JPEG2000 Encoding With Perceptual Distortion Control

Zhen Liu, Lina J. Karam, *Senior Member, IEEE*, and Andrew B. Watson

***Abstract*—In this paper, a new encoding approach is proposed to control the JPEG2000 encoding in order to reach a desired perceptual quality. The new method is based on a vision model that incorporates various masking effects of human visual perception and a perceptual distortion metric that takes spatial and spectral summation of individual quantization errors into account. Compared with the conventional rate-based distortion minimization JPEG2000 encoding, the new method provides a way to generate consistent quality images at a lower bit rate.**

***Index Terms*—Distortion control, embedded coding, human visual system (HVS), JPEG2000, perception.**

I. Introduction

JPEG2000 is the new still image compression standard. Like other image and video compression standards, JPEG2000 only specifies the decoder compatible bit-stream syntax, and has left enough room for innovations in the encoder and decoder design. This work is motivated by the following observations of current JPEG2000 code design.

Current publicly available JPEG2000 code (Verification Model [1], JASPER [2], JJ2000 [3], and Kakadu [4]) all adopted the rate-based distortion minimization encoding approach. This approach requires the end user to specify the desired bit rate (for SNR progressive, a set of increasing rates). The encoder tries to meet the rates while minimizing the distortion. The adopted distortion is typically a weighted mean squared error (MSE)-based distortion. For a lot of applications, the end user prefers a consistent quality. Different images when coded at the same bit rate can result in different visual qualities. Therefore, finding a rule to determine the bit rate that results in a desired image quality is not an easy task. For visual information, the human eye is the ultimate receiver. When considering image quality, one definitely should take the human visual system (HVS) into consideration.

JPEG2000 [4] consists of two encoding stages known as tier-1 and tier-2 coding, respectively. After a wavelet-transform stage, the image subbands are divided into equal-size coding blocks. In order to minimize the (weighted) MSE for the desired bit rate, a rate-distortion optimization process is adopted between the tier-1 and tier-2 stages. In tier-1 coding, each coding block (typically of size 64 × 64 or 32 × 32) is independently bit-plane coded from the most significant bit-plane (MSB) to the least significant bit-plane (LSB) using three coding passes (except for the MSB which is coded using only one "clean up" coding pass). For M bitplanes, this results in a total number of $(3M - 2)$ coding passes. An embedded bit-stream is then generated for each coding block. The distortion reduction and rate increase associated with each coding pass is collected. This information is then used by a post-compression rate-distortion optimization (PCRD-opt) stage, which is a rate control procedure used to determine each coding block's contribution to the final bit-stream. In tier-2 coding, those included coding passes from each coding block are organized into a final bit-stream; those not included are discarded. With a carefully optimized discrete wavelet transform (DWT) implementation, the embedded block coding tends to dominate the whole encoding time [5], [6]. So, current encoders waste computational power and memory on those finally discarded coding passes.

Visual progressive weighting (VIP) [4, Ch. 16, Sec. 16.1.1] and visual masking [4, Ch. 16, Sec. 16.1.4] have been provided as an option in some JPEG2000 implementations (Verification Model [1] and Kakadu [4]) by setting the weights of the weighted MSE distortion based on the human visual system Contrast Sensitivity Function (CSF) and some local visual masking effects [7]. However, the existing Part 1 JPEG2000-compliant VIP scheme [4, Ch. 16] uses a single weight for an entire coding block; this weight can only be progressively adjusted for different quality layers. In addition, as stated in [4, Ch. 16], a main limitation of the existing JPEG2000 perceptual weighting schemes is that it is only possible to alter the number of coding passes contributed by whole coding blocks to any given quality layer. But, even when using the smaller size 32 × 32 coding blocks, these coding blocks correspond to large portions of the relevant image components. In fact, significant subbands in the wavelet decomposition could consist of very few and even one coding block, which would then prohibit the exploitation of the local masking variations in these subbands. These limitations are partially addressed by the visual optimization tools proposed in [8]; but these latter tools are not compatible with the baseline (Part 1) JPEG2000 decoder and are, hence, included in Part 2 of the standard. Furthermore, the existing tools do not fully exploit variations in the local characteristics of the visual data such as local light adaptation.

In addition, the existing JPEG2000 (Part 1 and Part 2) schemes are "rate-based" schemes, meaning that they attempt

Manuscript received January 9, 2004; revised June 14, 2005. 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 Dr. David S. Taubman.

Z. Liu is with Qualcomm, San Diego, CA 92121 USA (e-mail: zhenl@qualcomm.com).

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

A. B. Watson is with NASA Ames Research Center, Moffett Field, CA 94035 USA (e-mail: abwatson@mail.arc.nasa.gov).

Digital Object Identifier 10.1109/TIP.2006.873460

to minimize the distortion (perceptual or MSE) at a given specified bit-rate and they all make use of the PCRD-opt stage. These existing schemes mainly make use of a specified bit-rate (or a set of bit-rates in the SNR progressive mode) rather than a specified distortion for determining the bit-stream truncation point for each coding block even when generating different quality layers (SNR progressive mode). More recently, strategies that truncate the bit-stream based on the rate-distortion slope values [4, Section 8.2.1], have been proposed in order to achieve more consistent quality [9]. However, none of the existing proposed tools provide a way to control the encoding process to reach a desired quality as we are proposing.

This paper presents a new distortion-based JPEG2000 encoding scheme that is fully compatible with Part 1 of the JPEG2000 standard, with the objective to minimize the bit rate for a desired target quality. The proposed scheme achieves the minimum bit-rate for a given desired distortion in a noniterative manner and directly as part of the tier-1 coding stage. So, there is no need for a tier-2 coding stage. Although any quality metric, including MSE-based ones, can be used with the proposed encoding scheme, we adopt a perceptual-based distortion metric in order to exploit the luminance and contrast masking properties of the HVS. The multiresolution wavelet decomposition and the two-tier coding structure of JPEG2000 make it extremely suitable to incorporate a locally adaptive HVS model into the JPEG2000 coder design.

Our method is based on a vision model that incorporates various masking effects of human visual perception and a perceptual distortion metric that takes spatial and spectral summation of individual quantization errors into account. Given the wavelet transformed coefficients, the coder adaptively computes the local distortion sensitivity profiles in the form of detection thresholds that adapt to the varying local frequency, orientation and spatial characteristics of the visual information. The derived thresholds are then used to control the embedded bitplane coding in order to meet the desired target perceptual quality.

This paper is organized as follows. In Section II, the adopted perceptual model is introduced. The perceptual detection thresholds and the perceptual distortion metric are described in details. In Section III, the existing rate-based JPEG2000 encoding approach is discussed. The new JPEG2000 encoding with perceptual distortion control is presented in Section IV. Experimental results and comparison with the conventional rate-based JPEG2000 encoding are reported in Section V. A conclusion is given in Section VI.

II. Perceptual Model

The first step in designing an image coder with perceptual distortion control is to select a perceptual vision model. The HVS perception can be modeled as a collection of channels that are selective in terms of frequency and orientation [10]. Each channel responds to a range of frequencies (frequency bandwidth) and orientations (orientation bandwidth) about some preferred center frequency and orientation (f_s, θ_s). Different channels have different center frequencies and orientations that span the full range of visible frequencies and the full 180° range of orientations. Measurement of the receptive fields in the visual cortex revealed that these multi-channel frequency- and orientation-selective components exhibit approximately a dyadic structure [11]. This can be approximated by the dyadic structure of the pyramid wavelet decomposition which can decompose the input image into frequency- and orientation-selective visual components that differ in terms of their sensitivity and visual masking properties.

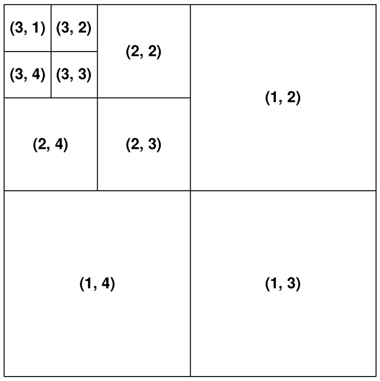


Fig. 1. Index of DWT subbands. Each subband is identified by a pair of integers (λ, θ), where λ is the level and θ is the orientation.

Perceptual-based coding algorithms attempt to discriminate between signal components which are and are not detected by the human receiver [12]. The main idea in perceptual coding are 1) to "hide" the coding distortion beneath the detection threshold, and 2) to augment the classical coding paradigm of redundancy removal with elimination of perceptually irrelevant signal information. This is typically achieved by exploiting the masking properties of the HVS [13]–[17] and establishing detection thresholds of just-noticeable distortion (JND) and minimally noticeable distortion (MND) based on psycho-physical masking phenomena [18], [19]. This usually requires computing and making use of image-dependent, locally varying masking thresholds. The two-tier coding structure of JPEG2000 gives the coder great flexibility in computing and using these locally varying thresholds without the need to send any side information. The clear separation of coding and bit stream formation provided by the JPEG2000 two-tier coding structure allows the precise computation of the locally varying masking thresholds and precise distortion control based on the true transformed coefficients.

A. JND Thresholds for Wavelet Coefficients

JPEG2000 makes use of the discrete wavelet transform, which decomposes the image into frequency and orientation selective subbands. The perceptual model needs one JND threshold, $t_{JND}(\lambda, \theta, i, j)$, for each DWT transformed coefficient at location (i, j) within subband (λ, θ), where λ is the transform level and θ is the orientation. The orientations are indexed as 1, 2, 3, 4 corresponding to the LL, HL, HH, and LH subbands, respectively, where low (denoted by L) and high (denoted by H) are in the order horizontal-vertical as illustrated in Fig. 1. In this work, three visual phenomena are modeled to compute the JND thresholds: contrast sensitivity, luminance

TABLE I
BASIS FUNCTION AMPLITUDE $A_{\lambda,\theta}$ FOR A SIX-LEVEL 9/7 DWT

	DWT decomposition level					
Orient	1	2	3	4	5	6
LL	0.62171	0.34537	0.18004	0.09140	0.045943	0.023013
LH,HL	0.67234	0.41317	0.22727	0.11792	0.059758	0.030018
HH	0.72709	0.49428	0.28688	0.15214	0.077727	0.039156

masking (also known as light adaption), and contrast masking. The JND thresholds $t_{JND}(\lambda,\theta,i,j)$ are thus computed as

$$t_{JND}(\lambda,\theta,i,j) = JND_{\lambda,\theta} a_l(\lambda,\theta,i,j) a_c(\lambda,\theta,i,j) \quad (1)$$

where $JND_{\lambda,\theta}$ is the base detection threshold for a subband (λ,θ), $a_l(\lambda,\theta,i,j)$ is the luminance masking adjustment, and $a_c(\lambda,\theta,i,j)$ is the contrast masking adjustment.

The base detection threshold $JND_{\lambda,\theta}$ is a measure, for each subband (λ,θ) (i.e., for each DWT basis function), of the smallest contrast that yields a visible signal over a background of uniform intensity. Contrast is a key concept in vision science because the response of the HVS depends much less on the absolute luminance than on the variation of the signal relative to the surround background, a property known as Weber-Fechner law [20]. Contrast is a measure of this relative variation. The necessary contrast to provoke a response from the HVS is defined as the detection threshold. The inverse of the detection threshold is the contrast sensitivity. Contrast sensitivity varies with frequency. With the increase of spatial frequency, the contrast sensitivity decreases. This important property of the HVS can be characterized by the contrast sensitivity function and can be measured using detection experiments.

In detection experiments, the tested subject is presented with test images and needs only to specify whether the target stimulus is visible or not visible. In [21], the base detection thresholds for the 9/7 discrete wavelet transform are measured using a two-alternative forced choice (2AFC) procedure. By fitting the data from the experimental results, a mathematical model for the JND threshold is formulated as

$$JND_{\lambda,\theta}(r) = \frac{1}{A_{\lambda,\theta}} a 10^{k\left\{\log\left(\frac{g_\theta f_0 2^\lambda}{r}\right)\right\}^2} \quad (2)$$

where a, k, f_0 and g_θ are constants, $A_{\lambda,\theta}$ is the amplitude of the DWT 9/7 basis function corresponding to level λ and orientation θ, and r is the visual resolution of the display in pixels/degree, which can be calculated as

$$r = dv \tan\left(\frac{\pi}{180}\right) \approx d\frac{v\pi}{180} \approx d\frac{v}{57.3} \quad (3)$$

where v is the viewing distance in cm and d is the display resolution in pixels/cm. Table I lists the $A_{\lambda,\theta}$ values for a 6-level DWT decomposition. Table II lists the constants a, k, f_0 and g_θ obtained with a background intensity of 128 [21]. The inverse of the base threshold defines the sensitivity of the eye in function of the DWT basis functions' frequency and orientation and with a background intensity of 128.

TABLE II
PARAMETERS FOR THE DWT THRESHOLD MODEL FOR THE Y CHANNEL [21]

a	k	f_0	g_{LL}	$g_{HL,LH}$	g_{HH}
0.495	0.466	0.401	1.501	1.0	0.534

In (2), the base detection threshold is obtained by fitting the data collected at one mean background intensity. However, the detection threshold varies with the background intensity levels, which is called light adaptation or luminance masking [20]. In image coding, the detection thresholds will depend on the mean luminance of the local image region and, therefore, a luminance masking correction factor must be derived and applied to the contrast sensitivity profile to account for this variation.

The effect of background intensity level upon the detection threshold is complex. It involves both vertical and horizontal shifts of the contrast sensitivity function [22]. In this work, the luminance masking adjustment is approximated using a power function [22]

$$a_l(\lambda,\theta,i,j) = \left(\frac{v_{\lambda_{\max},LL,i',j'}}{v_{\text{mean}}}\right)^{a_T} \quad (4)$$

where $\lambda_{\max}$ is the highest level of the DWT decomposition and is set to 5 in this work, v_{mean} is the LL subband constant corresponding to the mean luminance of the display (128 for an unsigned 8-bit image). In (4), $v_{\lambda_{\max},LL,i',j'}$ is the value of the DWT coefficient, in the LL subband, that spatially corresponds to location (λ,θ,i,j). In this case, i' and j' can be calculated as $i' = \lfloor i/2^{\lambda_{\max}-\lambda} \rfloor$ and $j' = \lfloor j/2^{\lambda_{\max}-\lambda} \rfloor$, where $\lfloor\,\rfloor$ is the operation of rounding to the nearest smaller integer. The parameter a_T controls the degree to which luminance masking occurs and takes a value of 0.649 [22]. Note that luminance masking can be suppressed by setting $a_T = 0$.

Another factor that will affect the detection threshold is the contrast masking which takes into account the fact that the visibility of one image component (the target) changes with the presence of other image components (the masker) [16], [23], [24]. Contrast masking 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 functions. Within the framework of image coding, the masker signal is usually represented by the subband coefficients of the input image to be coded while the target signal is represented by the distortion or coding noise. The contrast masking effect can be modeled as

$$a_c(\lambda,\theta,i,j) = a_{c\_\text{self}}(\lambda,\theta,i,j) a_{c\_\text{neig}}(\lambda,\theta,i,j) \quad (5)$$

where $a_{c\_\mathrm{self}}(\lambda, \theta, i, j)$ is the self contrast masking adjustment factor, and $a_{c\_\mathrm{neig}}(\lambda, \theta, i, j)$ is the neighborhood contrast masking adjustment factor.

The self contrast masking adjustment, $a_{c\_\mathrm{self}}(\lambda, \theta, i, j)$, is a measure of the increase in the detection threshold at the location (λ, θ, i, j) due to a sufficiently large coefficient at that location. For the DWT coefficients, it is modeled by the nonlinear transducer model introduced by Teo and Heeger [25], which is a piecewise linearization of the contrast masking results demonstrated by Foley [23], [24]. This model can be expressed as [26]

$$a_{c\_\mathrm{self}}(\lambda, \theta, i, j) = \max\left\{1, \left(\frac{|v(\lambda, \theta, i, j)|}{JND_{\lambda,\theta} a_l(\lambda, \theta, i, j)}\right)^{\epsilon}\right\} \quad (6)$$

where $v(\lambda, \theta, i, j)$ is the DWT coefficient value at location (λ, θ, i, j). For the LL subband, contrast masking is suppressed by setting $\epsilon = 0$. For other subbands, ϵ is set to 0.6.

The neighborhood contrast adjustment, $a_{c\_\mathrm{neig}}(\lambda, \theta, i, j)$, takes into consideration the fact that, in the reconstructed images, the signal (DWT basis function) determined by a wavelet transformed coefficient is superimposed on other signals determined by the neighboring wavelet coefficients. There is some masking effect contributed from spatially neighboring signals due to the phase uncertainty, receptive field size, as well as possible longer range effects that increase detection [27]. It is obtained by a slight modification of the model adopted in [8] as follows:

$$\begin{aligned} &a_{c\_\mathrm{neig}}(\lambda, \theta, i, j) \\ &= \max\left\{1, \sum_{k \in \text{neighbors of } (\lambda,\theta,i,j)} \frac{\left|\frac{v_k}{JND_{\lambda,\theta} a_l(\lambda,\theta,i,j)}\right|^{\zeta}}{N_{i,j}}\right\} \end{aligned} \quad (7)$$

where the neighborhood consists of the coefficients in the same subband that lie within a window centered at the location (i, j), $N_{i,j}$ denotes the number of coefficients in that neighborhood, v_k are the neighboring DWT coefficient values, and ζ is a constant that controls the influence of the amplitude of each neighboring coefficient. In (7), the parameter ζ, together with $N_{i,j}$, are used to control the degree of neighborhood masking. However, as Watson points out in [22], the contrast masking is strongest when both components are of the same spatial frequency, orientation, and location. Therefore, a contrast masking model based solely on self contrast masking provides a good conservative model with much reduced complexity.

B. Perceptual Distortion Metric

While the JND threshold profile provides a localized measure of the noise threshold for a single wavelet coefficient, a perceptual distortion metric that also accounts for spatial and spectral summation of individual quantization errors is needed. In this work, the probability summation model is adopted [26], [28], [29].

The probability summation model considers a set of independent detectors, one at each location (λ, θ, i, j). The probability, $p_{(\lambda,\theta,i,j)}$, of detecting a distortion at that location, is then the probability that detector (λ, θ, i, j) will signal the occurrence of a distortion. The probability $p_{(\lambda,\theta,i,j)}$ is determined by the psychometric function, which is commonly modeled as an exponential of the form

$$p_{(\lambda,\theta,i,j)} = 1 - \exp\left(-\left|\frac{e(\lambda, \theta, i, j)}{t_{JND}(\lambda, \theta, i, j)}\right|^{\beta}\right) \quad (8)$$

where $e(\lambda, \theta, i, j)$ is the quantization error at location (λ, θ, i, j), $t_{JND}(\lambda, \theta, i, j)$ denotes the detection threshold at location (λ, θ, i, j), and β is a parameter whose value is chosen to maximize the correspondence of (8) with the experimentally determined psychometric function for a given type of distortion. In the psychophysical experiments that examine summation over space, a value of β of about four has been observed to correspond well to probability summation [29]. Notice that in (8), a quantization error, $e(\lambda, \theta, i, j)$, that has a magnitude equal to the JND threshold results in a detection probability $p(\lambda, \theta, i, j) = 0.63$ [28]. This detection probability usually is referred to as the perceptually lossless coding point.

A less localized probability of error detection can be computed by adopting the "probability summation" hypothesis which pools the localized detection probabilities $p_{(\lambda,\theta,i,j)}$ over a region of interest [29]. In the human visual system, highest visual acuity is limited to the size of the foveal region and covers approximately $\alpha = 2^\circ$ of visual angle. Let $\mathcal{F}_{(n_1,n_2)}$ denote the area in the spatial domain that is centered at location (n_1, n_2) and covers 2° of visual angle. The number of pixels contained in that foveal region, $\mathcal{N}(\mathcal{F}_{(n_1,n_2)})$, can be computed as

$$\mathcal{N}\left(\mathcal{F}_{(n_1,n_2)}\right) = \left(\left\lfloor 2dv \tan\left(\frac{\alpha}{2}\right)\right\rfloor\right)^2 \approx \left(\lfloor 2r \rfloor\right)^2 \quad (9)$$

where r is the display visual resolution (in pixels/degree) calculated in (3).

Let F denote the set of DWT coefficients whose values affect the $\mathcal{F}_{(n_1,n_2)}$ reconstruction. Then, $\mathcal{P}_{\mathcal{F}_{(n_1,n_2)}}$, the probability of detecting a distortion in the foveal region centered at (n_1, n_2), can be written as

$$\mathcal{P}_{\mathcal{F}_{(n_1,n_2)}} = 1 - \prod_{(\lambda,\theta,i,j) \in F} \left(1 - p_{(\lambda,\theta,i,j)}\right). \quad (10)$$

This probability summation pools the probability of detecting an error in a foveal region over all DWT coefficients that affect its reconstruction. It is based on the following two assumptions. First, a distortion is detected in a foveal region if and only if at least one detector affecting its reconstruction signals the presence of a distortion, i.e., if and only if at least one of the distortions $e(\lambda, \theta, i, j)$ is above threshold and, therefore, considered visible. Second, the probabilities of detection 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.

Substituting (8) in (10) results in

$$\mathcal{P}_{\mathcal{F}_{(n_1,n_2)}} = 1 - \exp\left(-\left(\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}}\right)^{\beta}\right) \quad (11)$$

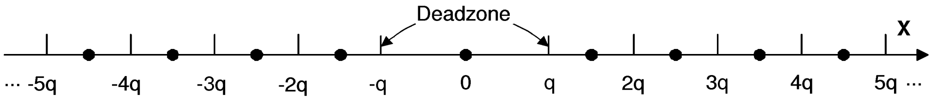


Fig. 2. Uniform scalar quantizer with deadzone twice the step size.

where

$$\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}} = \left(\sum_{(\lambda,\theta,i,j)\in F} \left| \frac{e(\lambda,\theta,i,j)}{t_{JND}(\lambda,\theta,i,j)} \right|^{\beta} \right)^{\frac{1}{\beta}}. \quad (12)$$

In (12), $\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}}$ takes the form of a Minkowski metric with exponent β. From (11), it can be seen that minimizing the probability of detecting a distortion in a foveal region is equivalent to minimizing the metric $\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}}$.

The distortion measure $\mathcal{D}$ for the whole image corresponds to the maximum probability of detecting a distortion over all possible foveal regions and is obtained by using a Minkowski metric with $\beta = \infty$ for inter-foveal pooling of the distortions $\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}}$

$$\mathcal{D} = \max_{(n_1,n_2)} \mathcal{D}_{\mathcal{F}_{(n_1,n_2)}}. \quad (13)$$

In (13), the maximum operation gives the worst case perceived error. Under this condition, the minimum bit rate for a given distortion $\mathcal{D}$ is achieved when all $\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}} = \mathcal{D}$.

III. Conventional JPEG2000 Encoding

In JPEG2000, an image is first divided into tiles. Each tile is then coded independently. For each tile, after DC level shifting and DWT, a scalar quantizer is first applied to each subband. The quantized samples in each subband are then divided into coding blocks. The usual code block size is 64 × 64 or 32 × 32. Each coding block is then independently bit-plane coded from the MSB to the LSB. Each bit-plane is fractionally coded using three coding passes: significant propagation, magnitude refinement and clean up (except the MSB, which is coded using the clean up pass). In the three passes, four coding operations, including significance coding, sign coding, magnitude refinement coding and cleanup coding, are used to code the sample's value in the current bitplane (0 or 1) and the sample's sign (positive or negative) when the sample becomes significant. Context-based binary arithmetic coding is extensively used. In this way, an embedded bitstream is generated for each code block. This is called the tier-1 coding. At the same time, the rate increase and the distortion reduction associated with each coding pass is recorded. This information is then used by the post compression rate-distortion (PCRD) optimization stage to determine each coding block's contribution to the different quality layers in the final bit-stream in order to reach the desired bit rates while optimizing the R-D performance.

Given the compressed bit-stream for each coding block and the rate allocation result, tier-2 coding is performed to form the final coded bit-stream. The basic unit in the final bit-stream is a packet. The formed packets can be of variable sizes. A packet consists of a packet header and a packet body. The packet header contains the following information: which coding blocks are included in this packet, number of bit planes for each newly included coding block (for which no encoded pass have been included in previous packets), number of coding passes included in this packet from each included coding block, and number of bytes contributed by each included coding block. The packet body contains the actually coded bits from each coding block.

This two-tier coding structure gives great flexibility to the final bit-stream formation. By determining how to assemble the sub-bitstreams from each coding block to form the final bitstream, different progression (SNR, resolution, position) order can be realized.

In the JPEG2000 post compression rate-distortion optimization process, the distortion metric can be the conventional MSE. In this case, the rate allocation attempts to minimize the peak signal to noise ratio (PSNR). One can also choose other perceptual distortion metrics that take various HVS properties into consideration [7]. Methods that truncate the bitstream based on distortion or rate-distortion slope values have been proposed for applications which require more consistent image quality [9].

From the above description, one can see that there are two layers of rate or distortion control in JPEG2000. First, the rate or distortion can be roughly controlled by the quantization step. This can be performed on a subband by subband basis. In Part 1 of the standard, only uniform scalar quantization with a deadzone twice the regular step size is included, as illustrated in Fig. 2. For the MSE distortion metric, the quantizer for each subband can be set to be inversely proportional to the amount of spatial-domain errors introduced by a coefficient unit error in the considered subband. This is also suggested in the informative part of the standard [30]. For perceptual criteria, one can choose the quantizer to be inversely proportional to the contrast sensitivity threshold of the considered subband, or be proportional to the visual detection thresholds [21].

Finer rate or distortion control is achieved by the selective inclusion of coding passes on a coding block basis. Given the rate-distortion information collected in the tier-1 coding for each coding pass of each coding block, the post compression rate-distortion optimization attempts to figure out the optimal coding pass inclusion strategy for different quality layers. First, a convex hull search is performed to find out the candidate (R, D) truncation points for each coding block. Each coding pass determines a point in the R-D plane. After the convex hull search, only those points on the convex hull are kept as candidate truncation points. The R-D slope values associated with each candidate truncation points are stored. Given the R-D curve for each coding block, a bisection search can be performed among all the coding blocks to find the R-D slope values that meet the desired bit rate. Those coding passes with steeper R-D slopes are included in the final bitsteam. Those coding passes that are not included in the final bitstream are ignored.

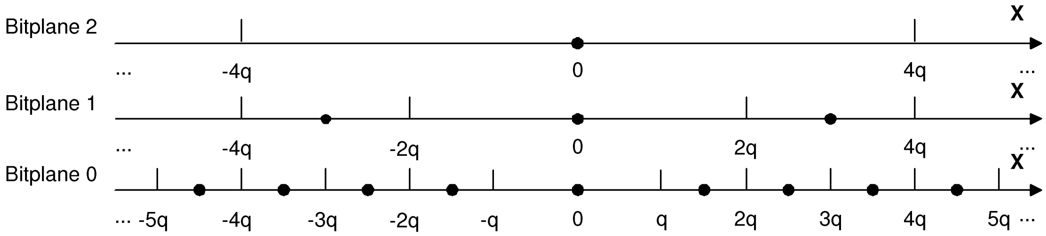


Fig. 3. Embedded uniform scalar quantizer with deadzone twice the step size.

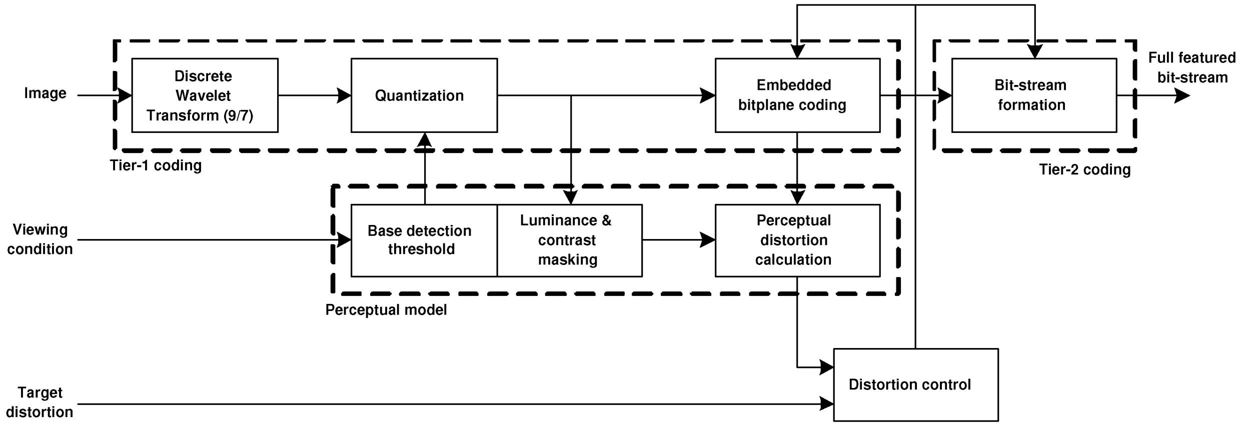


Fig. 4. Block diagram of the JPEG2000 encoding with perceptual distortion control.

From above, we can see that parts of the bitstream generated in the tier-1 coding are discarded. Since the embedded bitplane coding is the most complex part in the JPEG2000 compression system, computational power and memory is wasted on encoding and storing those finally discarded coding passes [5], [6]. Rate control methods [5], [6] and some JPEG2000 implementations such as Kakadu [4], [31] attempt to reduce this wasted complexity. It is desirable, however, to develop a rule to pre-determine the number of coding passes that should be encoded for each coding block before the embedded bitplane coding starts or make this decision in parallel with the embedded bitplane coding. However, under the current post compression rate-distortion optimization framework, it is very difficult to make this decision with sufficient accuracy to avoid discarding the encoded data.

IV. JPEG2000 Encoding With Perceptual Distortion Control

One question we should ask is whether achieving a certain bit rate is so important for a compression algorithm design. For certain applications such as transmission over channels with limited bandwidth, one definitely needs to reach a precise rate control. But for a lot of other applications, the quality and, especially, the perceived quality matters. In this work, we take a different perspective to this rate-distortion optimization problem. The goal is to reach the lowest bit rate for a desired perceptual quality. Depending on the perceptual model and complexity, several encoding approaches can be proposed.

If only the base detection threshold for each subband is considered, approximate distortion control can be easily implemented in one of two ways. Note that the embedded bitplane coding of JPEG2000 is essentially a family of uniform scalar quantizers with dead-zones twice the regular step size as illustrated in Fig. 3. At bitplane level n, the effective quantization step size is $2^n q$. For the midpoint reconstruction, the quantization error at the end of bitplane n is less than $2^n q$ for insignificant coefficients and $2^{n-1} q$ for significant coefficients, respectively. Thus, with the quantization step q set to 1 in the quantization stage, one just needs to code each coding block in a subband down to the following bitplane level

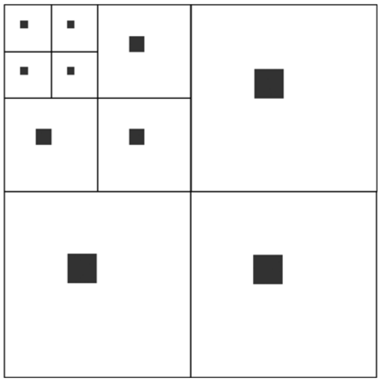


Fig. 5. DWT coefficients affecting the foveal region reconstruction.

$$B_{\lambda,\theta}(r) = \lfloor \log_2 JND_{\lambda,\theta}(r) \rfloor \tag{14}$$

for perceptual lossless coding. This is the threshold point at which the distortion will become just perceptible. This perceptual lossless point constitutes a natural target for high quality image coding, which is important in applications where loss in quality can not be tolerated, such as medical imaging. This perceptual lossless point can also be used as a benchmark for performance comparison with other coders. For other perceptual

Fig. 6. Images used in performance comparison. From left to right: 512 × 512 Lena, 720 × 576 Goldhill, 2048 × 2560 Woman, and Bike images (size of all images reduced to fit within the space).

qualities, one can increase the $B_{\lambda,\theta}(r)$ by an appropriate positive integer number.

The same result can also be achieved by using a quantizer for each subband that will result in a quantization error that is just at the threshold of visibility and then coding all the bitplanes after quantization. Therefore, with r denoting the display resolution, for perceptual-lossless compression, one just needs to set the quantizer step size on a subband-by-subband basis to

$$Q_{\lambda,\theta}(r) = JND_{\lambda,\theta}(r) \tag{15}$$

which will result in a quantization error that is just at the threshold of visibility. Other perceptual quality can be reached by scaling up the $Q_{\lambda,\theta}(r)$ for each subband by a common factor, resulting in supra-threshold coding at multiples of the JND level. After quantization, all bitplanes are coded in tier-1 coding and included in the final bitstream. In the JPEG2000 bitstream syntax, there is the QCD marker segment where one can specify the quantizer stepsize for each subband.

If the luminance masking and contrast masking are also considered, more precise distortion control can be performed. Fig. 4 shows a block diagram of the proposed JPEG2000 encoding with precise distortion control. After the discrete wavelet transform, the transformed wavelet coefficients are first quantized. The quantizer for subband (λ, θ) is set to the base detection threshold

$$Q_{\lambda,\theta}(r) = JND_{\lambda,\theta}(r). \tag{16}$$

The luminance masking and contrast masking adjustments are then calculated using the quantized wavelet coefficient values. So, the luminance masking adjustment (4) can be calculated as

$$a_l(\lambda,\theta,i,j) = \left(\frac{\hat{v}_{\lambda_{\max},LL,i',j'}}{\hat{v}_{\text{mean}}}\right)^{a_T} \tag{17}$$

where $\hat{v}_{\lambda_{\max},LL,i',j'} = v_{\lambda_{\max},LL,i',j'}/Q_{\lambda_{\max},LL}$, and $\hat{v}_{\text{mean}} = v_{\text{mean}}/Q_{\lambda_{\max},LL}$ with $Q_{\lambda_{\max},LL} = JND_{\lambda_{\max},LL}$.

The self contrast masking (6) and neighborhood contrast masking (7) can be expressed as

$$\begin{aligned} &a_{c\_\text{self}}(\lambda,\theta,i,j) \\ &\quad = \max\left\{1, \left(\frac{|\hat{v}(\lambda,\theta,i,j)|}{a_l(\lambda,\theta,i,j)}\right)^{\epsilon}\right\} \qquad (18) \\ &a_{c\_\text{neig}}(\lambda,\theta,i,j) \\ &\quad = \max\left\{1, \sum_{k\in\text{neighbors of }(\lambda,\theta,i,j)} \frac{\left|\frac{\hat{v}_k}{a_l(\lambda,\theta,i,j)}\right|^{\zeta}}{N_{i,j}}\right\} \qquad (19) \end{aligned}$$

where $\hat{v}(\lambda,\theta,i,j) = v(\lambda,\theta,i,j)/Q_{\lambda,\theta}$ with $Q_{\lambda,\theta} = JND_{\lambda,\theta}$.

The distortion expressed in (12) becomes

$$\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}} = \left(\sum_{(\lambda,\theta,i,j)\in F} \left|\frac{\hat{e}(\lambda,\theta,i,j)}{a_l(\lambda,\theta,i,j)a_c(\lambda,\theta,i,j)}\right|^{\beta}\right)^{\frac{1}{\beta}} \tag{20}$$

where $\hat{e}(\lambda,\theta,i,j)$ is the error caused by the bitplane coding of the quantized wavelet coefficients and by bitplane truncation and can be shown to be equal to $e(\lambda,\theta,i,j)/JND_{\lambda,\theta}$.

For a given viewing condition, the luminance and contrast masking adjustments can be calculated on a coefficient-by-coefficient basis using (17)–(19). Since the masking thresholds are just needed at the encoder side for distortion control, the actual quantized DWT coefficients can be used to derive the masking threshold profile at the encoder since these are not needed at the decoder.

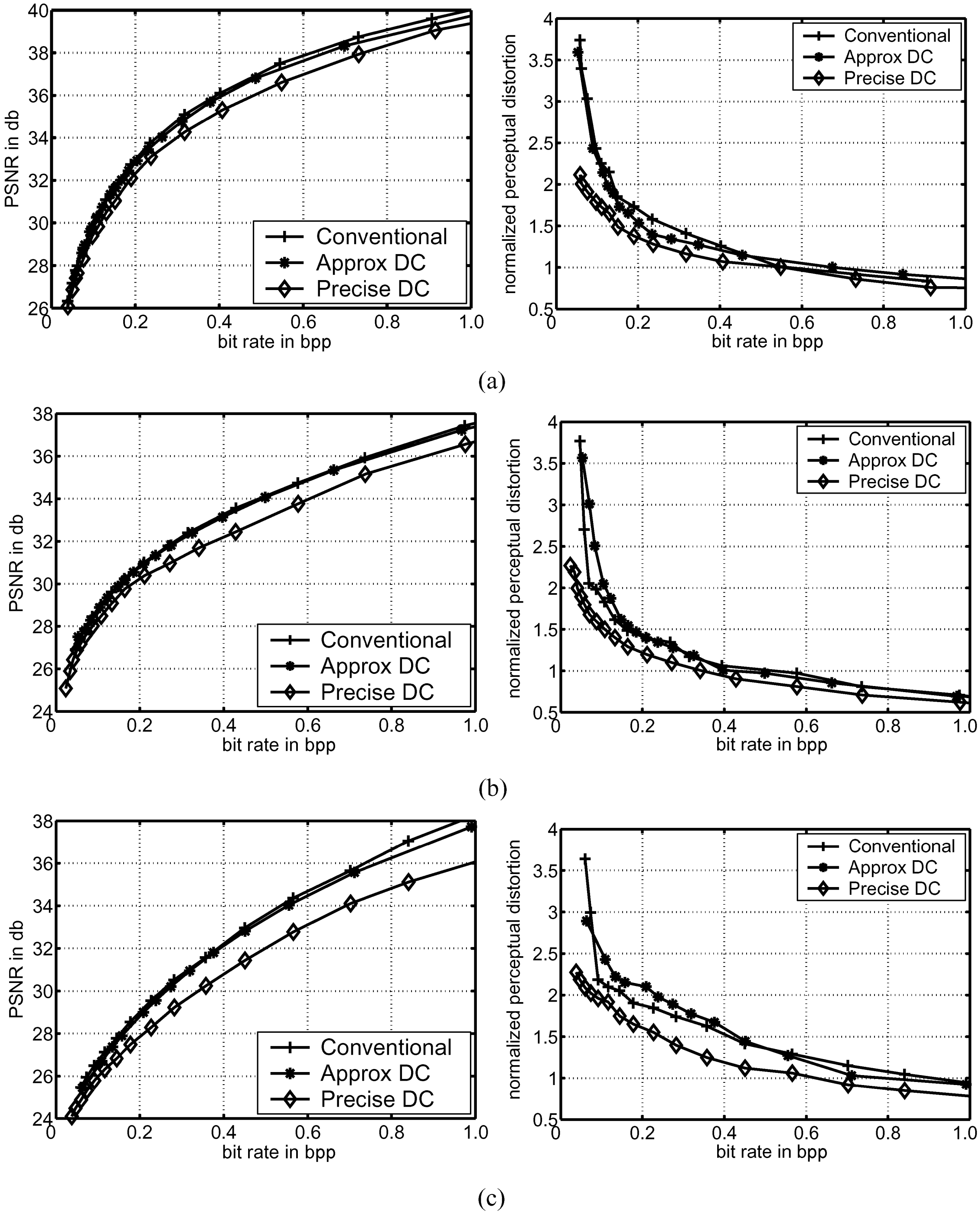
The size of the foveal region in the spatial domain can be calculated using (9). In addition, as defined in Section II-B, F represents the set of all DWT coefficients affecting the foveal region reconstruction. Let $F(\lambda,\theta)$ represent the DWT coefficients in subband (λ,θ) that affect the foveal region reconstruction. Then, $F = \bigcup_{(\lambda,\theta)} F(\lambda,\theta)$. $F(\lambda,\theta)$ can be determined in two ways. If we consider the filter length and boundary extension, $F(\lambda,\theta)$ can be determined as done for the mask generation in the Region of Interest (ROI) coding mode of JPEG2000 [30]. For simplicity, $F(\lambda,\theta)$ can also be approximated as a horizontal and vertical down-sampling by a factor of 2 of the foveal region with each level of DWT decomposition, as shown in Fig. 5.

Let $N(F)$ and $N(F(\lambda,\theta))$ denote the number of coefficients in F and $F(\lambda,\theta)$, respectively. Note that $N(F) = \sum_{\lambda}\sum_{\theta} N(F(\lambda,\theta))$. Define $\mathcal{D}_{F(\lambda,\theta)}$ to be

$$\mathcal{D}_{F(\lambda,\theta)} = \left(\sum_{(\lambda,\theta,i,j)\in F(\lambda,\theta)} \left|\frac{\hat{e}(\lambda,\theta,i,j)}{a_l(\lambda,\theta,i,j)a_c(\lambda,\theta,i,j)}\right|^{\beta}\right)^{\frac{1}{\beta}}. \tag{21}$$

Then $\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}}$ in (20) can be expressed as

$$\mathcal{D}_{\mathcal{F}_{(n_1,n_2)}} = \left(\sum_{\lambda}\sum_{\theta} \mathcal{D}_{F(\lambda,\theta)}{}^{\beta}\right)^{\frac{1}{\beta}}. \tag{22}$$



(a)

(b)

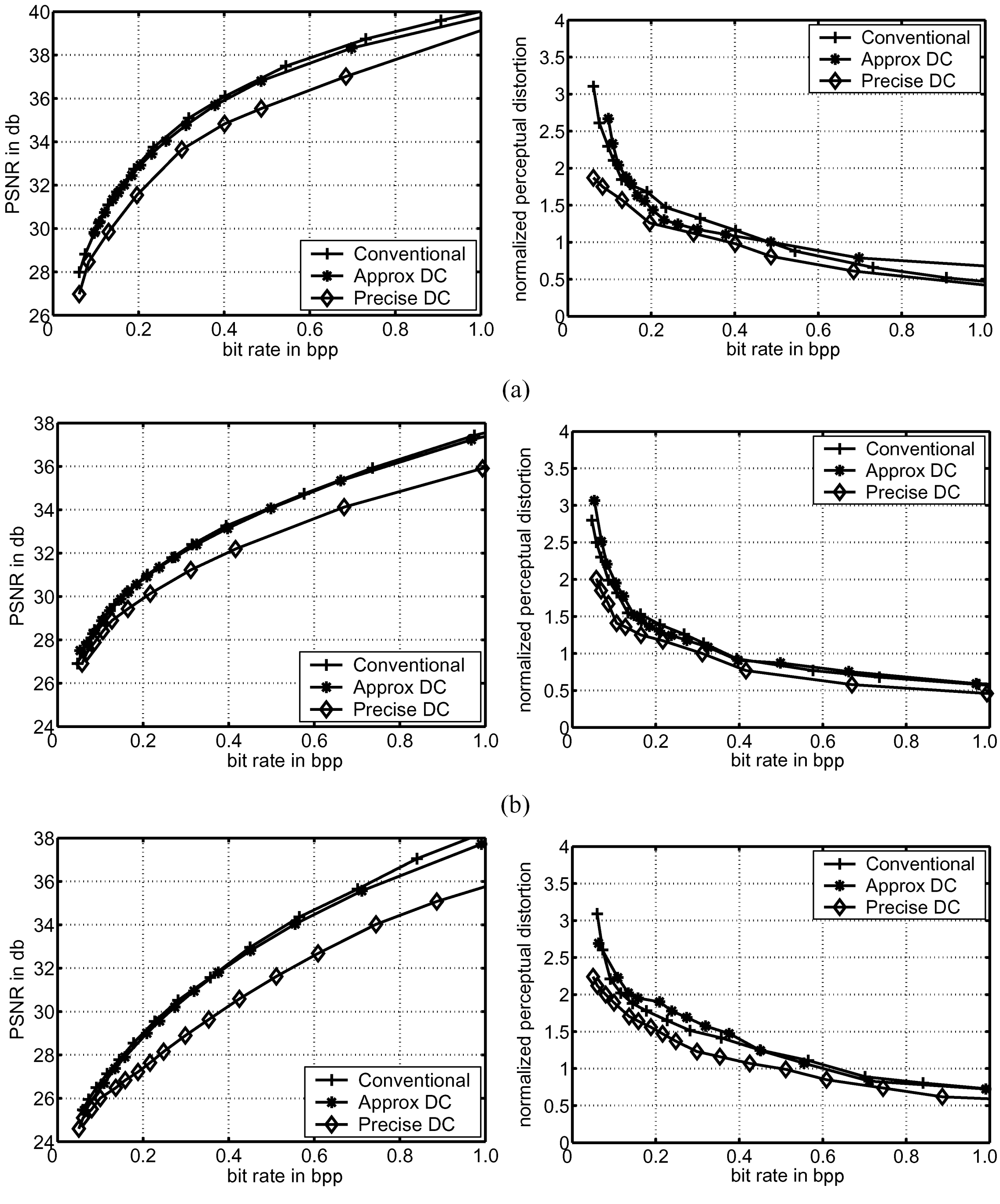
(c)

Fig. 7. Comparison of conventional JPEG2000 coding (Conventional), approximate distortion control (Approx DC) and precise distortion control (Precise DC) without neighborhood masking in terms of PSNR and perceptual distortion $\mathcal{D}_{\mathrm{nor}}$. (a) 512 × 512 Lena; (b) 720 × 576 Goldhill; (c) 2048 × 2560 Woman.

The perceptual distortion $\mathcal{D}$ of (13) can be made to be equal to a desired target distortion $\mathcal{D}_T$ by controlling the coding of $F(\lambda, \theta)$ such that

$$\mathcal{D}_{F(\lambda,\theta)} \leq \left(\frac{N\left(F(\lambda,\theta)\right)}{N(F)} \mathcal{D}_T^{\beta} \right)^{\frac{1}{\beta}}. \quad (23)$$

It is possible for $F(\lambda, \theta)$ to be smaller than a coding block. In this case, there are several sets of detectors, in the coding block, each covering $N(F(\lambda, \theta))$ wavelet coefficients. Before coding starts on a coding block, the initial perceptual distortions $\mathcal{D}_{F(\lambda,\theta)}$ are computed using (21) for the considered sets of detectors. Since, at the start, nothing is coded yet, the error $\hat{e}(\lambda, \theta, i, j)$ in (21) is equal to the initial coefficient value $\hat{v}(\lambda, \theta, i, j)$ before bit-plane coding. During the embedded bitplane coding, whenever a coefficient is identified as significant or refinement coded, the perceptual distortion $\mathcal{D}_{F(\lambda,\theta)}$ is updated. At the end of each coding pass, if the distortion $\mathcal{D}_{F(\lambda,\theta)}$ of all the sets of detectors, in the considered coding block, are below the desired distortion as in (23),



(a)

(b)

(c)

Fig. 8. Comparison of conventional JPEG2000 coding (Conventional), approximate distortion control (Approx DC) and precise distortion control (Precise DC) with neighborhood masking in terms of PSNR and perceptual distortion $\mathcal{D}_{\text{nor}}$. (a) 512 × 512 Lena; (b) 720 × 576 Goldhill; and (c) 2048 × 2560 Woman.

coding of that coding block can be stopped and coding of the next coding block can be started. In this way, one can ensure that the distortion of any foveal region is below the target distortion $\mathcal{D}_T$. On the other hand, if the size of $F(\lambda, \theta)$ is bigger than the coding block, there is only one set of detectors that covers the whole coding block. In this latter case, one needs to control the coding of the coding block such that $\mathcal{D}_{\text{coding block}} \leq \mathcal{D}_{F(\lambda,\theta)}(N(\text{coding block})/N(F(\lambda,\theta)))^{1/\beta}$, where $\mathcal{D}_{\text{coding block}}$ is the distortion in the considered coding block, and $N(\text{coding block})$ denotes the number of wavelet coefficients in that coding block. Note that $\mathcal{D}_{\text{coding block}}$ is computed as in (21) but with $F(\lambda, \theta)$ replaced with the considered coding block region.

In contrast to the conventional JPEG2000 encoding, the proposed distortion controlled approach stops the coding, for each coding block, when the target distortion is reached; and therefore, there is no need to code all the bitplanes in tier-1 coding and no need for post compression rate-distortion optimization.

The final compressed bitstream of JPEG2000 is constructed by packing the independently coded bitstreams of each coding

TABLE III
BIT RATES FOR THE CONVENTIONAL JPEG2000 ENCODING AND THE PROPOSED JPEG2000 CODING WITH PRECISE PERCEPTUAL DISTORTION CONTROL FOR $\mathcal{D}_{\text{nor}} = 1.0$ CORRESPONDING TO HIGH-QUALITY OR ALMOST TRANSPARENT CODING

	Without Neighborhood Masking			With Neighborhood Masking		
Image	Conv	Prec	Gain	Conv	Prec	Gain
Lena	0.5	0.428	14.4%	0.48	0.4	16.6%
Goldhill	0.394	0.341	13.4%	0.360	0.305	15.2%
Woman	0.712	0.586	17.7%	0.628	0.510	18.7%
Bike	0.883	0.708	19.8%	0.785	0.618	21.2%

block together in some order. By using the perceptual distortion as the metric of ordering, a perceptual distortion progressive bitstream can also be generated, if desired, where the visually important information is included first [32]. In this bitstream, the first layer, L_1, contains contributions from coding blocks that make the perceptual distortion of each foveal region below $\mathcal{D}_1$, subsequent layers, L_i, contains additional contributions from coding blocks that make the perceptual distortion of each foveal region below $\mathcal{D}_i$. Because the generated bitstream is fully compatible with the JPGE2000 Part 1 standard, any JPEG2000 compliant decoder can be used to get the reconstructed image.

As indicated in Section I, visual progressive weighting (VIP) and visual masking [4, Ch. 16] have been provided as an option in some JPEG2000 implementations. The Cvis technique, which is described in [7], attempts to model masking aspects of the HVS and has been incorporated as part of the Verification Model (VM) implementation of the JPEG2000 standard [1]. In addition to introducing a different *locally adaptive* masking model for perceptual thresholds and perceptual distortion computation, one main difference between the proposed method and Cvis is that the proposed method truncates the code-block based on perceptual distortion thresholds evaluated at each point in the code-block. On the other hand, Cvis is a PCRD-based truncation scheme which determines the code-block truncation point based on an average distortion over the code-block. In the proposed work, the distortions are computed adaptively based on local perceptual thresholds by summing over foveal regions (which can be smaller than the coding blocks). The coding stops when the distortions over the foveal regions (and not the coding block) are less than the target perceptual distortion. Coding results and comparison with Cvis are presented in Section V.

V. CODING RESULTS

The JASPER [2] JPEG2000 implementation is modified to incorporate the proposed perceptual distortion control. For approximate distortion control, the quantizer step size of a subband is set to the base detection threshold of that subband as discussed in Section IV. The encoding was optimized for a viewing distance of 60 cm and a 31.5 pixels-per-cm (80 pixels-per-inch) display, which results in a display visual resolution of 32.9 pixels/degree. The images used for the performance comparisons are shown in Fig. 6. They are the 512 × 512 Lena, 720 × 576 Goldhill, 2048 × 2560 Woman and 2048 × 2560 Bike images.

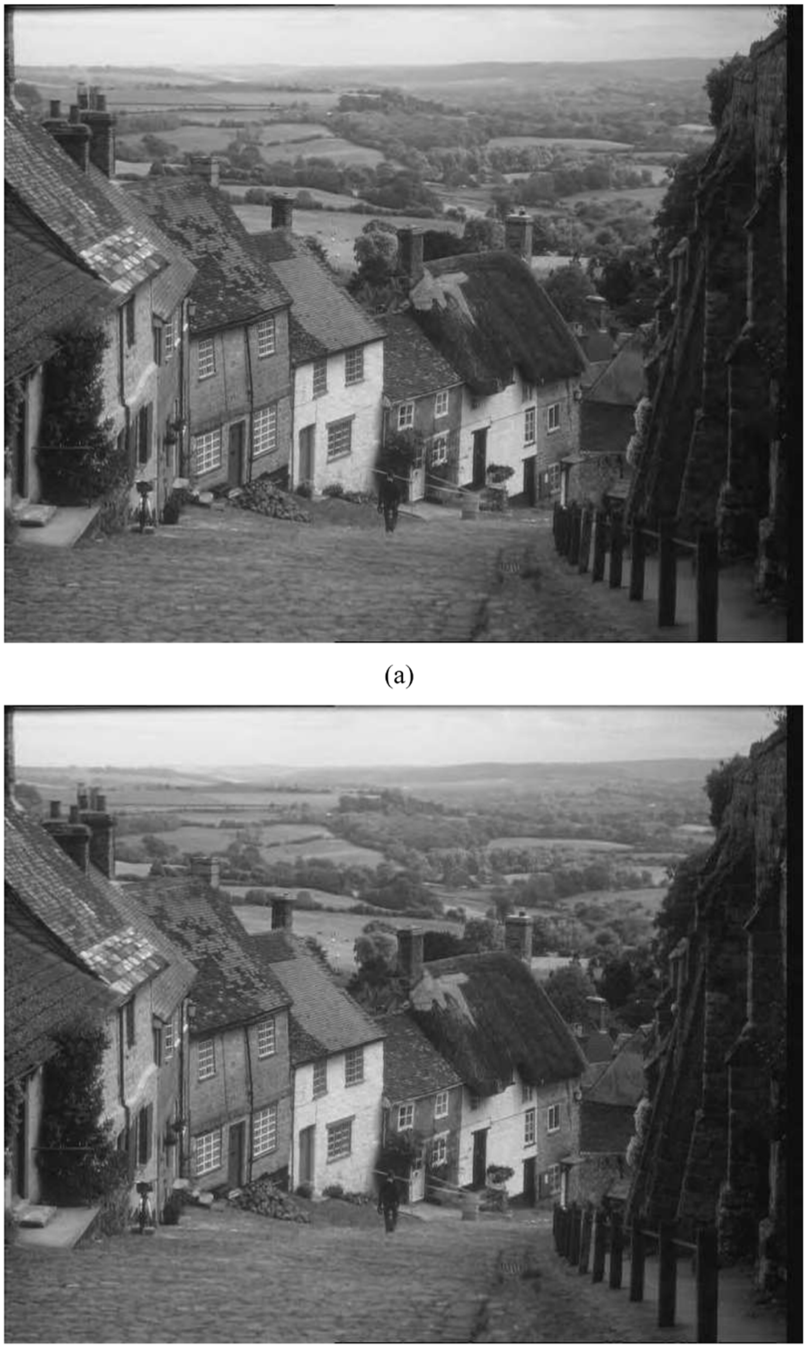
Figs. 7 and 8 compare the performance of the proposed precise perceptual distortion control, and approximate distortion control encoding methods with the conventional rate-based JPEG2000 encoding in terms of PSNR and the normalized perceptual distortion measure defined as

$$\mathcal{D}_{\text{nor}} = \frac{\mathcal{D}}{(N(F))^{\frac{1}{\beta}}} \tag{24}$$

where F is the area over which errors are pooled as given by (12). Since the bitstream generated by the perceptual distortion control is fully compatible with the JPEG2000 bitstream syntax, the JASPER JPEG2000 decoder is used in all cases. From Fig. 7, one can see that the PSNR value obtained by the approximate distortion control is very close to that obtained by the conventional JPEG2000 coding with reduced computational complexity. Since the proposed JPEG2000 encoding with precise distortion control is optimized to control the encoding to the local JND profile, and not to minimize the MSE, the MSE-based PSNR does not well indicate the achieved performance for the precise distortion control. However, if the images are compared in terms of the perceptual distortion measure, the proposed JPEG2000 encoding with precise distortion control clearly outperforms the conventional JPEG2000 coder and the proposed JPEG2000 encoding with approximate distortion control. Note that approximate distortion control exhibits a slightly decreased performance as compared to the rate-based JPEG2000 but at a much lower complexity.

Table III compares the bit rates (in bits per pixel) obtained using the proposed JPEG2000 encoding with precise perceptual distortion control with those obtained using the conventional rate-based JPEG2000 coder for almost transparent compression with $\mathcal{D}_{\text{nor}} = 1.0$. The compression gain is computed as (25), shown at the bottom of the page.

$$\text{compression gain} = \frac{\text{Conventional bit rate} - \text{Precise distortion control bit rate}}{\text{Conventional bit rate}} \tag{25}$$



(a)

(b)

Fig. 9. Coding results for the 720 × 576 Goldhill image with the proposed JPEG2000 coding with precise perceptual distortion control for a target $\mathcal{D}_{\rm nor} = 1.0$, and comparison with conventional JPEG2000 coding. (a) Conventional JPEG2000, $\mathcal{D}_{\rm nor} = 1.06$, rate $= 0.394$ bpp. (b) Proposed JPEG2000 coding with precise perceptual distortion control, $\mathcal{D}_{\rm nor} = 1.0$, rate $= 0.341$ bpp.

Figs. 7, 8, and Table III also show the results obtained using the proposed perceptual JPEG2000 encoding with precise distortion control with and without neighborhood masking. One can see that the neighborhood masking does not significantly change the performance of the proposed JPEG2000 encoding with precise distortion control and only self contrast masking. These results confirm the claim in [22] that the contrast masking is strongest when both components are of the same spatial frequency, orientation, and location. Due to the high computation complexity of the neighborhood masking, a contrast masking model based solely on self contrast masking is adopted and the remaining reported results are obtained without neighborhood masking.

Fig. 9 compares the reconstructed images for high-quality or almost transparent encoding for the 720 × 576 Goldhill image. For the proposed JPEG2000 encoding with precise distortion control, images are obtained with target distortion $\mathcal{D}_{\rm nor} = 1.0$. This results in a bit rate of 0.341 bits per pixel (bpp) and actual distortion $\mathcal{D}_{\rm nor} = 1.0$. The conventional JPEG2000 encoding result is obtained by trying to match the distortion obtained with

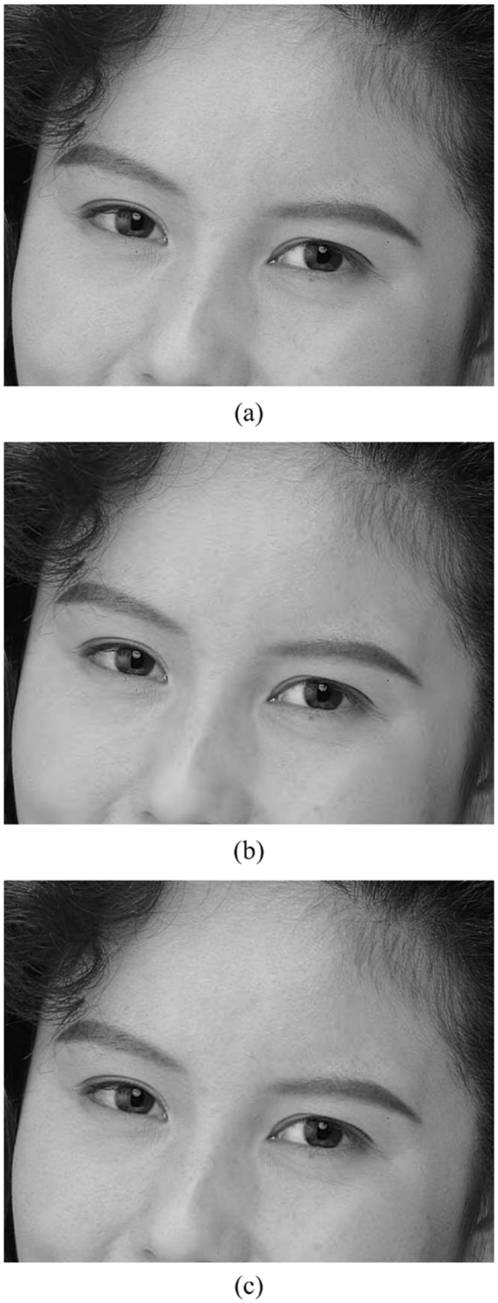


Fig. 10. Coding results for the 2048 × 2560 Woman image with the proposed JPEG2000 coding with precise perceptual distortion control for a target $\mathcal{D}_{\rm nor} = 1.0$, and comparison with conventional JPEG2000 coding. (a) Portion of original Woman image. (b) Conventional JPEG2000, $\mathcal{D}_{\rm nor} = 1.29$, rate $= 0.586$ bpp. (c) Proposed JPEG2000 coding with precise perceptual distortion control, $\mathcal{D}_{\rm nor} = 1.06$, rate $= 0.586$ bpp.

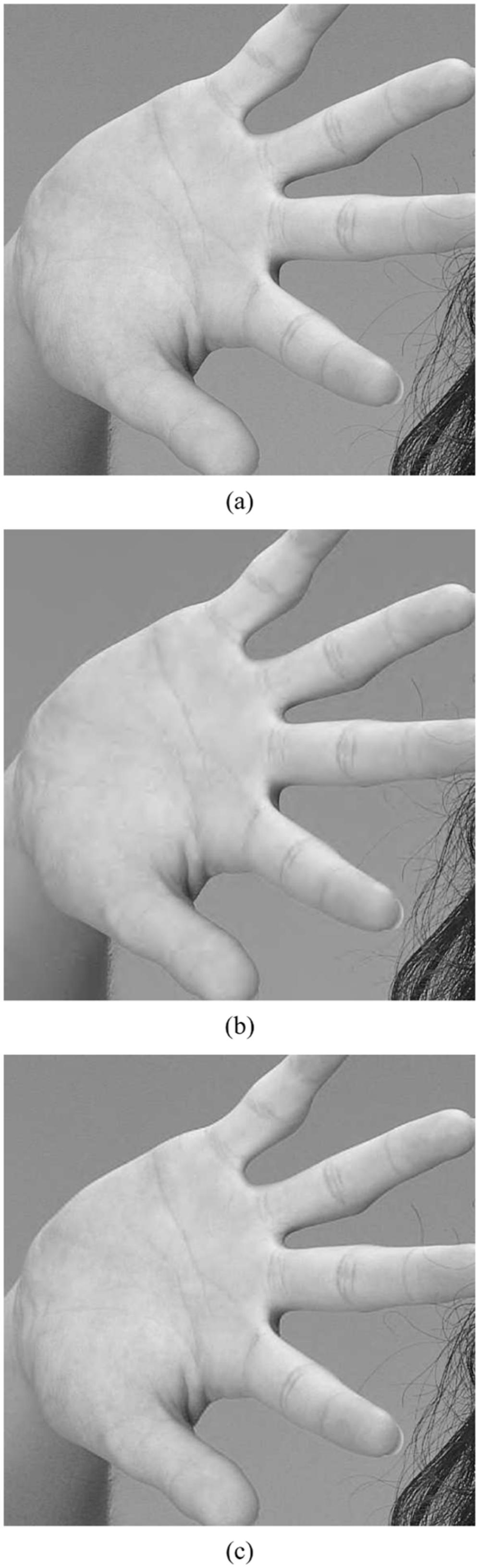


Fig. 11. Coding results for the 2048 × 2560 Woman image. (a) Portion of original Woman image. (b) Conventional JPEG2000, $\mathcal{D}_{\rm nor} = 1.29$, rate $= 0.586$ bpp. (c) Proposed JPEG2000 coding with precise perceptual distortion control, $\mathcal{D}_{\rm nor} = 1.06$, rate $= 0.586$ bpp.

the proposed JPEG2000 encoding with precise distortion control. This results in a bit rate of 0.394 bpp and an actual distortion $\mathcal{D}_{\rm nor} = 1.06$.

Figs. 10–12 compare the reconstructed 2048 × 2560 Woman and Bike images obtained using the conventional JPEG2000 coding with those obtained using the proposed JPEG2000 coding with precise perceptual distortion control at the same bit rate. For the proposed JPEG2000 coding with precise distortion control, the images are coded with the target distortion $\mathcal{D}_{\rm nor}$ set to 1.0. The conventional JPEG2000 images are obtained by matching the bit rate obtained using the proposed JPEG2000 coding with perceptual distortion control.

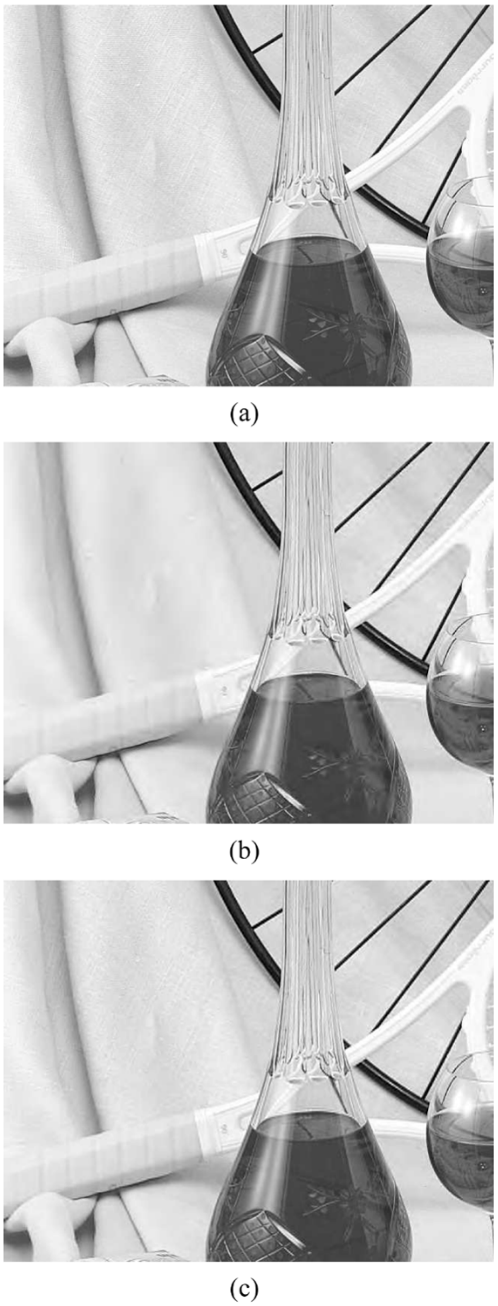


Fig. 12. Coding results for the 2048 × 2560 Bike image with the proposed JPEG2000 coding with precise perceptual distortion control for a target $\mathcal{D}_{\rm nor} = 1.0$, and comparison with conventional JPEG2000 coding. (a) Portion of original Bike image. (b) Conventional JPEG2000, $\mathcal{D}_{\rm nor} = 1.33$, rate $= 0.708$ bpp. (c) Proposed JPEG2000 coding with precise perceptual distortion control, $\mathcal{D}_{\rm nor} = 1.10$, rate $= 0.708$ bpp.

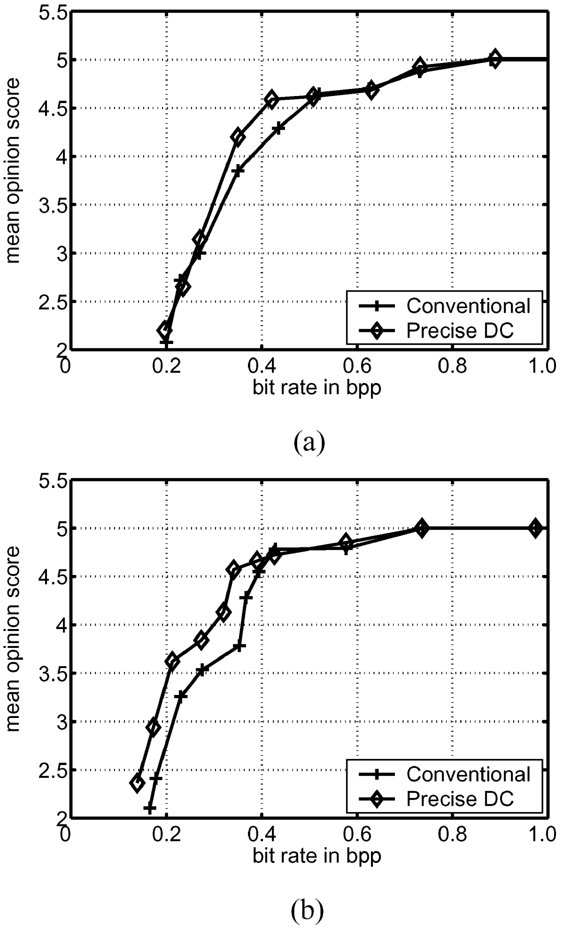
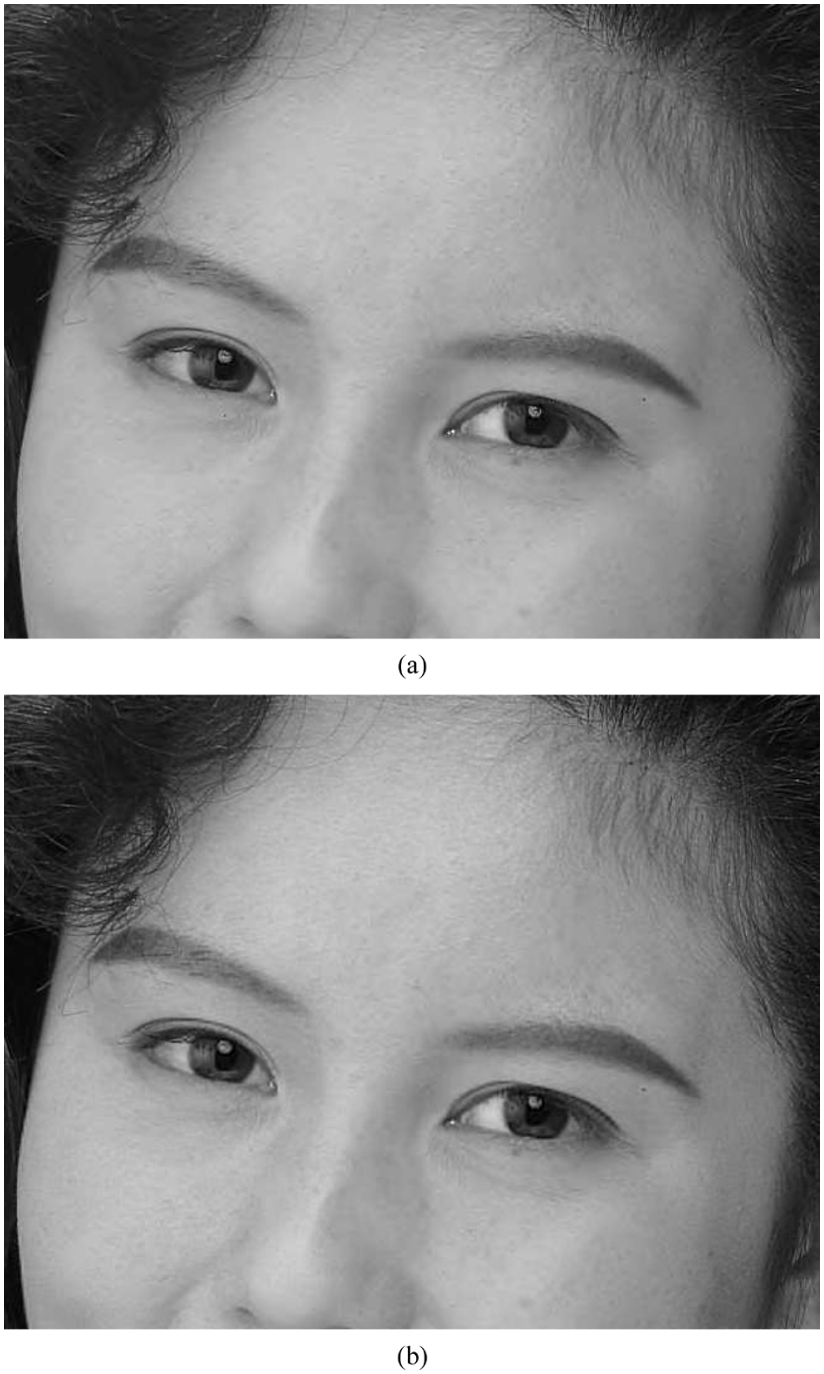


Fig. 13. Mean opinion score as a function of bit rate for the 512 × 512 Lena and 720 × 576 Goldhill image coded using conventional JPEG2000 encoding and JPEG2000 encoding with precise perceptual distortion control. (a) MOS for 512 × 512 Lena image; (b) MOS for 720 × 576 Goldhill image.

Portions of the Woman image are shown in Figs. 10 and 11. Clear differences can be seen when the resulting images are viewed on a CRT monitor. The proposed JPEG2000 coding with precise perceptual distortion control preserves the low amplitude texture much better than the conventional one. In the conventional JPEG2000, the skin texture is almost erased, but is relatively well preserved in the proposed JPEG2000 coding with precise perceptual distortion control, especially in the forehead and cheek areas (Fig. 10), and in the hand and hair areas (Fig. 11). Similar observations can be made from Fig. 12, which shows coding results for the Bike image. When these images are compared on a screen, it can be observed that the conventional JPEG2000 encoding totally wipes out the texture on the background table cloth, the tennis handle and the top surface of the liquid, and that it blurs the boundary of the bike wheel. In comparison, the JPEG2000 coding with precise perceptual distortion control preserves this texture information much better and presents the high frequency contours of the bike wheel very crisply.

A set of subjective impairment tests have been conducted to compare the coding performance in terms of perceived quality for the proposed JPEG2000 encoding with precise perceptual distortion control and the conventional JPEG2000 encoding. The Double Stimulus Impairment Scale (DSIS) method [33] is adopted. In the test, a sequence of image pairs consisting of the original image and reconstructed image at various bit rates are displayed on a computer monitor with a display resolution of 80 pixels per inch. The subjects were asked to rate the amount of impairment in the decoded images using a five-level scale [33]. The categories on the scale and their numerical values are "imperceptible" (5), "perceptible, not annoying" (4), "slightly annoying" (3), "annoying" (2), and "very annoying" (1). The suggested viewing distance was 60 cm (23.6 inches). The group of viewers consisted of 15 individuals with normal or corrected



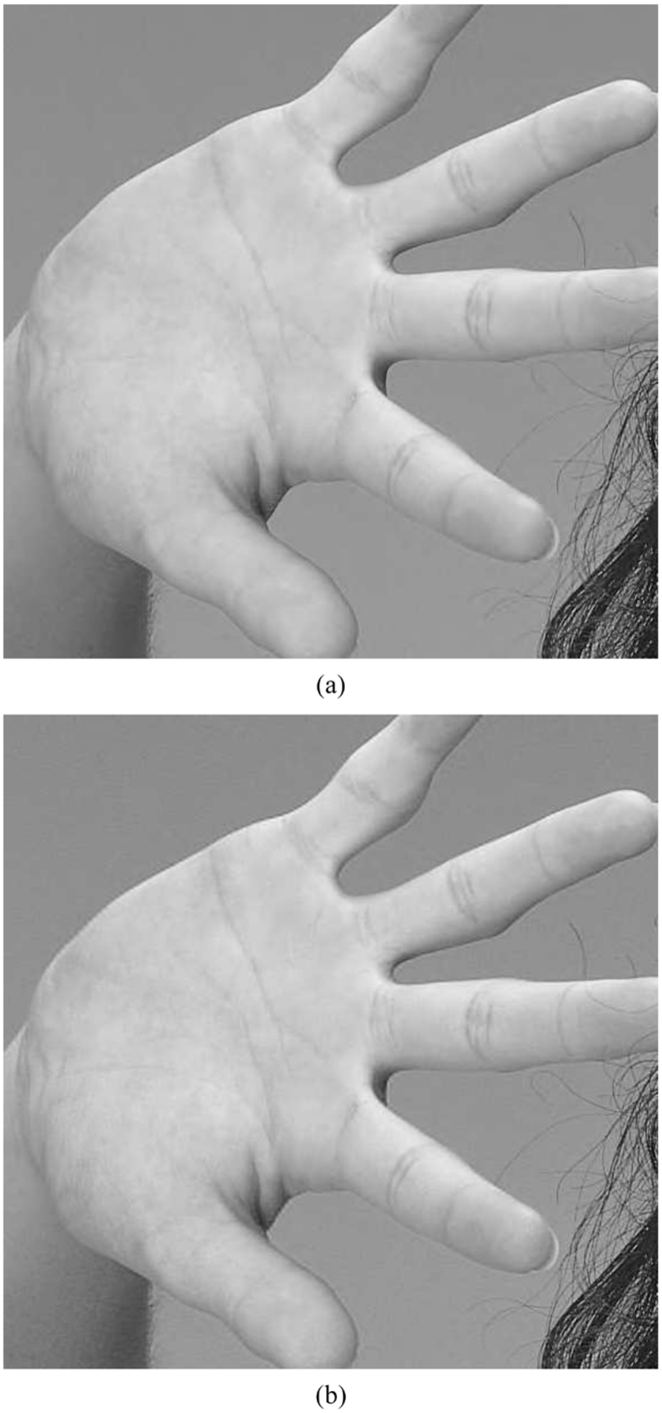
(a)

(b)

Fig. 14. Coding results for the 2048 × 2560 Woman image with the proposed JPEG2000 coding with precise perceptual distortion control for a target $\mathcal{D}_{\rm nor} = 1.0$, and comparison with existing JPEG2000 coding with visual masking (VM with Cvis on) at the same resulting bit-rate; the face portion of the coded woman image is shown. (a) Existing JPEG2000 with visual masking (VM with Cvis), rate $= 0.586$ bpp. (b) Proposed JPEG2000 coding with precise perceptual distortion control, rate $= 0.586$ bpp.

to normal vision. The mean opinion score (MOS) results for the impairment test for the 512 × 512 Lena and 720 × 576 Goldhill images are shown as a function of bit rate in Fig. 13. They indicate that the proposed JPEG2000 encoding with precise distortion control results in an improved perceived image quality as compared to the conventional JPEG2000 encoding method.

In order to compare the performance of the proposed perceptual JPEG2000-based coding scheme with existing visual weighting schemes that are compatible with Part I of JPEG2000, the 2048 × 2560 Woman image was also coded, at the rate of 0.586 bpp, using the VM8.0 JPEG2000 implementation with the Cvis option [7] (run with an exponent of 0.5 as recommended [4, Ch. 16, Sec. 16.1.4]). The Cvis option incorporates visual masking into the JPEG2000 encoding. Portions of the resulting coded Woman image are shown in Figs. 14 and 15. The corresponding original images are shown in Figs. (10a) and (11a), respectively. As before, the proposed perceptual JPEG2000 coding preserves better the skin texture in the forehead, cheek, hand, wrist, and hair areas. In addition, the edges and lines of the hand and hair are better preserved by


(a)

(b)

Fig. 15. Coding results for the 2048 × 2560 Woman image with the proposed JPEG2000 coding with precise perceptual distortion control for a target $\mathcal{D}_{\rm nor} = 1.0$, and comparison with existing JPEG2000 coding with visual masking (VM with Cvis on) at the same resulting bit-rate; the hand portion of the coded woman image is shown. (a) Existing JPEG2000 with visual masking (VM with Cvis), rate $= 0.586$ bpp. (b) Proposed JPEG2000 coding with precise perceptual distortion control, rate $= 0.586$ bpp.

the proposed scheme as compared to the existing JPEG2000 encoding with visual masking (VM with Cvis).

VI. Conclusion

A distortion-controlled JPEG2000 encoding method is presented with the goal to achieve consistent reconstructed image quality while reaching the lowest possible bit rate for the desired perceptual distortion. For this purpose, a vision model that takes into account various masking effects of human visual perception and a perceptual distortion metric that takes spatial and spectral summation of individual quantization errors are incorporated into the JPEG2000 coder design. The proposed new encoding method is compatible with Part I of the JPEG2000 standard and the generated bitstream can be decoded by any JPEG2000 decoder. Compared with the conventional rate-based JPEG2000 encoding, the proposed method provides a way to generate consistent quality images at a lower bit rate.

References

[1] *JPEG2000 Verification Model 7.0 Software*, iSO/IEC/JTC1/SC29/WG1 N1685.

[2] *JASPER Software Reference Manual*, iSO/IEC/JTC1/SC29/WG1 N2415.

[3] *JJ2000 V. 4.2*, iSO/IEC/JTC1/SC29/WG1 N2136.

[4] D. S. Taubman and M. W. Marcellin, *Image Compression Fundamentals, Standards and Practice: JPEG2000*. Norwell, MA: Kluwer, 2002.

[5] T. Masuzaki, "Adaptive rate control for JPEG2000 image coding in embedded systems," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, 2002, pp. 77–80.

[6] T. H. Chang, "Computation reduction technique for lossy JPEG2000 encoding through EBCOT tier-2 feedback processing," in *Proc. IEEE Int. Conf. Image Processing*, vol. 3, 2002, pp. 85–88.

[7] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Process.*, vol. 9, no. 7, pp. 1158–1170, Jul. 2000.

[8] W. Zeng, S. Daly, and S. Lei, "An overview of the visual optimization tools in JPEG2000," *Signal Process.: Image Commun.*, pp. 85–104, 2002.

[9] Kakadu User's Group Archives. [Online]. Available: http://www.kakadusoftware.com

[10] L. Karam, "An analysis/synthesis model for the human visual based on subspace decomposition and multirate filter bank theory," in *Proc. IEEE Int. Symp. Time-Frequency Time-Scale Analysis*, 1992, pp. 559–562.

[11] J. G. Daugman, "Two-dimensional spectral analysis of the cortical receptive field profiles," *Vis. Res.*, vol. 20, no. 10, pp. 847–856, 1980.

[12] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422, Oct. 1993.

[13] A. B. Watson and J. A. Solomon, "A model of visual contrast gain control and pattern masking," *J. Opt. Soc. Amer.*, vol. 14, pp. 2397–2391, 1997.

[14] D. J. Swift and R. A. Smith, "Spatial frequency masking and Weber's law," *Vis. Res.*, vol. 23, pp. 495–506, 1983.

[15] J. Ross and H. D. Speed, "Contrast adaptation and contrast masking in human vision," in *Proc. Roy. Soc. Lond. B*, 1991, pp. 61–69.

[16] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *J. Opt. Soc. Amer. A*, vol. 70, no. 12, pp. 1458–1471, 1980.

[17] D. G. Pelli, "Effects of Visual Noise," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 1981.

[18] R. J. Safranek, "A comparison of the coding efficiency of perceptual models," *Proc. SPIE*, vol. 2411, pp. 83–91, 1995.

[19] T. Pappas, T. Michel, and R. Hinds, "Supra-threshold perceptual image coding," in *IEEE Int. Conf. Image Processing*, 1996, pp. 237–240.

[20] F. Kingdom and P. Whittle, "Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing," *Vis. Res.*, vol. 36, no. 6, pp. 817–829, 1996.

[21] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Process.*, vol. 6, no. 8, pp. 1164–1175, Aug. 1997.

[22] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proc. Hum. Vis., Visual Process., Digital Display IV*, 1993, pp. 202–216.

[23] 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," in *Comput. Vis. Based Neurobiol.*, vol. 2054, 1994, pp. 32–42.

[24] 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.

[25] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Processing*, 1994, pp. 982–986.

[26] I. Honsch and L. J. Karam, "Adaptive image coding with perceptual distortion control," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 213–222, Mar. 2002.

[27] S. Daly, W. Zeng, J. Li, and S. Lei, "Visual masking in wavelet compression for JPEG2000," *Image Video Commun. Process.*, vol. 3, pp. 1016–1026, 2000.

[28] A. B. Watson, "Probability summation over time," *Vis. Res.*, vol. 19, pp. 515–522, 1979.

[29] 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.

[30] [Online]. Available: http://www.jpeg.org/CDs15444.htm

[31] D. Taubman, "Software architectures for jpeg2000," in *Proc. IEEE Int. Conf. Digital Signal Processing*, vol. 1, 2002, pp. 197–200.

[32] J. Lin, "Visual progressive coding," *Vis. Commun. Image Process.*, vol. 3653, no. 116, pp. 1143–1154, 1999.

[33] Methodology for Subjective Assessment of Quality Of Television Pictures, International Telecommunication Union, Geneva, Switzerland, 2000.

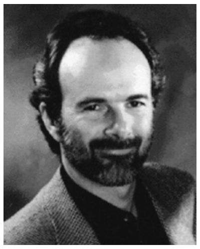
Zhen Liu received the B.S. and M.S. degrees in telecommunications from the Nanjing University of Posts and Telecommunication, China, in 1995 and 1998, respectively, and the Ph.D. degree in electrical engineering from Arizona State University (ASU), Tempe, in 2003.

He is currently a Senior Engineer with the Digital Signal Processing Department, Qualcomm, Inc., San Diego, CA. His research interests are in the areas of image and video compression, transmission, and communication. From 1999 to 2003, he was a Graduate Research Assistant in the Image, Video, and Usability Lab, ASU, and a Graduate Teaching Assistant in the Electrical Engineering Department, ASU. He was an intern at HP Labs and Ricoh California Research Center, where he worked on compound document compression during the summer of 2000 and 2003, respectively.

Lina J. Karam (S'91–M'95–SM'03) received the B.E. degree in computer and communications engineering from the American University of Beirut, Beirut, Lebanon, in 1989, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, in 1992 and 1995, respectively.

She is currently an Associate Professor with the Electrical Engineering Department, 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 Tech. She worked at Schlumberger Well Services on problems related to data modeling and visualization and in the Signal Processing Department of AT&T Bell Labs on problems in video coding during 1992 and 1994, respectively.

Dr. Karam is the recipient of an NSF CAREER Award. She served as the Chair of the IEEE Communications and Signal Processing Chapters in Phoenix in 1997 and 1998. She also served as an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 1999 to 2003. She is currently an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS. She is an elected member of the IEEE Circuits and Systems Society's DSP Technical Committee, the IEEE Signal Processing Society's IMDSP Technical Committee, and a member of the IEEE Signal Processing Society's Conference Board. She is also a member of the Signal Processing and Circuits and Systems Societies of the IEEE.



Andrew B. Watson is the Senior Scientist for Vision Research at NASA Ames Research Center, Moffett Field, CA, where he works on models of vision and their application to visual technology. He has developed widely used models of motion perception and spatial pattern perception, as well as methods of efficient data collection. He has published extensively on topics such as spatial and temporal sensitivity, motion perception, image quality, and neural models of visual coding and processing. He is the author of five patents in areas such as image compression, video quality, and detection of artifacts in the display manufacturing process. In 2001, he founded the *Journal of Vision*, for which he now serves as Editor-in-Chief. He also serves as an Associate Editor for the journal *Displays*.

Dr. Watson is a Fellow of the Optical Society of America. In 1990, he received NASA's H. Julian Allen Award for an outstanding scientific paper, and in 1993, he was appointed to Ames Associate Fellow for exceptional scientific achievement.