Joint perceptually-based Intra prediction and quantization for HEVC

Guoxin Jin*‡, Robert Cohen†, Anthony Vetro† and Huifang Sun†
* Northwestern University, Evanston, IL, USA
E-mail: gjin@u.northwestern.edu Tel: +1-847-4915410

† Mitsubishi Electric Research Laboratories, Cambridge, MA USA
E-mail: {cohen,avetro,hsun}@merl.com Tel: +1-617-6217503

***Abstract*—This paper proposes a new coding scheme which jointly applies perceptual quality metrics to prediction, quantization and rate-distortion optimization (RDO) within the High Efficiency Video Coding (HEVC) framework. A new prediction approach which uses template matching is introduced. The template matching uses a structural similarity metric (SSIM) and a Just-Noticeable Distortion (JND) model. The matched candidates are linearly filtered to generate a prediction. We also modify the JND model and use Supra-threshold Distortion (StD) as the distortion measurement in RDO. Experimental results showing improvements for coding textured areas are presented as well.**

I. Introduction

Image and video coding is one of the most critical techniques in modern multimedia signal processing. The state-of-the-art image codecs such as JPEG 2000 and JPEG XR, and video codecs such as H.264/AVC and the imminent High Efficiency Video Coding (HEVC) [1] standard can compress image or video at considerably low bit rates with good quality. However, the quality and mode-dependent decisions inside the codecs are typically measured using mean-square error, which is related to PSNR, or mean-absolute error, which are well known metrics that do not necessarily related to the ultimate signal receiver, i.e. human vision system (HVS). The successful use of perceptual coding for audio signal compression, e.g. MP3, proved that by exploiting the characteristics of human perceptual system, redundant information can be discarded without noticeable distortion, and thus the bit-rate of the coded signal can be significantly reduced. Although spatial or frequency component correlations are widely considered with techniques such as motion compensation, template matching, and adaptive prediction, perceptually optimized coders are less studied. Additionally, the distortion that could not be perceived by the HVS is usually not measured objectively [2].

To study HVS, many researchers conduct subjective tests using simple excitation such as uniform luminance blocks, sinusoid gratings and Gabor patches to determine the detection threshold of distortion or of the signal. These experimental results are related to Just-Noticeable-Distortion (JND) and they are modeled mathematically in order to be used in the image/video codecs. Theoretically as long as the distortion or signal level is below JND, it should not be perceived by the HVS (perceptually lossless). In an image/video coding context, ideally the coder would only allocate bits for signaling portions of the image for which the distortion that greater than or equal to JND or the corresponding Supra-threshold Distortion (StD) [2].

A. Human Vision System and Just Noticeable Distortion

HVS is a very complex system for which much is left to be understood. At the lower level, HVS is known to perform a sub-band decomposition. Also HVS does not consider different visual information, e.g. intensity and frequency, as having the same importance [3].

Psychophysics studies shows four aspects affect the detection threshold of distortion (or signal) in HVS. They are luminance adaptation, contrast sensitivity, contrast masking and temporal masking [2], [4].

Luminance adaptation indicates the nonlinear relationship between perceived luminance and true luminance displayed [2]. Luminance adaptation is also called luminance masking since the luminance of the signal masks the distortion. The luminance adaptation is usually tested by showing a patch of uniform luminance as the excitation against the background which has different luminance. The detection sensitivity is modeled by the Weber-Fechner Law [5] such that when the excitation is just noticeable, the luminance difference between the patch and background divided by luminance of background is a constant. In other words, the brighter the background is, the higher the detection threshold will be, meaning that the sensitivity to distortion is lower. However, due to the ambient illumination of display devices [6], the masking in very dark regions would be stronger than that in very bright regions.

Contrast sensitivity refers to the reciprocal of the contrast detecting threshold, which is the lowest contrast at which the viewer can just barely detect the difference between the single frequency component (a sinusoidal grating) and the uniform luminance background. Contrast here means the peak-to-peak amplitude of sinusoidal grating [3]. The sensitivity varies in depending upon the background luminance. As early as 1967, a experiment led by van Nes and Bouman [7]

‡This work was done while at MERL

showed sinusoids of light for different wavelengths (Red, Green and Blue) and different luminance levels to human viewers. They found that when the background luminance is low (< 300 Td) the detection threshold obeys de Vries-Rose law with respect to frequency, in which the threshold increases in proportion to the reciprocal of the square root of the luminance. When the background luminance is high (> 300 Td) then the detection threshold follows the Weber-Fechner law. The contrast sensitivity function has important impact on later research of perceptual image/video coding.

Contrast masking is the effect of reducing the perceivability of the distortion (or signal) by the presence of a masking signal. For example, many coding artifacts in the complex regions such as tree leaves and sand are less visible than those in the uniform regions such as the sky. In this case the high spatial-frequency components in complex regions mask the high spatial-frequency components in the artifacts. The masker usually has a similar spatial location and spatial-frequency components as the distortion (or signal) [2]. Therefore, contrast masking is sometimes been called texture masking [4], [6]. Contrast masking was studied Legge and Foley [8], who experimented with sinusoidal gratings having different frequencies and grate widths. The results show that the detection threshold for the high contrast masker follows a power law, and the low contrast masker reduces the detection threshold. By quantitatively measuring the detection threshold for different background luminance of varieties of subjects, the *threshold modulation curves*, namely the adjusted contrast sensitivity function (CSF) is plotted.

Temporal masking in video coding refers to the reduction in the perception of distortion (or signal) in the presence of high temporal frequencies. When the motion is fast, details in individual video frames become more difficult to detect. In addition to depending upon the temporal frequency, temporal masking also depends upon a function of spatial frequency [9]. The sensitivity is modeled as a band-pass filter at low spatial frequencies and a low-pass filter at high spatial frequencies [10].

Beside the four aspects mentioned above, some studies also involve a foveation model [11], [12], especially for video coding. Because of the nature of HVS [3], the resolution is extremely high at the fixation point in the region of interest. Thus, foveation can be modeled as a low pass filter around the fixation point in the region of interest. The region away from fixation points can be coded using lower rates.

Since the just noticeable distortion (JND) is a nonlinear function of HVS characteristics, a pooling strategy for JND using above factors can be multiplication [13], maximum [6] or summation [14].

Once the encoder has the JND model for the image or for the local blocks of image, an uniform quantizer with a step size twice the JND value is usually used to quantize the image itself, the transformed image, or a prediction residue depending on the coding scheme and JND model. More details will be discussed in Section III.

B. Perceptual Quality Metrics

Besides JND, another important technique that can improve perceptual image/video coding is the quality metric which approximates subjective characteristics of a viewer. Instead of considering quality only from the Signal-to-Noise Ratio (SNR) point of view, many pioneering metrics based on the nature of HVS has been proposed in the past two decades. An up-to-date survey has been done by Lin and Kuo [4] where the perceptual quality metric is classified into model based and signal driven. The model-based perceptual quality metrics use the idea of filter banks similar to HVS such as the Cortex Transform, Gabor filters, and steerable filters to decompose the image into sub-bands and analyze the perceptual characteristics to quantitatively determine the quality. A good example of a model-based metric is the visible difference predictor (VDP) in [15]. Since model-based metrics are computationally expensive, signal-driven metrics, which do not try to build a HVS model, are preferred. Signal-driven metrics extract only the characteristics or statistics of a signal for evaluation purposes.

Within the category of signal-driven metrics, structural approaches show great success in image processing. The quality should be measured in the sense of structural similarity between the original image and the coded version. A good coder can reconstruct totally different images or image blocks in the MSE sense, without effecting the viewing quality [16]. The pitfall of MSE or PSNR is that images are strongly locally correlated 2D signals which do not convey information at only the single pixel level. In fact the information containers are shapes, patterns, colors, edges and so forth. A good metric should maximize similarities and be invariant to translation, scaling and rotation. Moreover, the metric should also invariant to light intensity and chroma changes. A well-known structural similarity metric (called SSIM) was first introduced for the spatial domain and later applied to sub-bands in [17]. The SSIM metric is defined as:

$$SSIM(\mathbf{x},\mathbf{y}) \triangleq \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (1)$$

The original image x (or block) and the reconstructed image y (or block) are decomposed into S levels and L orientations using steerable filters [18]. As a result, there will be $S \times L$ sub-bands. The local mean, variance of corresponding sub-bands, and the covariance between x and y in each sub-band are computed using a small sliding window. For each sub-band, the local SSIM score is computed using (1). The overall SSIM score is a floating-point number between 0 and 1, which is the arithmetic average of all scores over all sub-bands. A more complex and accurate metric called structure texture similarity (STSIM) [19] improved upon SSIM by introducing statistics between sub-bands with different orientations and scale. STSIM also discards the σ_{xy} term from SSIM. In this paper, we use SSIM for simplicity.

C. Proposed Method

The typical flow for compressing pictures in state-of-the-art video coders such as HEVC [1] and H.264/AVC [20]

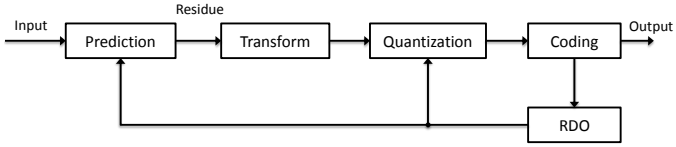


Fig. 1. Common Predictive Encoder.

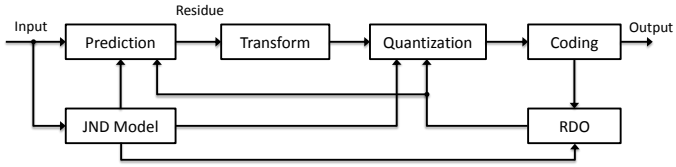


Fig. 2. Proposed Encoder.

is prediction, transform, quantization and entropy coding, controlled via rate-distortion optimization (RDO) as shown in Fig. 1. The approach proposed in this paper, as shown in Fig. 2, exploits a local SSIM and JND-based model in the predictor, quantizer, and during RDO.

H.264/AVC and HEVC iteratively use RDO to find the optimal prediction and quantization schemes. Existing perceptual based image coding methods either focus on adjusting the quantization matrix based on JND metrics or on using perceptual-based metrics to perform RDO. However, to our knowledge, there is not a method that applies perceptual based metrics to find a perceptually-optimal prediction jointly with a perceptually adjusted quantizer.

D. Organization

The remainder of this paper is organized as follows: Section II gives an overview of related work on perceptual coding. Section III proposes a modification to an existing quantization method based on a JND model. A proposed perceptual prediction scheme is described in detail in Section IV. Experimental results are shown in Section V. Conclusions are given in VI.

II. Related Work

Many researchers have studied the perceptual distortion visibility model or equivalent CSF model since the 1960s. Data for JND models come from psychophysical experiments. Using these models, researchers proposed variants of coding algorithms. In general the models are classified into spatial domain models which use local pixel values to determine the detection threshold of distortion, and sub-band domain models which usually adjust the CSF to find the distortion tolerance in different sub-bands.

A. Sub-Band Domain JND Model and Perceptual Coding

Safranek and Johnston [21] introduced a contrast masking model in a generalized quadrature mirror filter (GQMF) sub-band domain given a uniform background gray level of 127. The baseline sub-band sensitivity and sensitivity adjustment for different luminance values are measured subjectively and tabulated. The texture masking is computed by the texture energy of each of the sub-bands. The overall sensitivity is the product of the baseline sensitivity, luminance adjustment and texture masking. Each sub-band is then DPCM-coded and quantized using the overall sensitivity as the quantization steps.

Peterson et al. [22] performed a subjective experiment on sensitivity for different colors (RGB) with 8x8 DCT basis functions. They found that the DC sensitivity plot has a 'U-shape' vs. the background luminance, and the AC sensitivity logarithmically increases with respect to the basis function magnitude. Based on the experiments, Peterson et al. generated a quantization matrix that could be used in a visually lossless coding scheme.

Based on van Nes and Bouman's results [7], Ahumada and Peterson [23] applied a parabolic fitting of 1D contrast sensitivity experimental results, and build a 2D CSF by orthogonally combining two 1D CSFs with consideration of the correlation between different dimensions. The CSF is a function of luminance which can be estimated from the image. A parabolic model was used to compute the CSF in DCT sub-bands and was applied to the quantization matrix.

Foley and Boynton [24] introduced a new contrast masking model for HVS. They found the response of HVS depends not only on the excitation of the receptive field, but also on inhibitory inputs. The relationship is modeled as excitation divided by inhibition called target threshold vs. masker contrast function (TvC). They use sinusoidal gratings, like [22], and Gabor patterns for different orientations, phases and frequencies of the signal. Both the target and masker are the linear sum of signals. Also, both target and masker are orientation-dependent functions.

A famous perceptual coder called DCTune [25] uses luminance adaptation [23] and contrast masking [8] to adjust the quantization matrix in DCT sub-bands. The distortion was measured as in [24]. He also introduced error pooling of JND via the Minkowski metric. Later, Watson [26] performed a series of subjective experiments to test the sensitivity of discrete wavelet transform (DWT) components. The results were fitted to the exponential model.

Visible Difference Predictor (VDP) [15] is an image quality metric considering luminance adaptation, contrast sensitivity and contrast masking. VDP assumes the luminance adaptation occurs locally around a fixation area. A 2D CSF model in the frequency domain is proposed. A Cortex Transform is used to decompose the image.

Hontsch and Karam [27] modified the JND model in [21] for a 16 sub-band GQMF decomposition. The model considers the luminance adaptation and Inter/Intra masking. The base luminance sensitivity is measured numerically without a closed form. Compared to the coding scheme in [21], the JND measurement is adapted locally without the need for transmitting side information. Later in [13], the authors also tried to use JND modeled by a product of luminance-adjusted CSF as in [23] and Contrast Masking Adjustment [24] in DCT sub-bands. The local luminance computation involves only data in the foveation region. A new quality metric is also proposed which is based upon the perceptual error-detecting probability. However, the work in [13] does not accurately model for

luminance adaptation in the dark and bright regions, as well as where HVS has lower visibility threshold at/around edges. Zhang et al. [28] improved upon [13] by expressing the JND as the product of the base threshold and elevation factor for each DCT sub-band.

Liu et al. [29] extended the work of [13] to JPEG 2000 encoding by incorporating JND into DWT sub-bands. The model is the product of contrast sensitivity as in [26], luminance adaptation [25] and contrast masking adjustment. The contrast masking consists of a self-contrast adjustment [13] and neighborhood contrast adjustment [30].

Leung and Taubman [31] assumed that movement in the scene can be tracked easily by the eye such that compressed artifacts are most noticeable at that moment. Thus, the authors use an adaptive masking slope to model the Supra-threshold Distortion (StD) in DWT sub-bands. The visual sensitivity is same as in [32].

A DCT-domain JND model for video coding is introduced by [9], which considers contrast sensitivity [23], a newer luminance adaptation model, and texture contrast masking. These aspects are assisted using edge detection and temporal contrast sensitivity, along with temporal frequency associated with eye movement [15].

Jia et al. [33] introduced a spatio-temporal JND model which considers the eye movement for video in the DCT sub-bands. This paper extends [28] to video. The model incorporates spatio-temporal CSF [23], eye movement, luminance adaptation and contrast [28].

Naccari and Pereira [34] used a JND model of a color video signal in the DCT domain. The model is an extension of their earlier work [35], assuming the same masking in chroma and luma, along with temporal masking [9]. In this work, the H.264/AVC quantization matrix is used as the baseline detection threshold, which is then adjusted depending on the average energy of DCT sub-bands.

In [32], a wavelet domain JND method is introduced. The author proposes a method to reduce the prediction error in DWT + DPCM compression for color images with respect to JND in DWT sub-bands. This approach uses JND-based quantization and to estimate JND in the decoder locally. The JND is modeled by luminance adaption [25], [26] and masking adjustments.

B. Spatial Domain JND Model and Perceptual Coding

In [6], the authors proposed a new spatial domain JND model as the maximum of spatial masking and luminance adaption. The authors also introduce the minimally noticeable distortion (MND) as a relaxation of JND, which can be used to control the rate-distortion optimizer.

Foveation Scalable Video Coding (FSVC) [11] is introduced by Wang et al. The critical frequency beyond which no frequency component is perceivable is used to threshold the foveation sensitivity function. Both spatial and DWT domain foveation sensitivity model are built to adapt prediction and scale the bit rate, by shrinking DWT coefficients, respectively.

The work in [14] focuses on motion compensation residues. Additionally, this model is suitable for Intra-frame prediction. It is based on a summation rule, as JND is linear combination of different effects. The authors introduced a spatial domain JND model called non-linear additively model for masking (NAMM) using luminance masking plus texture masking minus cross-effect. The luminance masking is the same as [12], which was initially defined in [6]. The texture masking only considers edge information via a product of gradient and edge intensity.

A motion perception uncertainty measurement is proposed in [36]. The motion self-information is measured by using an approximate power law model. The background likelihood measures the uncertainty of the scene. The importance map is weighted by the difference of self-information and background likelihood.

The authors of [12] proposed a Foveated-JND model as a function of spatial JND, temporal JND and foveated JND, for video coding in the spatial domain. The spatial JND and temporal JND are based on [6]. If multiple foveation points exist, the total JND is the minimum overall.

Gao et al. [37] based their JND model on [14]. Depth Image Based Rendering (DIBR) is used to synthesize JND maps for Multiview Video Coding (MVC). The 2D JND map is then projected onto 3D world coordinates with help of a depth map. The 3D JND map is then back-projected to other camera views. The residue of 4x4 block is thresholded by synthesized JND maps.

C. Perceptual Based Rate-Distortion Optimization

In HEVC [1] and H.264/AVC [20], the encoder uses rate-distortion optimization (RDO) to output the best picture quality with a rate less than a given constraint. This process can be expressed as

$$\min\{D\} \quad \text{subject to} \quad R \leq R_c, \tag{2}$$

where D is the distortion measurement, usually based on MSE. In order to incorporate a perceptual quality measurement, [38] uses $D = 1 - \mathrm{SSIM}$ in H.264/AVC, where SSIM is the structural similarity metric introduced by [17], as in (1). This algorithm produces a better perceptual quality than H.264/AVC.

D. Perceptual Based Template Matching

Template-matching techniques generally use MSE or Mean Absolute Error as a matching criteria [39]. Recent research involves sparse reconstruction as the constraint to solve the matching problem [40]. Lan et al. [41] use template matching to find candidates in the reconstructed image, and they then use the candidates to train a KLT matrix adaptively for transform coding. A structural texture similarity metric (STSIM) [19] is combined with MSE as the structurally-lossless template matching criteria in [16].

III. PERCEPTUAL QUANTIZATION

In this paper, we propose a new perceptual template matching metric using SSIM and JND models. We therefore need to choose a JND model. Because HEVC quantizes prediction residues in the transform domain, the model from Ahumdada and Peterson [23] is feasible for this purpose, with some proposed modifications.

Given an $N\times N$ block in the original image as the coding target, HEVC uses non-uniform scalar quantization to quantize the DCT coefficient residue $\varepsilon_{m,n} = X_{m,n} - \hat{X}_{m,n}$, where $m, n = 0, 1, 2, ..., N-1$ are indexes of DCT sub-bands, $X_{m,n}$ are the DCT coefficients of target block x at sub-band or coefficient location (m, n), and $\hat{X}_{m,n}$ are the DCT coefficients of the prediction of target block at sub-band (m, n). Similar to other Intra-frame coding algorithms, it is well known [25] (DCTune) that the DC coefficient and low frequency AC coefficients should be quantized finer than the high frequency coefficients in order to achieve the best perceptual rate-distortion performance.

For scalar quantization, since the maximum quantization error for a given quantization step is equal to the half of the quantization step size [13], and because JND theoretically is the least distortion that the HVS could perceive, the quantization step size for sub-band (m, n) should be twice the JND $T_{m,n}$, as follows:

$$Q_{m,n} = 2T_{m,n}. \tag{3}$$

Here, we use the JND model in DCT sub-bands T_{DCT}, which is computed by the product of the contrast sensitivity function adjusted by luminance and scaled by DCT coefficients, and a contrast masking adjustment a_{CM} [13]:

$$T_{m,n} = T_{DCT}(m, n) \times a_{CM}(m, n). \tag{4}$$

The CSF is an exponential function depending on the background luminance and sub-band frequency [23], excluding the DC sensitivity for which both m and n are equal to zero:

$$\log_{10} T_{CSF}(m, n) = \log_{10}\left(\frac{T_{min}}{r + (1-r)\cos^2\theta_{m,n}}\right) + K(\log_{10} f_{m,n} - \log_{10} f_{min})^2. \tag{5}$$

The scaled version of CSF which will be used in (4) is [23]

$$T_{DCT} = \frac{MT_{CSF}(m, n)}{2\alpha_i\alpha_j(L_{max} - L_{min})}, \tag{6}$$

where M is the number of gray levels (e.g 256), and L_{max} and L_{min} are the maximum and minimum display luminance. Given a block with N transform coefficients, α_i or α_j are DCT normalization factors:

$$\alpha_i = \frac{1}{\sqrt{N}}\begin{cases}1, & i = 0\\ \sqrt{2}, & i \neq 0\end{cases}$$

$$\theta_{m,n} = \arcsin\frac{2f_{i,0}f_{0,j}}{f_{i,j}^2} \tag{7}$$

$$f_{i,j} = \frac{1}{2N}\sqrt{\left(\frac{i}{\omega_x}\right)^2 + \left(\frac{j}{\omega_y}\right)^2}. \tag{8}$$

Here, ω_x and ω_y are the horizontal width and vertical height of a single pixel in degrees of visual angle [42],

$$w_x = (360/\pi) \times \arctan\left(W_{\text{screen}}/2D_v W_{\text{resolution}}\right)$$
$$w_y = (360/\pi) \times \arctan\left(H_{\text{screen}}/2D_v H_{\text{resolution}}\right)$$

where D_v is the viewing distance from the center of screen.

The expression in (5) is a mathematical approximation of data in [7] using a parabola. Here, f_{min} is the spatial-frequency where the detection threshold is the smallest, i.e. where the sensitivity is the highest. The detection threshold at f_{min} is T_{min} and the bandwidth of the CSF is determined by K. With L being the local background luminance $L = L_{min} + (L_{max} - L_{min})/M$, we get

$$f_{min} = \begin{cases} f_0 L^{\alpha_f} L_f^{-\alpha_f}, & L \leq L_f \\ f_0, & L > L_f \end{cases}$$
$$T_{min} = \begin{cases} L^{\alpha_T} L_T^{1-\alpha_T}/S_0, & L \leq L_T \\ L/S_0, & L > L_T \end{cases}$$
$$K = \begin{cases} f_0 L^{\alpha_K} L_K^{1-\alpha_K}, & L \leq L_K \\ K_0, & L > L_K. \end{cases}$$

The constants Ahumdada and Peterson [23] used in this model are the result of least square fitting of experiment results in [7]: α_T=0.649, α_f=0.182, α_K=0.0706, L_T=13.45 cd/m^2, L_f=L_K=300 cd/m^2, S_0=94.7, f_0=6.78 cyc/deg and K_0=3.125.

However, there are two problems with this CSF model. First, for the DC coefficient, the CSF is undefined, since when both m, n equal zero, $\theta_{m,n}$ in (7) approaches infinity. Second, in the low frequency sub-bands $f_{m,n}$ is quite small as compared to f_{min}, such that the second term in (5) is extremely large. As a result, the model tends to over-estimate the detection threshold in low frequency sub-bands. In order to deal with the first issue, a local luminance adaptation model is used to estimate the detection threshold in DC sub-band. Either [9] or [6] can be used here to model the DC threshold. We use the latter model with local mean μ_x for better quality in HEVC.

$$T_{CSF}(0, 0) = \begin{cases} 17(1 - \sqrt{\mu_x/127} + 3), & if\ \mu_x \leq 127 \\ 3(\mu_x - 127)/128 + 3, & if\ \mu_x > 127 \end{cases} \tag{9}$$

To address the second problem, we clip the minimum value of the spatial frequency to 1 to prevent $f_{min}/f_{m,n}$ from becoming too large.

The contrast masking adjustment is defined as [13],

$$a_{CM}(m, n) = \begin{cases} \max\left(1, \left|\frac{E(|X|)}{T_{DCT}(m, n)}\right|^{\epsilon}\right), & m \neq 0,\ n \neq 0 \\ 1, & otherwise, \end{cases} \tag{10}$$

where $E(|X|)$ is the average magnitude of local DCT coefficients and ϵ=0.6. The contrast masking adjustment uses the excitation/inhibition model of [24], which indicates how much more distortion can be tolerated with the presence of signal (X) in the spatial-frequency domain.

IV. Perceptually-based Intra prediction

Given a block of the original image to be coded as the target, HEVC [1] uses previously-decoded pixel values in the reconstructed image as part of the prediction process. The reconstructed pixels immediately above and to the left of the target, if available, are used as the input of the predictor. There are 33 prediction directions ranging from approximately $-135°$ to $45°$. Two extra modes, DC mode and Planar mode are used to generate flat or bilinear interpolated predictions. Rate-Distortion Optimization (RDO) in HEVC is used to determine the best prediction mode and the best quadtree structure of a block.

HEVC Intra prediction, as well as with many other image coders such as H.264/AVC Intra prediction, consider the spatial correlation in the local 2D signal. As a result, a "good" prediction that minimizes the rate-distortion cost is expected. The faithfulness of the prediction is measured in the mean square error (MSE) sense such that the best prediction will give least MSE with respect to the target. In some cases, metrics with lower complexity than MSE are used to choose a subset of Intra prediction modes. However, in Section I-B we described that MSE or similar absolute error metrics do not always faithfully represent the perceptual quality of an image. To obtain a predictor that is more consistent with the HVS-related aspects of images, a perceptual quality metric can be used while selecting coding modes. Moreover, authors of [16] showed that one can find many structurally lossless candidates to replace the target in the reconstructed image, rather than relying on only the neighboring pixels adjacent to the target. In this paper, we propose new a Intra prediction scheme which uses function of JND and SSIM as the quality metric and tries to find a prediction from the structurally lossless candidates.

As shown in Fig. 3, the proposed predictor use the neighboring previously-decoded blocks above and to the left of the target block as the template. The template is matched using a perceptual quality metric to find up to K candidates having the best quality. A filter is used to combine the best candidates to generate a prediction of the target. There is some flexibility in how this is accomplished. First, the number of best candidates K can vary depending on the filter and the quality metric. For example, the predictor would only need one candidate without filtering or K=16 candidates with a median filter. Second, this scheme can be used as an extra mode in HEVC in addition to the existing 35 modes, or this scheme can be the only mode that HEVC uses, to reduce the bits needed to signal the mode index. Third, since HEVC uses a quad-tree structure to recursively code a target block, a cost function is used to determine the structure of quad-tree. This cost function will be described later in Section IV-D. In HEVC, the cost function is MSE or similar. In our coding scheme, we will use perceptual quality metrics.

The detailed prediction algorithm is shown in Fig. 4. A target block x has both left b_1 and upper b_2 neighboring blocks having the same block size as the target. For each candidate in the search region, which is generally 3 to 5 times the block

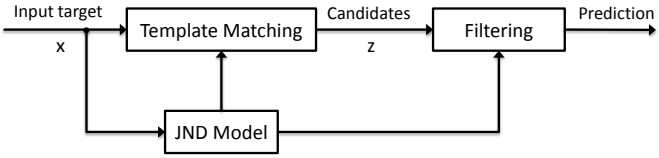


Fig. 3. Prediction Model

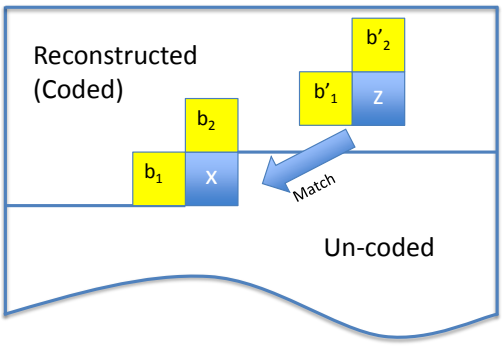


Fig. 4. Template Matching

size, a quality associated with the candidate is estimated as

$$D = \frac{D_Q(b_1, b_1') + D_Q(b_2, b_2')}{2}, \tag{11}$$

where b_1' and b_2' are left and upper blocks of the candidate, and D_Q is the perceptual quality measurement. This measurement is a function of structural similarity metrics and JND distortion measurement. For example

$$D_Q(x, y) = \frac{\text{SSIM}(x, y)}{\gamma log(D_{StD})}, \tag{12}$$

or

$$D_Q(x, y) = \text{SSIM}(x, y),$$

where SSIM is defined in (1), γ is a constant scale factor and D_{StD} is the supra-threshold distortion of y as compared to x [13]. Given x is the target block and y is the predicted block, let X, Y represent the DCT transform of x, y respectively, the supra-threshold distortion for the block is

$$D_{StD} = \left\| \frac{|X - Y|}{T} \right\| = \left\| \frac{\varepsilon}{T} \right\|, \tag{13}$$

where T is the local JND in DCT sub-bands for this target block from (4) as described in Section III.

If the target x has only one of the upper or left neighboring blocks, then for each candidate only the available neighboring block is used to compute D in (11). If the target x is the first block processed in the image, then a uniform block with pixel values of 128 is used as the prediction. All the perceptual quality measurements of candidates in the search region are computed, except for the initial target. The measurements are sorted and the candidates with best K measurements are selected. If in some case the number of valid candidates is less than K, then $K' < K$ candidates are also acceptable.

Once the best candidates have been determined, we propose two prediction schemes. The prediction $\hat{x}$ of target x is

a linear combination of up to K best candidates $z_k, k = \{0, 1, ..., K-1\}$. The following subsections show two algorithms to find the weights w_k for linearly combining the candidates z_k.

A. Joint Prediction and Quantization with Weight Signaling (WS)

Let the DCT transform of the target be $X_{m,n}$ and the DCT transform of candidates be $Z_{k,m,n}$, $m, n = \{0, 1, 2, ..., N-1\}, k = \{0, 1, 2, ..., K-1\}$. The prediction in spatial domain is denoted as $\hat{x}_{i,j}$, $i, j = \{0, 1, 2, ..., N-1\}$ and its DCT transform as $\hat{X}_{m,n}$. The residue of DCT coefficients is then

$$\varepsilon_{m,n} = X_{m,n} - \hat{X}_{m,n}, \quad m, n = \{0, 1, ..., N-1\}. \tag{14}$$

We would like to have the prediction minimizes the total quantization levels. At sub-band (m, n), the quantization level is the residue $\varepsilon_{m,n}$ in DCT sub-band (m,n) divided by the quantization step $T_{m,n}$ (ignore rounding if possible). Thus the predictor will try to find the weights $w_k, k = 0, 1, ..., K-1$ that minimize the p-norm of quantization levels:

$$\begin{aligned} w_k^* &= \arg\min_{w_k} \sum_{m,n} \left| \frac{\varepsilon_{m,n}}{2T_{m,n} \max\left(\left|\frac{E(|X|)}{T_{m,n}}\right|^{\epsilon}, 1\right)} \right|^p \\ &= \arg\min_{w_k} \sum_{m,n} \left| \frac{\sum_{i,j}(\sum_k w_k z_{i,j} - x_{i,j})\psi_{i,m}\psi_{j,n}}{2T_{m,n} \max\left(\left|\frac{E(|X|)}{T_{m,n}}\right|^{\epsilon}, 1\right)} \right|^p \\ &= \arg\min_{w_k} \sum_{m,n} \left| \frac{\sum_k w_k Z_{m,n} - X_{m,n}}{2T_{m,n} \max\left(\left|\frac{E(|X|)}{T_{m,n}}\right|^{\epsilon}, 1\right)} \right|^p, \end{aligned} \tag{15}$$

where as in (10), $\epsilon = 0.6$, $E(|X|)$ is the mean magnitude of sub-band coefficients of target, and $\psi_{i,m}, \psi_{j,n}$ comprise the DCT basis. The expression (15) comes from the linearity of DCT. The exponent p is typically 1 or 2.

B. Joint Prediction and Quantization Weighted by Quality Predictor (WQ)

A simpler predictor can be built by assuming that the high quality candidates should have higher weights.

$$w_k = \frac{D_k}{\sum_k D_k}, \quad k = 0, 1, ..., K-1, \tag{16}$$

where $D_k = D(x, z_k)$ is the perceptual quality measurement between target x and candidate z_k in (11).

C. Comparison of Predictors

An example of the prediction using (11) in HEVC is shown in Fig. 5. The corresponding predictor using the original HEVC prediction method is shown in Fig. 6. We can see that in some regions such as on the grass and on parts of the horse, the prediction using the perceptual metric (11) better resembles the original texture than with the HEVC predictor. However, as the search candidates are limited to a small transition area, the edges are less preserved.

Both the Joint Prediction and Quantization with Weight Signaling (WS) and Joint Prediction and Quantization Weighted by Quality (WQ) methods find up to K weights. As a result, the prediction will be a linear combination of candidates:

$$\hat{x} = \sum_k w_k^* z_k, \quad k = 0, 1, ..., K-1. \tag{17}$$

Since the candidates are computed via template matching using a perceptual quality metric in the reconstructed image as discussed in Section I-B, both the encoder and decoder can find the same candidates, eliminating the need for the encoder to signal the locations of these candidates. WS requires transmitting weights w_k^*, so that $\hat{x}$ can be computed directly from the best K candidates and weights received by the decoder. WQ, however, does not require extra bits to be signaled, since the weights are a function of quality measurements. In both schemes, the local mean of the magnitudes of target DCT coefficients $E(|X|)$ is used. However, this value is not available at the decoder. To resolve this issue, the decoder can estimate $E(|X|)$ using the mean magnitude DCT coefficients of candidates and/or of the neighboring block of targets as in [13], [35].

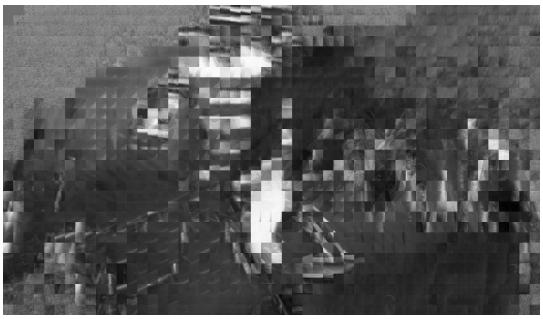


Fig. 5. Prediction using perceptual metric in HEVC with JND scaling $\zeta = 2.5$.

Fig. 6. Prediction using unmodified HEVC with $QP = 28$.

D. Rate-Distortion Optimization

Since we use perceptual quality metrics in both prediction and quantization, the metric used in Rate-Distortion Optimization (RDO), which controls the encoder as shown in Fig. 2, should also be modified correspondingly. The QP parameter in HEVC is no longer effective for this purpose in the proposed scheme. In order to control the rate of the coder, the idea of minimally noticeable distortion (MND) [6] is modified and used here. Therefore, T_{CSF} is scaled by a constant $\zeta \geq 1$:

$$|\varepsilon_{m,n}| < \zeta \cdot T_{CSF}(m, n) \cdot a'_{CM}(m, n), \tag{18}$$

where

$$a'_{CM}(m,n) = \begin{cases} max\big(1, \big|\frac{E(|X|)}{\zeta T_{DCT}(m,n)}\big|^{\epsilon}\big), & m \neq 0,\ n \neq 0 \\ 1, & otherwise. \end{cases} \tag{19}$$

Thus, the larger ζ is, the lower rate going to be. The cost function in HEVC is then

$$C = D_{StD} + \lambda R, \tag{20}$$

where D_{StD} defined in (13) and λ is a constant.

V. Experimental results

In order to use the model in Section III, viewing conditions must be specified. The authors used a NEC LCD220WX display with resolution 1680×1050 and screen size 472mm×292mm. The viewing distance D_v depends on the frame size:

$$D_v = 4H_{frame}(H_{screen}/H_{resolution}). \tag{21}$$

The display was calibrated using The Lagom LCD monitor test pages [43]. The maximum luminance of the display device was 300 cd/m^2. After calibration, the display luminance was set to L_{max}=150 cd/m^2. Since the contrast ratio of display is 1000:1, the minimum luminance is L_{min}=0.3 cd/m^2. The bit depth is 8, so M=256. The viewing angle was approximately 0°.

For the simulations, the maximum Coding Unit Size (MaxCU) is set to 32. The Max CU partition Depth is 1, so only 32x32 and 16x16 CUs are used. The Transform Unit (TU) size is restricted to 16. The remaining coding parameters use the default All-Intra Main common test conditions from [44]. HM 6.1 is used as a code base. The luma channel is coded using proposed approach, and chroma channels are coded using the original HEVC methods. γ in (12) is empirically set to 0.2. The value of λ in (20) was decreased by 20% from the original value used in the HEVC encoder, as determined in separate experiments.

Three test sequences were coded: *RaceHorses* (832×480, 20 frames), *Cactus* (cropped to 1920×1024, 20 frames) and *SlideEditing* (cropped to 1280×704, 10 frames). Because WS incurs a rate penalty from signaling weights, we will focus on WQ in this section.

Results for *RaceHorses* are shown in Fig. 7. The original frame is shown in 7(a), and Fig. 7(b) was coded using HEVC with QP=28. The WQ-coded frame using ζ=2.5 is shown in Fig. 7(c). The average bit rates over 20 Intra frames for these examples are 9.2 kbps and 9.6 kbps, respectively. With HEVC, some areas such as the horse are smoothed, as shown in Fig. 8(b) and Fig. 8(a)). The proposed approach does a better job at reproducing the textures, although there still are some blocking artifacts. Moreover, in the pad under the saddle (Fig. 8(c)) and grass region (Fig. 8(d)), the contrast of the proposed approach is improved as compared to HEVC. Note that the PSNR using HEVC in Fig. 7(b) is 35.73 dB and the PSNR for the proposed method in Fig. 7(c) is 33.65 dB. The fact that proposed approach has better perceptual quality shows PSNR is does not always reflect the perceived quality of an

(a) Original

(b) HEVC, QP=28, 35.73 dB, StD=77408; 9207 kbps

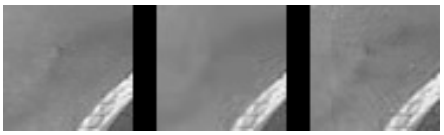
(c) WQ, ζ=2.5, 33.65 dB, StD=48022; 9617 kbps

Fig. 7. *RaceHorses* coded using HEVC and the proposed WQ approach

image. During the encoding process, using (13) to compute the perceptual measure of the prediction error, the total StD over all blocks for this case is 77408 and 48022 for HEVC and WQ, respectively. Additional results are shown in Fig. 9 for *Cactus* and Fig. 10 for *SlideEditing*.

VI. Conclusions

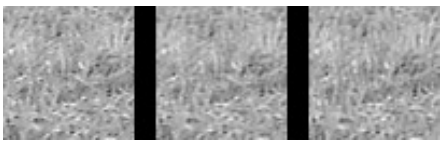
We proposed a new coding scheme which jointly applies SSIM and JND models for prediction, quantization and rate-distortion optimization within HEVC. This work focused on introducing a new predictor which uses template matching along with a perceptual quality measurement to select multiple prediction candidates. The perceptual quality measurement was derived from a joint SSIM and JND model. The selected candidates were filtered either by minimizing the supra-threshold distortion during quantization, or via a linear combination weighted by perceptual quality. We modified the existing JND model and use super-threshold distortion in rate-distortion optimization as well. The proposed approach yields a lower PSNR than HEVC at similar rates. However, the



(a) horse/pad

(b) horse

(c) Saddle



(d) Grass

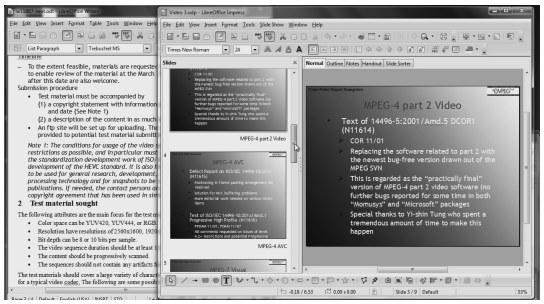
Fig. 8. *RaceHorses* comparison in detail: original (left), HEVC (middle), WQ (right)

(a) Original

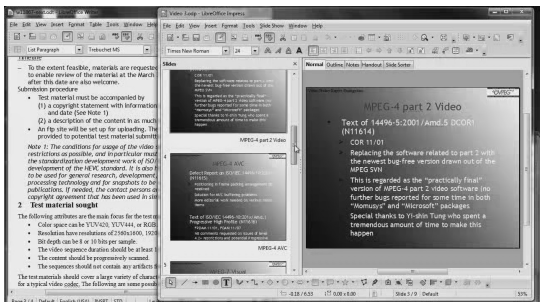
(b) HEVC, QP=31, 35.5 dB, StD=181413; 34874 kbps

(c) WQ, ζ=2.5, 32.5 dB, StD=90675; 36190 kbps

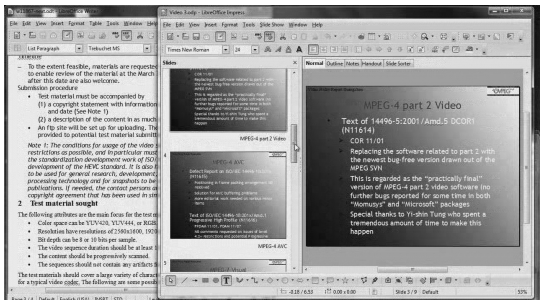
Fig. 9. *Cactus* coded using HEVC and the proposed WQ approach



(a) Original



(b) HEVC, QP=36, 27.7 dB, StD=145421; 18208 kbps



(c) WQ, ζ=2.5, 24.5 dB, StD=91705; 18652 kbps

Fig. 10. *SlideEditing* coded using HEVC and the proposed WQ approach

proposed approach exhibits a better perceptual quality than HEVC. The perceptual quality could also be measured by supra-threshold distortion.

Modeling the perceptual quality of the Human Vision System (HVS) is still an open question. SSIM alone cannot accurately achieve the quality measurement performance of HVS. The existing JND models derived from simple psychophysics studies may not describe HVS thoroughly. As a result, perceivable distortion which is not characterized by the JND model may still exist. Also, bits may be unnecessarily allocated for coding imperceptable distortion. We noticed at low rates, for which the JND scaling ζ is large, there were many blocking artifacts. This would be caused by suboptimal tuning between the proposed coding approach and HEVC tools. The constant λ in rate-distortion optimization should also be further studied.

The proposed approach which performs joint perceptual optimization of prediction and quantization could be generalized for other image and video coding techniques. New JND models and perceptual quality metrics can also be adopted into this paradigm.

REFERENCES

[1] B. Bross, et al., "High efficiency video coding (HEVC) text specification draft 6," JCTVC-H1003, 8th JCT-VC Meeting, San Jose, CA, Feb. 2012.

[2] T. Pappas, R. Safranek, and J. Chen, "Perceptual criteria for image quality evaluation," in *The Essential Guide to Image Processing*, A. C. Bovic, Ed. Academic Press, Jun. 2009.

[3] S. E. Palmer, *Vision Science*. MIT Press, May 1999.

[4] W. Lin and C.-C. J. Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, vol. 22, no. 4, pp. 297–312, May 2011.

[5] E. B. Goldstein, *Sensation and Perception*. Wadsworth, Inc., 1980.

[6] C.-H. Chou and Y.-C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 5, no. 6, pp. 467 –476, Dec. 1995.

[7] F. L. van Nes and M. A. Bouman, "Spatial modulation transfer in the human eye," *J. Opt. Soc. Am.*, vol. 57, no. 3, pp. 401–406, Mar. 1967.

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

[9] Z. Wei and K. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 3, pp. 337–346, Mar. 2009.

[10] J. G. Robson, "Spatial and temporal contrast-sensitivity functions of the visual system," *J. Opt. Soc. Am.*, vol. 56, no. 8, pp. 1141–1142, Aug. 1966.

[11] Z. Wang, L. Lu, and A. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.

[12] Z. Chen and C. Guillemot, "Perceptually-friendly H.264/AVC video coding," in *IEEE International Conference on Image Processing (ICIP)*, Nov. 2009, pp. 3417 –3420.

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

[14] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 6, pp. 742–752, Jun. 2005.

[15] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital images and human vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, Oct. 1993, pp. 179–206.

[16] G. Jin, Y. Zhai, T. Pappas, and D. Neuhoff, "Matched-texture coding for structurally lossless compression," in *IEEE International Conference on Image Processing (ICIP)*, Orlando, Florida, Sep. 2012.

[17] Z. Wang, A. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, Apr. 2004.

[18] J. Portilla and E. P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *Int. J. Comput. Vision*, vol. 40, no. 1, pp. 49–70, Oct. 2000.

[19] J. Zujovic, T. Pappas, and D. Neuhoff, "Structural similarity metrics for texture analysis and retrieval," in *IEEE International Conference on Image Processing (ICIP)*, Cairo, Egypt, Nov. 2009, pp. 2225–2228.

[20] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[21] R. Safranek and J. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May. 1989, pp. 1945–1948 vol.3.

[22] H. Peterson, H. Peng, J. Morgan, and W. Pennebaker, "Quantization of color image components in the DCT domain," in *SPIE, Proc. in Human Vision, Visual Processing, and Digital Display*, vol. 1453, Jun. 1991, pp. 210–222.

[23] J. Albert J. Ahumada and H. A. Peterson, "Luminance-model-based DCT quantization for color image compression," *SPIE, Human Vision, Visual Processing, and Digital Display III*, vol. 1666, p. 365, Sep. 1992.

[24] J. M. Foley and G. M. Boynton, "New model of human luminance pattern vision mechanisms: analysis of the effects of pattern orientation, spatial phase, and temporal frequency," in *Proc. SPIE*, T. B. Lawton, Ed., vol. 2054, no. 1. SPIE, Mar. 1994, pp. 32–42.

[25] A. Watson, "Visually optimal DCT quantization matrices for individual images," in *Data Compression Conference, 1993. DCC '93.*, Mar. 1993, pp. 178–187.

[26] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Am. A*, vol. 14, no. 9, pp. 2379–2391, Sep. 1997.

[27] I. Hontsch and L. Karam, "Apic: adaptive perceptual image coding based on subband decomposition with locally adaptive perceptual weighting," in *IEEE International Conference on Image Processing (ICIP)*, vol. 1, Oct. 1997, pp. 37–40 vol.1.

[28] X. Zhang, W. Lin, and P. Xue, "Improved estimation for just-noticeable visual distortion," *Signal Processing*, vol. 85, no. 4, pp. 795–808, Apr. 2005.

[29] Z. Liu, L. Karam, and A. Watson, "JPEG2000 encoding with perceptual distortion control," *IEEE Trans. Image Process.*, vol. 15, no. 7, pp. 1763–1778, july 2006.

[30] W. Zeng, S. Daly, and S. Lei, "An overview of the visual optimization tools in JPEG 2000," *Signal Processing: Image Communication*, vol. 17, no. 1, pp. 85–104, Oct. 2002.

[31] R. Leung and D. Taubman, "Perceptual optimization for scalable video compression based on visual masking principles," *IEEE Trans. Cir. and Sys. for Video Technol.*, vol. 19, no. 3, pp. 309–322, Mar. 2009.

[32] K.-C. Liu, "Prediction error preprocessing for perceptual color image compression," in *EURASIP Journal on Image and Video Processing*, Mar. 2012.

[33] Y. Jia, W. Lin, and A. Kassim, "Estimating just-noticeable distortion for video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 820–829, Jul. 2006.

[34] M. Naccari and F. Pereira, "Advanced H.264/AVC-based perceptual video coding: Architecture, tools, and assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 6, pp. 766–782, Jun. 2011.

[35] ——, "Comparing spatial masking modelling in just noticeable distortion controlled H.264/AVC video coding," in *11th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, Apr. 2010, pp. 1–4.

[36] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Am. A*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[37] Y. Gao, X. Xiu, J. Liang, and W. Lin, "Perceptual multiview video coding using synthesized just noticeable distortion maps," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, May. 2011, pp. 2153 –2156.

[38] S. Wang, A. Rehman, Z. Wang, S. Ma, and W. Gao, "Ssim-motivated rate-distortion optimization for video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 4, pp. 516–529, Apr. 2012.

[39] Y. Hashidume and Y. Morikawa, "Lossless image coding based on minimum mean absolute error predictors," in *SICE, 2007 Annual Conference*, Sep. 2007, pp. 2832–2836.

[40] A. Dremeau, M. Turkan, C. Herzet, C. Guillemot, and J.-J. Fuchs, "Spatial intra-prediction based on mixtures of sparse representations," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, Oct. 2010, pp. 345–349.

[41] C. Lan, J. Xu, F. Wu, and G. Shi, "Intra frame coding with template matching prediction and adaptive transform," in *IEEE International Conference on Image Processing (ICIP)*, Sep. 2010, pp. 1221–1224.

[42] University of Toronto. (2008) Visual angle calculator. Affect and Cognition Laboratory. [Online]. Available: http://www.aclab.ca/wiki/images/b/b3/Visual_Angle_Calculator.xls

[43] H.-K. Nienhuys. (2008) The lagom LCD monitor test pages. [Online]. Available: http://www.lagom.nl/lcd-test/

[44] F. Bossen, "Common test conditions," JCTVC-H1100, 8th JCT-VC Meeting, San Jose, CA, Feb. 2012.