)^{\beta_b} \right)^{1/\beta_b} \\
 &\leq \left(\sum_{(b', m_1, m_2) \in \mathcal{F}_{(b, n_1, n_2)}} \left(\frac{w_b}{G} \right)^{\beta_b} \right)^{1/\beta_b} \\
 &\leq \frac{w_b}{G} (\mathcal{N}(\mathcal{F}_{(b, n_1, n_2)}))^{1/\beta_b} \quad (25)
 \end{aligned}$$

where $\mathcal{N}(\mathcal{F}_{(b, n_1, n_2)})$ denotes the number of subband coefficients contained in the foveal region centered at location (n_1, n_2) in subband b . In the before last step, the definition of the step size (23) was used and the JND threshold has been approximated by its estimate that is computed based on the already quantized coefficients, $\hat{t}_{JND}(b, n_1, n_2) \approx t_{JND}(b, n_1, n_2)$.

Using the saturation point $w_b = 255$, an estimate of the upper bound for G is obtained which can be expressed as

$$G \lesssim (\mathcal{N}(\mathcal{F}_{(b, n_1, n_2)}))^{1/\beta_b} \frac{255}{D_T}. \quad (26)$$

In order to allow enough margin to prevent saturation when adjusting the step size multipliers, w_b , G is set to the 75th percentile of the bound estimate

$$G = \frac{192}{D_T} (\mathcal{N}(\mathcal{F}_{(b, n_1, n_2)}))^{1/\beta_b}. \quad (27)$$

REFERENCES

- [1] J. L. Mannon and J. D. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inform. Theory*, vol. 20, no. 4, pp. 525–536, 1974.
- [2] N. B. Nill, "A visual model weighted cosine transform for image compression and quality assessment," *IEEE Trans. Commun.*, vol. COM-33, pp. 551–557, July 1985.
- [3] J. D. Eggerton and M. D. Srinath, "A visually weighted quantization scheme for image bandwidth compression at low data rate," *IEEE Trans. Commun.*, vol. COM-34, pp. 840–847, Aug. 1986.
- [4] J. A. Saghri, P. S. Cheatham, and A. Habibi, "Image quality measure based on a human visual system model," *Opt. Eng.*, vol. 28, no. 7, pp. 813–818, 1989.
- [5] B. Macq, "Weighted optimum bit allocations to orthogonal transforms for picture coding," *IEEE J. Select. Areas Commun.*, vol. 10, pp. 875–883, June 1992.
- [6] S. Daly, "Digital image compression and transmission system with visually weighted transform coefficients," U.S. Patent 4 780 761, 1995.
- [7] N. Farvardin, X. Ran, and C. C. Lee, "Adaptive DCT coding of images using entropy-constrained trellis coded quantization," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1993, pp. 397–400.
- [8] J. H. Kasner, M. W. Marcellin, and B. R. Hunt, "Universal TCQ in wavelet image coding," in *Proc. 31st Asilomar Conf. Signals, Systems, Computers*, 1997, pp. 1279–1283.

- [9] I. Höntsch, L. Karam, and R. Safranek, "A perceptually tuned embedded zerotree image coder based on set partitioning," in *Proc. IEEE Int. Conf. Image Processing*, 1997.
- [10] R. J. Safranek and J. D. Johnston, "A perceptually tuned subband image coder with image dependent quantization and post-quantization," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1989, pp. 1945–1948.
- [11] T. D. Tran and R. Safranek, "A locally adaptive perceptual masking threshold model for image coding," in *IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 1996.
- [12] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 31–43, 1991.
- [13] D. LeGall, "MPEG: A video compression standard for multimedia applications," *Commun. ACM*, vol. 34, no. 4, pp. 46–58, 1991.
- [14] A. J. Ahumada and H. A. Peterson, "Luminance-model-based DCT quantization for color image compression," in *Proc. Human Vision, Visual Processing, Digital Display III*, 1992, pp. 365–374.
- [15] H. A. Peterson, A. J. Ahumada, and A. B. Watson, "An improved detection model for DCT coefficient quantization," in *Proc. Human Vision, Visual Processing, Display VI*, 1993, pp. 191–201.
- [16] J. A. Solomon, A. J. Ahumada, and A. B. Watson, "Visibility of DCT basis functions: Effects of contrast masking," in *Proc. Data Compression Conf.*, 1994, pp. 361–370.
- [17] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," in *Soc. Information Display Dig. Tech. Papers XXIV*, 1993, pp. 946–949.
- [18] —, "DCT quantization matrices visually optimized for individual images," in *Proc. Human Vision, Visual Processing, Digital Display IV*, B. E. Rogowitz, Ed., 1993, pp. 202–216.
- [19] —, "Visually optimal DCT quantization matrices for individual images," in *Proc. Data Compression Conf.*, 1993, pp. 178–187.
- [20] R. Rosenholtz and A. B. Watson, "Perceptual adaptive JPEG coding," in *IEEE Int. Conf. Image Processing*, 1996, pp. 901–904.
- [21] I. Höntsch and L. J. Karam, "Locally adaptive perceptual image coding," *IEEE Trans. Image Processing*, vol. 9, pp. 1472–1483, Sept. 2000.
- [22] A. B. Watson, "Probability summation over time," *Vis. Res.*, vol. 19, pp. 515–522, 1979.
- [23] J. G. Robson and N. Graham, "Probability summation and regional variation in contrast sensitivity across the visual field," *Vis. Res.*, vol. 21, pp. 409–418, 1981.
- [24] J. M. Foley and G. M. Boynton, "A new model of human luminance pattern vision mechanisms: Analysis of the effects of pattern orientation, spatial phase and temporal frequency," *Proc. SPIE*, vol. 2054, pp. 32–42, 1994.
- [25] J. M. Foley, "Human luminance pattern-vision mechanisms: Masking experiments require a new model," *J. Compar. Neurol.*, vol. 11, no. 6, pp. 1710–1719, 1994.
- [26] A. B. Watson, "Image data compression having minimum perceptual error," U.S. Patent 5 426 512, 1995.

- [27] S. Bellofiore, L. J. Karam, W. Metz, and T. Acharya, "A flexible and user-friendly image quality assessment system," in *Int. Conf. Signal Image Processing*, 1997, pp. 51–54.
- [28] S. Bellofiore and L. J. Karam, "Image and video evaluation system final report," Tech. Rep., Arizona State Univ., Tempe, July 1998.

Ingo Höntsch (S'93–M'99) received the Dipl.-Ing. degree in electrical engineering from the Dresden Institute of Technology, Dresden, Germany, in 1993, and the Ph.D. degree in electrical engineering from Arizona State University (ASU), Tempe, in 1999.

He was a Research Assistant with ASU and was with SAP AG. In 1999, he joined the Institut für Rundfunktechnik (IRT), Munich, Germany, where is currently a Member of Research Staff with the Television Production Technology Department. His research interests include image and video compression, human visual perception, multimedia, signal processing, and, more recently, video production systems integration, technology migration, and interoperability. He serves as Chair of the EBU P/PITV Working Group.

Dr. Höntsch is a member of the IEEE Signal Processing and Communications Societies and SMPTE.



Lina J. Karam (S'91–M'95) received the B.E. degree in electrical engineering from the American University of Beirut, Beirut, Lebanon, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1992 and 1995, respectively.

She is currently an Associate Professor with the Department of Electrical Engineering, Arizona State University, Tempe. Her research interests are in the areas of image and video processing, image and video coding, error-resilient source coding, and digital filtering. From 1991 to 1995, she was a Graduate Research Assistant with the Graphics, Visualization, and Usability Center and then with the Department of Electrical Engineering, Georgia Institute of Technology, Atlanta. In 1992, she was with Schlumberger Well Services working on problems related to data modeling and visualization. In 1994, she was with the Signal Processing Department, AT&T Bell Labs, working on problems in video coding.

Dr. Karam is the recipient of an NSF CAREER Award. She served as Chair of the IEEE Communications and Signal Processing Chapters in Phoenix, AZ, in 1997–1998. She is currently an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING and a member of the IEEE Circuits and Systems Technical Committee. She is a member of the IEEE Signal Processing, Circuits and Systems, and Communications Societies.