Perceptually-Friendly H.264/AVC Video Coding Based on Foveated Just-Noticeable-Distortion Model

Zhenzhong Chen, *Member, IEEE,* and Christine Guillemot, *Senior Member, IEEE*

***Abstract*—Traditional video compression methods remove spatial and temporal redundancy based on the signal statistical correlation. However, to reach higher compression ratios without perceptually degrading the reconstructed signal, the properties of the human visual system (HVS) need to be better exploited. Research effort has been dedicated to modeling the spatial and temporal just-noticeable-distortion (JND) based on the sensitivity of the HVS to luminance contrast, and accounting for spatial and temporal masking effects. This paper describes a foveation model as well as a foveated JND (FJND) model in which the spatial and temporal JND models are enhanced to account for the relationship between visibility and eccentricity. Since the visual acuity decreases when the distance from the fovea increases, the visibility threshold increases with increased eccentricity. The proposed FJND model is then used for macroblock (MB) quantization adjustment in H.264/advanced video coding (AVC). For each MB, the quantization parameter is optimized based on its FJND information. The Lagrange multiplier in the rate-distortion optimization is adapted so that the MB noticeable distortion is minimized. The performance of the FJND model has been assessed with various comparisons and subjective visual tests. It has been shown that the proposed FJND model can increase the visual quality versus rate performance of the H.264/AVC video coding scheme.**

***Index Terms*—Bit allocation, foveation model, H.264/advanced video coding (AVC), human visual system (HVS), just-noticeable-distortion (JND), video coding.**

I. Introduction

TRADITIONAL compression methods, from JPEG [1] to JPEG2000 [2], from H.261/263 [3], [4], MPEG-1/2/4 [5]–[7] to H.264/advanced video coding (AVC) [8], attempt to remove spatial and temporal statistical redundancies of the visual signal for compression. The commonly used techniques include transform, motion estimation/compensation, intra/inter prediction, entropy coding, etc. To improve the performance of video compression systems, the human visual system needs to be well understood and carefully utilized. Early attempts to remove *perceptual redundancy* in signal compression by exploiting human perception mechanisms can be found in [9], [10]. A number of image and video coding methods aiming to account for psychovisual properties of the human visual system (HVS) in the quantization and rate allocation problems have then been proposed [11]–[19]. More recently, the authors in [20] have proposed a visual distortion sensitivity model, exploiting the non-uniform spatio-temporal sensitivity characteristics of the HVS, for video coding. Macroblocks (MBs) visual distortion sensitivity is analyzed based on both motion and textural structures. Fewer bits are allocated to MBs in which higher distortion can be tolerated. The bit rate of the whole video can be reduced accordingly. This approach has been further extended in [21] by considering extra spatial and temporal cues based on a better distortion sensitivity model. These cues include motion attention, spatial-velocity visual sensitivity, and visual masking. It is known that human cannot perceive the fine-scale variation of visual signal due to the psychovisual properties of the HVS and the visual sensitivity of HVS can be measured by a spatio-temporal contrast sensitivity function (CSF) [22]. The spatio-temporal CSF shows the contrast required to detect a flickering grating of different spatial and temporal frequencies. In addition, the sensitivity of the HVS depends on the background luminance. The just-noticeable-distortion (JND) provides cues for measuring the visibility of the HVS [10]. JND refers to the maximum distortion which cannot be perceived by the HVS. It describes the perceptual redundancy of the picture by providing the visibility threshold. The JND model generally exploits the visibility of the minimally perceptible distortion by assuming that the visual acuity is consistent over the whole image. The JND model is used in [23], [24] for image/video coding and transmission. It has been extensively studied and applied in image and video compression and transmission [23]–[27].

However, it is also well known that the HVS is not only a function of the spatio-temporal frequency but is also highly space-variant [28]. The HVS consists of three major components: 1) eyes, 2) visual pathways, and 3) visual centers of the brain. When light enters the pupil, it is focused and inverted by the cornea and lens. Then it is projected onto the retina. The retina is a thin layer of neural cells which are responsible for converting light signals into neural signals,

Manuscript received November 21, 2008, revised May 15, 2009 and August 24, 2009. Date of publication March 18, 2010; date of current version June 3, 2010. This work was carried out during the tenure of an ERCIM "Alain Bensoussan" Fellowship Program. This paper was recommended by Associate Editor Prof. W. Gao.

Z. Chen was with INRIA/IRISA, Campus Universitaire de Beaulieu, 35042 Rennes Cedex, France and is now with the School of Electrical and Electronic Engineering, Nanyang Technological University, 639798, Singapore (e-mail: zzchen@ieee.org).

C. Guillemot is with the Institut National de Recherche en Informatique et Automatique, Campus Universitaire de Beaulieu, Rennes Cedex 35042, France (e-mail: christine.guillemot@inria.fr).

Digital Object Identifier 10.1109/TCSVT.2010.2045912

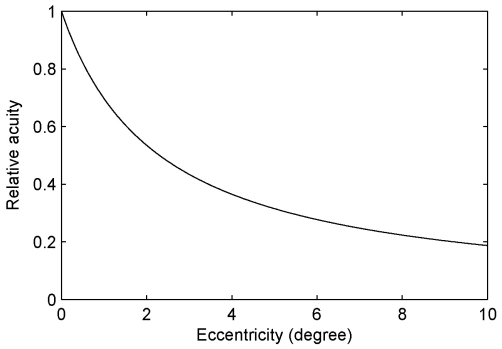


Fig. 1. Visual acuity.

which then transit to the brain through the optic nerve. The retina consists of two classes of photoreceptors: 1) rods, and 2) cones. The rods work under low-ambient light and extract luminance information. The cones are able to perceive finer details and more rapid changes in images since their response times to stimuli are faster than rods. At the center of the retina is the fovea which consists of compact cones and is responsible for our sharp central vision. The HVS is space-variant since the retina in human eye does not have a uniform density of photoreceptor cells. The fovea has the highest density of sensor cells on the retina. The density of the cone and retinal ganglion cells drops with the increased retinal eccentricity. When the visual stimulus is projected on the fovea, it is perceived at the highest resolution. The visual acuity decreases with increased distance, or eccentricity, as shown in Fig. 1, from the fixation point. Therefore, a foveation model considered for video compression will lead to a more intelligent representation of visual scenes [29]–[33]. As the visual acuity is higher at the foveation point, an object in the scene located at the foveation point should be coded with a better quality [34]. Foveated image and video compression algorithms [31], [32], [35] have thus been proposed to deliver high-quality video at reduced bit rates by seeking to exploit this property which results from a nonuniform sampling of the human retina. Similar ideas have been exploited to enhance the quality of regions of interests (ROI) in image and video [36]–[41]. In [42], skin color-based face detection is used to locate the ROI, and weighting factors derived from the JND model are used for adapting the rate in the different regions. With the evolution of computational techniques of visual attention [43]–[45] based on feature analysis, attention properties have been considered in recent image and video coding applications [34], [46], [47]. The perceptual quality is improved in these image/video coding systems by improving the quality of the region of attention.

This paper presents a foveated JND (FJND) model which incorporates both visual sensitivity and foveation features of the HVS. Since the perceptual acuity decreases with increased eccentricity, the visibility threshold of the pixel of the image increases when the distance between the pixel and the fixation point increases. The spatio-temporal JND (STJND) model can thus be improved by accounting for the relationship between visibility and eccentricity. A foveation model is first developed and then incorporated to the STJND model. The proposed FJND model is then used in H.264/AVC video coding. For each MB, the quantization parameter is optimized based on the distortion visibility thresholds given by the FJND model. Macroblocks in which larger distortion can be tolerated are more coarsely quantized, and the saved bit rate can be allocated to MBs where visual sensitivity is higher. The Lagrange multiplier in the rate-distortion optimization is thus adapted so that the MB noticeable distortion is minimized. The reminder of this paper is organized as follows. The next section presents the background and motivation of our work. The FJND model is introduced in Section II. The use of the FJND model in H.264/AVC video coding is presented in Section III. Experimental results and the conclusion are given in Sections IV and V, respectively.

II. FJND Model

As we have mentioned in the above section, we consider both the visual sensitivity and space-variant properties of the HVS. The spatio-temporal CSF describes relative sensitivity of the HVS to the distortion at different spatio-temporal frequencies. As indicated by Weber's law [48], the perceptible luminance difference of a stimulus depends on the surrounding luminance level, i.e., the HVS is sensitive to luminance contrast rather than the absolute luminance level. For example, the distortion is most visible against a mid-gray background. The distortion in the very dark or very bright background is more difficult to be distinguished. The JND model measures the visibility of the HVS according the characteristics of the visual signal. As studied in the literature, a JND model could be built based on the visual sensitivity according to luminance contrast and masking effects. Masking effect is complicated and generally refers to the perceptibility of one signal in the presence of another signal in its spatial, temporal, or spectral vicinity [10]. In addition, the visibility is related to the retinal eccentricity. The visual acuity decreases with increased distance between the fixation point and the pixel and the JND threshold increases accordingly. Therefore, we define the FJND model as a combination of the spatial JND (SJND), temporal JND (TJND), and foveation models

$$\mathrm{FJND}(x, y, t, v, e) = f(\mathrm{SJND}(x, y), \mathrm{TJND}(x, y, t), \mathrm{F}(x, y, v, e)) \quad (1)$$

where $\mathrm{FJND}(x, y, t, v, e)$, $\mathrm{SJND}(x, y)$, $\mathrm{TJND}(x, y, t)$, and $\mathrm{F}(x, y, v, e)$ denote FJND, SJND, TJND, and the foveation model, respectively. t is the frame index, v is the viewing distance, and e is the eccentricity for the point (x, y) relative to the fixation point (x_f, y_f).

Before introducing the proposed foveation and FJND models, let us first review the spatial and temporal JND described in [23] and [24], respectively.

A. Spatial JND Model

The perceptual redundancy in the spatial domain is mainly based on the sensitivity of the HVS due to luminance contrast and spatial masking effect. Various computational JND models

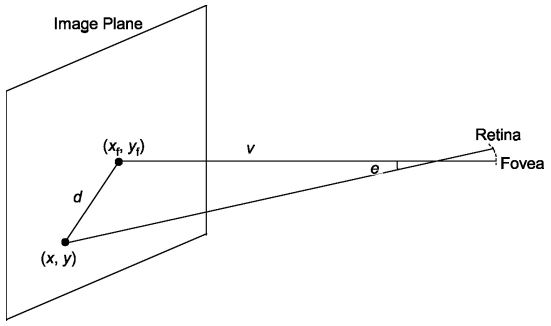


Fig. 2. View geometry.

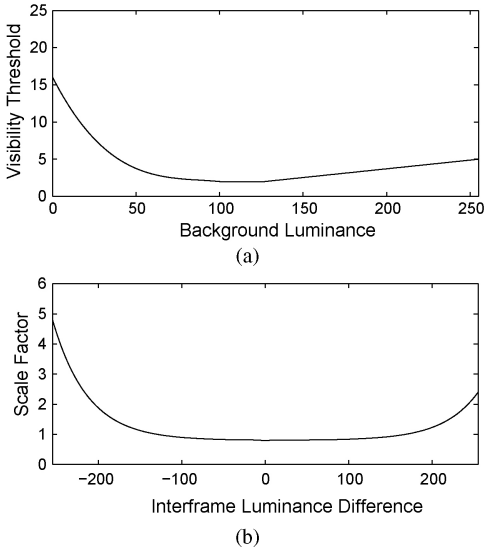


Fig. 3. (a) Spatial JND as a function of background luminance. (b) Temporal JND as a scale factor of spatial JND defined as a function of inter-frame luminance difference.

have been developed. In [23], [25], the JND models were built in spatial (pixel) domain. To incorporate the CSF into the JND model, some other models were proposed in sub-band, discrete cosine transform (DCT), and wavelet domains [12], [13], [49], [50]. In this paper, we have used the spatial JND model developed in [23]. Chou and Li [23] have found that the spatial JND threshold could be modelled as a function of luminance contrast and spatial masking

$$\mathrm{SJND}(x, y) = \max\{f_1(bg(x, y), mg(x, y)), f_2(bg(x, y))\} \quad (2)$$

where $f_1(bg(x, y), mg(x, y))$ and $f_2(bg(x, y))$ are functions to estimate the spatial masking and luminance contrast, respectively. The quantity $f_1(bg(x, y), mg(x, y))$ is defined as

$$f_1(bg(x, y), mg(x, y)) = mg(x, y) \times \alpha(bg(x, y)) + \beta(bg(x, y)) \quad (3)$$

where $mg(x, y)$ is the maximum weighted average of luminance differences derived by calculating the weighted average of luminance changes around the pixel (x, y) in four directions as

$$mg(x, y) = \max_{k=1,2,3,4} \left\{ |\mathrm{grad}_k(x, y)| \right\} \quad (4)$$

where

$$\mathrm{grad}_k(x, y) = \frac{1}{16} \sum_{i=1}^{5} \sum_{j=1}^{5} p(x-3+i, y-3+i) \times G_k(i, j). \quad (5)$$

The operators G_k are defined in Fig. 4. The quantities $\alpha(bg(x, y))$ and $\beta(bg(x, y))$ depend on the background luminance and specify the linear relationship between the visibility threshold and the luminance difference [or luminance contrast around the point of coordinate (x, y)], hence model the spatial masking [9]. These quantities can be expressed as

$$\alpha(bg(x, y)) = bg(x, y) \times 0.0001 + 0.115 \quad (6)$$

and

$$\beta(bg(x, y)) = \mu - bg(x, y) \times 0.01 \quad (7)$$

where μ is the slope of the function at higher background luminance level, and $bg(x, y)$ is the average background luminance calculated by a weighted low-pass filter B (see Fig. 5), as

$$bg(x, y) = \frac{1}{32} \sum_{i=1}^{5} \sum_{j=1}^{5} p(x-3+i, y-3+i) \times B(i, j). \quad (8)$$

The function $f_2(bg(x, y))$ computes the visibility threshold from the luminance contrast as

$$f_2(bg(x, y)) = \begin{cases} T_0 \times \left(1 - \left(\frac{bg(x,y)}{127}\right)^{1/2}\right) + \epsilon, & bg(x, y) \le 127 \\ \gamma \times (bg(x, y) - 127) + \epsilon, & bg(x, y) > 127 \end{cases} \quad (9)$$

where T_0 is the visibility threshold when the background luminance level is 0 and ϵ denotes the minimum visibility threshold. This function shows that the visibility threshold is a root relationship for low-background luminance and is a linear relationship for higher background luminance. To quantify the function parameters, we have conducted the experiments recommended in [23]. In this paper, we have assumed the view distance v to be three times the picture width. In the experiments, a 16×16 square has been located in the center of an image with constant gray level $\mathcal{G}$. For the image at each possible gray level $\mathcal{G}$, noise of fixed amplitude $\mathcal{A}$ has been randomly added or subtracted to each pixel in the square. Therefore, the amplitude of the pixel in the square area was either $\mathcal{G} + \mathcal{A}$ or $\mathcal{G} - \mathcal{A}$ and bounded by the maximum and minimum luminance values. The amplitude of the noise has been adjusted from 0 and increased by 1 until the noise was becoming noticeable. By doing so, the visibility threshold for each background luminance level has been determined. In this paper, the T_0, γ, μ, and ϵ have been set to 14, 3/128, 1/4 and 2, respectively [Fig. 3(a)].

B. Temporal JND Model

In addition to the spatial masking effect, the temporal masking effect should also be considered to build the STJND model for video signal. Usually, larger inter-frame luminance

0	0	0	0	0
1	3	8	3	1
0	0	0	0	0
−1	−3	−8	−3	−1
0	0	0	0	0

(a)

0	0	1	0	0
0	8	3	0	0
1	3	0	−3	−1
0	0	−3	−8	0
0	0	−1	0	0

(b)

0	0	1	0	0
0	0	3	8	0
−1	−3	0	3	1
0	−8	−3	0	0
0	0	−1	0	0

(c)

0	1	0	−1	0
0	3	0	−3	0
0	8	0	−8	0
0	3	0	−3	0
0	1	0	−1	0

(d)

Fig. 4. Matrix G_k. (a) G_1. (b) G_2. (c) G_3. (d) G_4.

1	1	1	1	1
1	2	2	2	1
1	2	0	2	1
1	2	2	2	1
1	1	1	1	1

Fig. 5. Matrix B.

difference results in larger temporal masking effect. To measure the temporal JND function, we have conducted the same experiments as in [24]. A video sequence at 30 frames/s has been constructed in which a square of luminance level $\mathcal{G}$ moves horizontally over a background of luminance level $\mathcal{B}$. Noise of amplitude $\mathcal{A}$ has been randomly added or subtracted to each pixel in small regions as defined in [24]. The distortion visibility thresholds have been determined as a function of the inter-frame luminance difference and background luminance. Based on the experimental results [Fig. 3(b)], the temporal JND is defined as

$$\mathrm{TJND}(x,y,t) = \begin{cases} \max\left(\tau, \frac{\mathcal{H}}{2}\exp\left(\frac{-0.15}{2\pi}(\Delta(x,y,t)+255)\right)+\tau\right) \\ \qquad \Delta(x,y,t) \le 0 \\ \max\left(\tau, \frac{\mathcal{L}}{2}\exp\left(\frac{-0.15}{2\pi}(255-\Delta(x,y,t))\right)+\tau\right) \\ \qquad \Delta(x,y,t) > 0 \end{cases} \tag{10}$$

where $\mathcal{H} = 8$ and $\mathcal{L} = 3.2$ are model parameters. The value $\tau = 0.8$ is based on the conclusion in [24] stating that the scale factor should be reduced to 0.8 as $\Delta(x,y,t) < 5$, in order to minimize the allowable distortion in stationary regions. The quantity $\Delta(x,y,t)$ denotes the average luminance difference between the frame t and the previous frame $t-1$ and is computed as

$$\Delta(x,y,t) = \frac{p(x,y,t) - p(x,y,t-1) + bg(x,y,t) - bg(x,y,t-1)}{2}.$$

The model shows that larger inter-frame luminance difference results in larger visibility threshold. The inequality $\mathcal{H} > \mathcal{L}$ indicates that high-to-low luminance changes cause more significant temporal masking effect than low-to-high luminance changes. The STJND is then defined as

$$\mathrm{STJND}(x,y,t) = [\mathrm{SJND}(x,y)] \cdot [\mathrm{TJND}(x,y,t)]. \tag{11}$$

The STJND model exploits the HVS visual sensitivity to luminance contrast, as well as the spatial and temporal masking effects. The STJND model provides the visibility threshold of each pixel of an image by assuming that the pixel is projected on the fovea area and is perceived at the highest visual acuity. However, if the pixel is not projected in the fovea region, the visual acuity becomes lower. The STJND model can only provide a local visibility threshold. To measure the global visibility threshold of the whole image, the visibility threshold of the pixel not only depends on the local JND threshold, which is modeled by the STJND model, but also depends on its distance from the nearest fixation point. Therefore, we propose to incorporate a foveation model in the STJND model to build the FJND model.

C. FJND Model

To study the HVS foveation behavior, we first consider the relationship between the visual acuity and the retinal eccentricity. We assume that the position of the fixation point is (x_f, y_f) as shown in Fig. 2. The retinal eccentricity e for any given point (x, y) can be calculated as

$$e = \tan^{-1}\left(\frac{d}{v}\right) \tag{12}$$

where v is the viewing distance, and d is the distance between (x, y) and (x_f, y_f) given by

$$d = \sqrt{(x-x_f)^2 + (y-y_f)^2}. \tag{13}$$

An analytical model has been developed in [51] to measure the contrast sensitivity as a function of eccentricity. The contrast sensitivity $CS(f,e)$ is defined as the reciprocal of the contrast threshold $CT(f,e)$

$$CS(f,e) = \frac{1}{CT(f,e)}. \tag{14}$$

The contrast threshold is defined as

$$CT(f,e) = CT_0 \exp\left(\chi f \frac{e+e_2}{e_2}\right) \tag{15}$$

where f is the spatial frequency (cycles/degree), e is the retinal eccentricity (degree). The parameters CT_0, χ, and e_2 are the minimum contrast threshold, the spatial frequency decay constant, and the half-resolution eccentricity constant, respectively. These quantities are model parameters set to $CT_0 = \frac{1}{64}$, $\chi = 0.106$, and $e_2 = 2.3$. The cutoff frequency f_c can be obtained, by setting CT to 1.0, as

$$f_c(e) = \frac{e_2 \ln\left(\frac{1}{CT_0}\right)}{\chi(e+e_2)} \tag{16}$$

and the corresponding eccentricity e_c is given by

$$e_c = \frac{e_2}{\chi f}\ln\left(\frac{1}{CT_0}\right) - e_2. \tag{17}$$

Wang *et al.* [32] suggested that the display cutoff frequency f_d should be half of the display resolution r, that is

$$f_d(v) = \frac{r}{2} \approx \frac{1}{2} \times \frac{\pi v}{180}. \tag{18}$$

Therefore, the cutoff frequency can be refined as

$$f_m(v,e) = \min\left(f_c(e(v)),\, f_d(v)\right). \tag{19}$$

We build the foveation model based on the cutoff frequency obtained from (19) which is a function of eccentricity. To

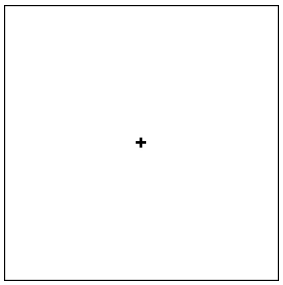


Fig. 6. FJND test.

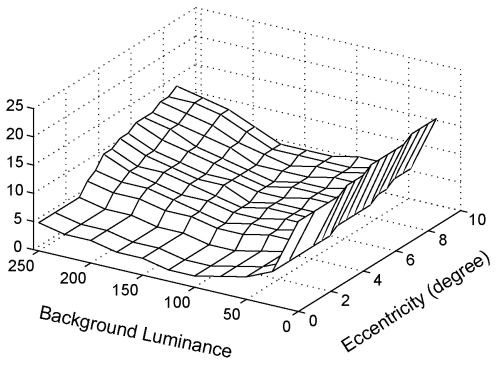


Fig. 7. FJND due to background luminance and eccentricity.

explore the relationship between the JND and eccentricity, we have conducted the experiment in a dark room to obtain the visibility threshold due to the foveation property of the HVS. A small square area, 16×16 pixels, has been randomly displayed on the screen based on predefined eccentricity for the fixation point, located in the center of the picture, and shown as the cross in Fig. 6. For each possible background luminance level $\mathcal{G}$, noise of fixed amplitude $\mathcal{A}$ has been randomly added or subtracted from the pixels within the small square area. Therefore, the amplitude of the pixel in the square area is either $\mathcal{G}+\mathcal{A}$ or $\mathcal{G}-\mathcal{A}$ and bounded by the maximum and minimum luminance values. The amplitude of the noise has been adjusted from 0 and increased by 1 each time. The visibility threshold for each background luminance level and eccentricity has then been determined. The experimental results are shown in Fig. 7. The visibility threshold is not only background-luminance dependent, but also eccentricity dependent. So the foveation model is defined as

$$\mathrm{F}(x, y, v, e) = W_f^{\eta(bg(x,y))}(v, e) \tag{20}$$

where $W_f(v, e)$ is the foveated weighting model defined as

$$W_f(v, e) = 1 + \left(1 - \frac{f_m(v, e)}{f_m(v, 0)}\right)^{\gamma}$$

with $\gamma = 1$ and where $\eta(bg(x, y))$ is a function of the background luminance defined as

$$\eta(bg(x, y)) = 0.5 + \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\log_2(bg(x, y) + 1)) - \mu)^2}{2\sigma^2}\right)$$

with $\mu = 7$ and $\sigma = 0.8$. The foveation model reflects the fact that when the eccentricity increases, the visibility threshold increases accordingly.

After obtaining the foveation model, we define the FJND model by combining the spatial JND, temporal JND, and the foveation model as

$$\mathrm{FJND}(x, y, t, v, e) = [\mathrm{SJND}(x, y)]^{\xi} \cdot [\mathrm{TJND}(x, y, t)]^{\psi} \cdot [\mathrm{F}(x, y, v, e)]^{\zeta} \tag{21}$$

where ξ, ψ and ζ are set to 1s as the temporal JND function $\mathrm{TJND}(x, y, t)$ and the foveation model $\mathrm{F}(x, y, v, e)$ scale the spatial JND model. The STJND model can be obtained by combining SJND and TJND functions. Similarly, the foveated spatial JND (FSJND) model can be obtained by combining the spatial JND model $\mathrm{SJND}(x, y)$ and the foveation model $\mathrm{F}(x, y, v, e)$ so the FSJND model could be used for images.

Since the human fixation points vary from one observer to the other, there might be multiple fixation points in practice [31]. Therefore, we need to adapt the FJND model to multiple fixation points by calculating the foveation model according the possible fixation points $(x_f^1, y_f^1), \ldots, (x_f^K, y_f^K)$. When there are multiple fixation points, we have

$$\mathrm{F}(x, y, v, e) = \min_{i \in \{1, \ldots, K\}} \mathrm{F}^i(x, y, v, e). \tag{22}$$

Since we assume a fixed viewing distance v to calculate $\mathrm{F}(x, y, v, e)$ for each pixel, $\mathrm{F}(x, y, v, e)$ can be calculated by only considering the closest fixation point which results in the smallest eccentricity e and the minimum foveation weight $\mathrm{F}(x, y, v, e)$.

The fixation points can be obtained from Itti's attention model [46]. This model is a bottom-up (stimulus-driven) method to locate the conspicuous visual area. The input visual scene is processed into multiscale low-level feature maps such as color, orientation, and motion. The cross-scale center-surround operation is implemented to simulate the center-surround neurons. Combining the feature maps into the scalar saliency map provides the fixation locations in the visual scene. In this paper, we aim to exploit the contribution of foveation properties of the HVS in the JND model. Therefore, other computational methods which can provide accurate fixation location information can also be employed. As ROI video coding has been widely discussed [40]–[42], [52], we also consider to use skin color detection to define the face as the fixation area. Then we apply the FJND model in ROI video coding based on the ROI-based FJND information.

III. Application of the FJND Model in H.264/AVC Video Coding

This section describes the use of the proposed FJND model for MB quantization adjustment in an H.264/AVC video coder. For each MB, the quantization parameter is optimized based on its FJND information. The Lagrange multiplier in the rate-distortion optimization is adapted so that the MB distortion is lower than the noticeable distortion threshold for the MB.

A. Macroblock Quantization Adjustment

As we have discussed, the visibility of the HVS is space variant. For a video frame, different areas have different

visibility thresholds due to luminance contrast, masking effects, retinal eccentricity, etc. Therefore, different MBs could be represented with different fidelity. Let us consider the distortion measure for each MB given by [53]

$$D = w\frac{Q^2}{\Lambda} \tag{23}$$

where w denotes the distortion weight and Λ is a constant. The distortion weight indicates that the MB can tolerate more or less distortion. Therefore, for a MB with higher distortion weight, a smaller quantization step size should be chosen to reduce the corresponding distortion. In this paper, the distortion weight is defined as a noticeable *perceptual* distortion weight based on the visibility thresholds given by the FJND model. The MB quantizer parameter is adjusted as follows. Let $Q_{\mathrm{r}} = f^{-1}(R(Q))$ denote the reference quantizer determined by the frame-level rate control [54]. The quantizer parameter for the MB of index i is defined as

$$Q_i = \sqrt{\frac{w_{\mathrm{r}}}{w_i}}Q_{\mathrm{r}} \tag{24}$$

where $w_{\mathrm{r}} = 1$ and w_i is defined as a sigmoid function to provide continuous output of the FJND value

$$w_i = a + b\frac{1 + m\exp(-c\frac{s_i - \bar{s}}{\bar{s}})}{1 + n\exp(-c\frac{s_i - \bar{s}}{\bar{s}})} \tag{25}$$

with $a = 0.7$, $b = 0.6$, $m = 0$, $n = 1$, $c = 4$ defined empirically. The quantity s_i is the average FJND of MB i, and $\bar{s}$ is the average FJND of the frame. The noticeable distortion weight w_i is obtained from the FJND information of the MB, by (25). A larger weight w_i indicates that the MB is more sensitive to noise. Such a MB may be perceived at higher visual acuity, e.g., projected on fovea due to attention, or cannot tolerate higher distortion due to low-luminance contrast or masking effects. So a smaller quantization parameter should be used to preserve higher fidelity. When the distance between the MB and the fixation point is large, i.e., the eccentricity is larger, or the MB is less sensitive to noise due to luminance contrast or masking effects, a smaller weight w_i will be obtained and a larger quantization parameter will be used. The gain brought by the proposed FJND model has been assessed using the H.264/AVC joint model (JM). Note that new algorithms have been introduced to determine the frame-level quantization parameter as the reference quantizer [55], [56]. These frame rate control algorithms outperform the JM rate control. However, the proposed method remains compatible with these new algorithms.

B. Rate-Distortion Optimization

The H.264/AVC video coding solution supports flexible block modes, e.g., SKIP, INTER16 × 16, INTER16 × 8, INTER8 × 16, INTER8 × 8, INTRA16 × 16, and INTRA4 × 4 for P-slice. The rate-distortion optimization (RDO) minimizes the Lagrangian cost function for mode selection [57]

$$J_{\mathrm{M}}(\mathrm{M}|Q, \lambda_{\mathrm{M}}) = D_{\mathrm{M}}(\mathrm{M}|Q) + \lambda_{\mathrm{M}} R_{\mathrm{M}}(\mathrm{M}|Q) \tag{26}$$

where D_{M} and R_{M} are the distortion and bit rate for various modes, respectively. M is the set of modes, Q is the quantizer, and λ_{M} is the Lagrange multiplier.

Since $J_{\mathrm{M}}(\mathrm{M}|Q, \lambda_{\mathrm{M}})$ is convex, the minimization of the Lagrangian cost function is given by

$$\frac{\partial J_{\mathrm{M}}(\mathrm{M}|Q, \lambda_{\mathrm{M}})}{\partial R_{\mathrm{M}}(\mathrm{M}|Q)} = \frac{\partial D_{\mathrm{M}}(\mathrm{M}|Q)}{\partial R_{\mathrm{M}}(\mathrm{M}|Q)} + \lambda_{\mathrm{M}} = 0 \tag{27}$$

leading to

$$\lambda_{\mathrm{M}} = -\frac{\partial D_{\mathrm{M}}(\mathrm{M}|Q)}{\partial R_{\mathrm{M}}(\mathrm{M}|Q)}. \tag{28}$$

Based on the noticeable distortion in (23) and conclusion in [57], we have the Lagrange multiplier for the MB i as

$$\lambda_i = 0.85 w_i \times 2^{(Q_i - 12)/3}. \tag{29}$$

So the Lagrange multiplier λ_i for the MB i is a function of the noticeable distortion weight.

IV. Experimental Results

To evaluate the performance of the FJND model, various comparisons and subjective visual tests have been conducted. The subjective visual quality assessment and impairment assessment tests have been performed in a typical laboratory viewing environment with normal lighting. The viewing distance has been set to approximately three times the image width. The display system was a 20 in SGI CRT display with resolution of 800 × 600. The assessors were carefully introduced the assessment methods, the quality or impairment factors such as occur, grading scales, and timing.

A. Validation Tests for FJND

Eleven observers, three females and eight males, have participated in the subjective tests. They were all non-expert with (or corrected-to-)normal visual acuity. The observers were asked to follow the simultaneous double stimulus for continuous evaluation (SDSCE) protocol, as in Rec. ITU-R BT.500 [58], to evaluate the impairment of the right video relative to its reference on the left (Fig. 11). The sequences used in the tests were *Akiyo*, *Stefan*, *Football*, *Bus*, and *Flower*, all in common intermediate format (CIF) format. The left reference video was the original video. The noise associated with the FJND visibility threshold has been randomly added or subtracted on each pixel of the original video frame. The distorted video was displayed on the right side. The observers have been watching two videos at the same time. They were asked to check whether the noise was noticeable, to vote during the mid-gray post-exposure period and record their grade (1–5), as recommended in Fig. 12. The impairment results in Table I indicate that no noticeable distortion was perceived. We have also calculated the average peak signal-to-noise ratios (PSNRs) for the non-FJND model and FJND model. The non-FJND model is the STJND described in the paper. The PSNR gap is up to 1.6 dB. This means that with the foveation properties of the HVS, more distortion can be tolerated. The subjective tests hence demonstrate the usefulness of the FJND model.

The JND model provides cues for signal compression algorithms to match the human perception properties [10] but it traditionally assumes that visual acuity is consistent over the

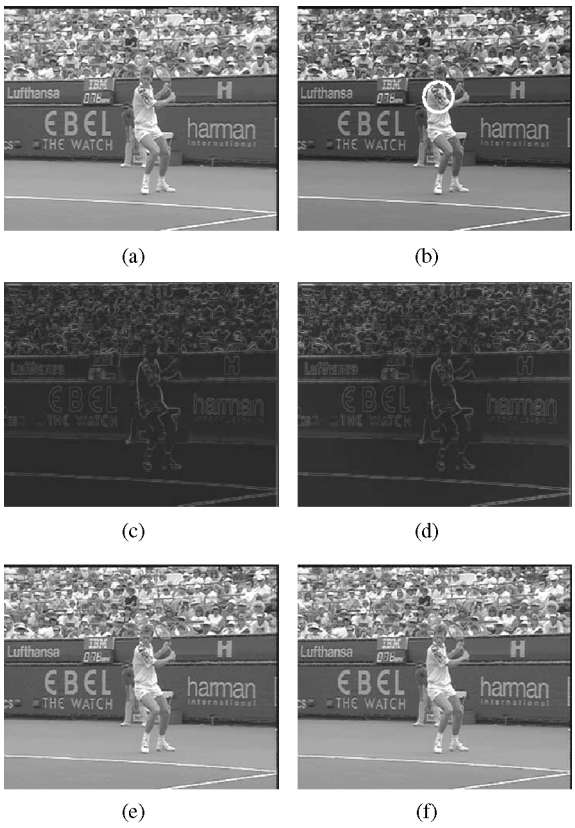


Fig. 8. (a) Original video frame of CIF video sequence *Stefan*. (b) One fixation point on video frame denoted by the circle. (c) Traditional JND map (×4 for display). (d) FJND map (×4 for display). (e) Distorted video frame associated with traditional JND noise (PSNR: 34.78 dB). (f) Distorted video frame associated with FJND noise (PSNR: 32.85 dB).

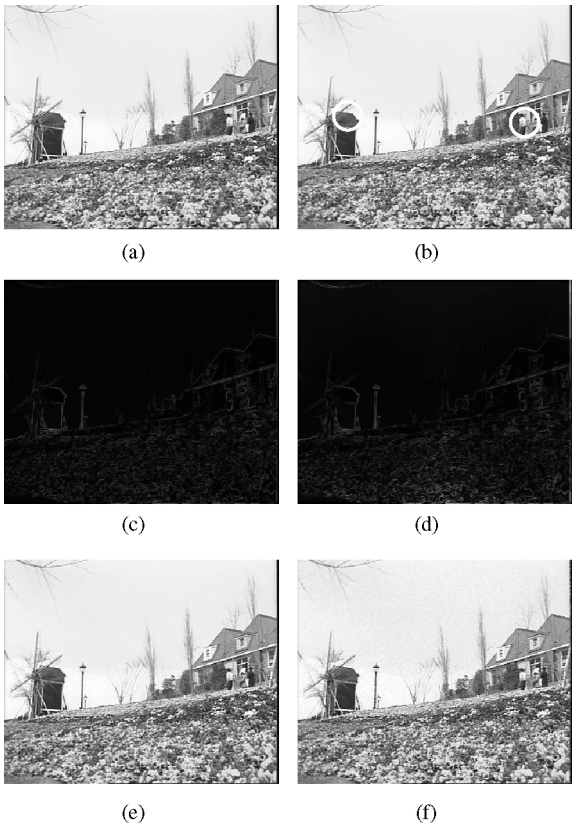


Fig. 9. (a) Original video frame of CIF video sequence *Flower*. (b) One fixation point on video frame denoted by the circle. (c) traditional JND map (×4 for display). (d) FJND map (×4 for display). (e) Distorted video frame associated with traditional JND noise (PSNR: 34.97 dB). (f) Distorted video frame associated with FJND noise (PSNR: 33.10 dB).

entire image. We show an example of the video sequence *Stefan* in Fig. 8. Fig. 8(e) shows the image in which random noise has been added on each pixel. The random noise is not larger than the JND visibility threshold [Fig. 8(c)] so that the distorted picture [Fig. 8(e)] is perceptually lossless. However, according to the HVS property, the perceptual acuity decreases with increased eccentricity. The visibility threshold of the pixel of the image increases when the distance between the pixel and the fixation point increases, as shown in Fig. 8(d). Image regions projected outside the fovea have higher visibility thresholds. Exploiting this so-called *foveation* property should help in further removing *perceptual redundancy*. The distorted picture [Fig. 8(d)] is perceptually lossless when the attention is fixed within the circle area in the picture as shown in Fig. 8(b). The PSNR is decreased from 34.78 dB [Fig. 8(d)] to 32.85 dB Fig. 8(f). More distortion can be tolerated. Fig. 10 shows the distributions of non-FJND and FJND visibility thresholds according to Fig. 8(c) and (d). We show an example of the video sequence *Flower* in Fig. 9 which contains multiple foveation points in one picture. As we can find from Fig. 8(e) and (f), the PSNR is decreased from 34.97 dB to 33.10 dB, which indicates that higher distortion can be tolerated.

B. Comparison With State-of-the-Art JND Models

We also compare the proposed FJND model with two other state-of-the-art JND models. The first one is the pixel-

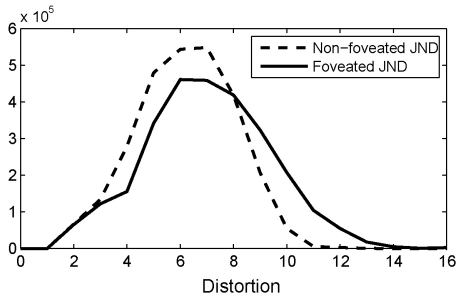


Fig. 10. Distributions of visibility thresholds of non-FJND and FJND.

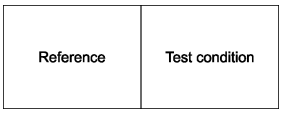


Fig. 11. SDSCE display format.

5	Imperceptible
4	Perceptible, but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Fig. 12. ITU-R impairment scales.

TABLE I
RESULTS OF COMPARISONS AND FJND VALIDATION TESTS

Test Sequence	Average PSNR (dB) of Jia's Method	Average PSNR (dB) of Yang's Method	Average PSNR (dB) for Non-FJND	Average PSNR (dB) for FJND	Mean Impairment Scale for FJND
Akiyo	41.55	35.01	37.16	35.55	5.0
Stefan	35.02	34.71	35.43	33.80	5.0
Football	34.59	36.74	36.17	35.01	5.0
Bus	35.64	32.54	33.70	32.44	5.0
Flower	36.79	37.70	34.78	33.14	5.0

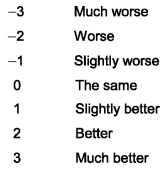


Fig. 13. ITU-R comparison scales.

domain JND model proposed by Yang [25] and the second one is the DCT-domain-based model proposed by Jia [59]. The noise associated with their JND visibility threshold has been randomly added or subtracted on each pixel of the original video frame. The average PSNRs obtained by adding noise according to the different models are shown in Table I. These results show that for most sequences, with the FJND model, one can add more noise (a lower PSNR is then obtained) for the same visual quality. Note that the proposed approach differs from the other methods by the fact that the space-variance properties of the HVS are exploited so that more distortion (lower PSNR and larger MSE) can be tolerated in the image. It has been observed that for the *Football* sequence, the DCT-domain-based method [59] has the lowest PSNR, and for the *Akiyo* sequence, the method in [25] has the lowest PSNR. This is because the JND models used in the methods described in [59] and [25] outperform the STJND model on which is based the proposed foveated model for these two test sequences. However, the foveation method proposed here is also applicable to these two STJND models.

C. Extension to ROI Applications

We have applied the FJND model to video coding with ROI. In both Yang's algorithm [42] and Liu's algorithm [52], the ROI is detected by skin color detection and the quantization parameter is adjusted accordingly. In Yang's algorithm [25], the sensitivity of the ROI is scaled to reflect the importance of the ROI. The perceptual sensitivity weight for MB is then calculated for the distortion model. The sensitivity weight is applied in H.264/AVC quantization adjustment in our comparisons. In Liu's approach [52], the ROI importance is indicated by the skin-tone area and used as weighing factor of MB for quantization determination. However, both of their algorithms do not account for the space-variant property of the HVS and cannot exploit more perceptual redundancies of regions outside ROI. When applying the proposed FJND, the detected ROI is assumed to be the fixation region. The performance of different algorithms has been comparatively assessed. Fig. 14 shows the detected ROI based on skin color for test sequence *Foreman*.

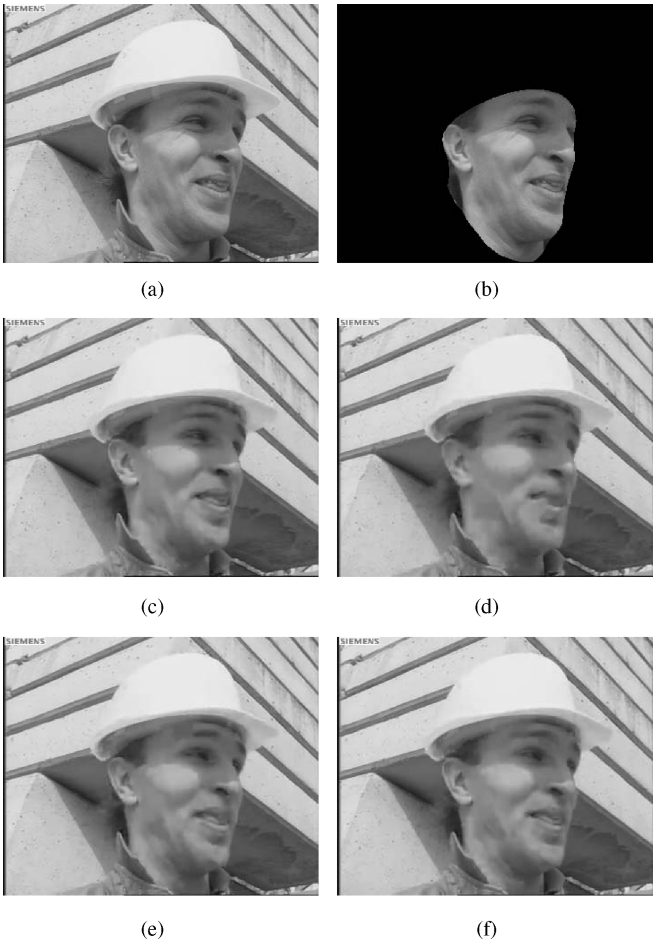


Fig. 14. ROI comparisons of *Foreman*. (a) Original frame. (b) Region of interests. (c) Reconstructed frame from FJND-based method. (d) Reconstructed frame from JM-based method. (e) Reconstructed frame from Liu's method. (f) Reconstructed frame from Yang's method.

The test sequences, *Akiyo*, *Foreman*, *Carphone*, *Silent*, and *Salesman*, of frame rate 30 frames/s, are coded using the different algorithms at a bit rate of 128 kb/s. The H.264/AVC software platform used is the KTA implementation [60]. Each sequence has been coded with IPPP mode (all P frames except the first I frame). The Simplified UMHexagonS with a search range of ± 32, high-complexity RDO, and context-based adaptive binary arithmetic coder (CABAC) coding have been used. The FJND model is used both for tuning the quantization parameter and in the RDO. However, in order to compare the coding performance obtained by using the FJND model with the ones obtained with Yang [42] and Liu's [52] methods,

TABLE II
RESULTS OF COMPARISONS WITH ROI ALGORITHMS

Test Sequence	FPSPNR (dB)					PSNR of ROI (dB)				
	JM	Liu	Yang	FJND I	FJND II	JM	Liu	Yang	FJND I	FJND II
Foreman	34.41	35.52	35.72	35.81	36.19	29.03	31.37	31.78	32.23	32.46
Akiyo	52.90	52.88	52.96	53.22	53.41	39.93	41.32	41.72	41.83	41.86
Carphone	37.99	38.11	38.72	39.35	39.45	34.34	36.32	35.87	36.52	36.59
Silent	44.55	44.65	44.68	44.70	44.83	38.08	39.14	38.97	39.89	39.95
Salesman	44.76	44.79	45.57	45.85	45.90	36.51	37.48	37.84	38.02	38.22

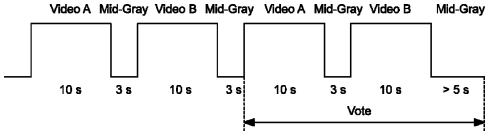


Fig. 15. DSCQS presentation structure.

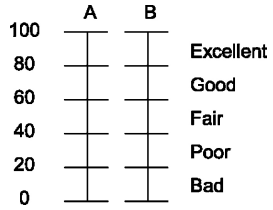


Fig. 16. ITU-R MOS scales.

which only take into account the perceptual model in the quantization parameter adjustment, we have conducted two sets of evaluations. The first one, denoted FJND I in Table II, uses the model in the MB quantization step adjustment only, without modifying the RDO. The second one, denoted FJND II in Table II, uses the model in both MB quantization adjustment and RDO.

Several successful objective metrics have been proposed in the literature [61]–[64] but they do not consider the foveation characteristics of the HVS. Wang *et al.* have proposed the foveation wavelet image quality index in [31]. In this paper, we follow the original idea of the peak signal-to-perceptible-noise ratio (PSPNR) in [24] and directly derive the foveated PSPNR (FPSPNR) based on the proposed FJND model

$$\text{FPSPNR} = 20\log_{10}\frac{255}{\sqrt{E\{[|p(x,y)-p'(x,y)|-\mathcal{F}(x,y,t,v,e)]^2\delta(x,y)\}}}$$

$$\delta(x,y) = \begin{cases} 1, & \text{if } |p(x,y)-p'(x,y)| \geq \mathcal{F}(x,y,t,v,e) \\ 0, & \text{if } |p(x,y)-p'(x,y)| < \mathcal{F}(x,y,t,v,e) \end{cases}$$

where $p(x, y)$ is the original pixel, $p'(x, y)$ is the reconstructed pixel, $\mathcal{F}(x, y, t, v, e)$ denotes the corresponding FJND threshold. As demonstrated in the simulation results (Table II), our approach outperforms the other two ROI approaches by higher FPSPNR and higher PSNR of ROI. The method does not simply use different visibility thresholds in the ROI and outside, but instead exploits the decreasing visual acuity with the eccentricity, i.e., with increasing distance of the MB from the ROI as indicated by our FJND model. Therefore, the saved bits from the MBs which have larger eccentricities from the ROI can be used for coding the ROI. The quality of ROI is improved accordingly. Sample results are presented in Fig. 14.

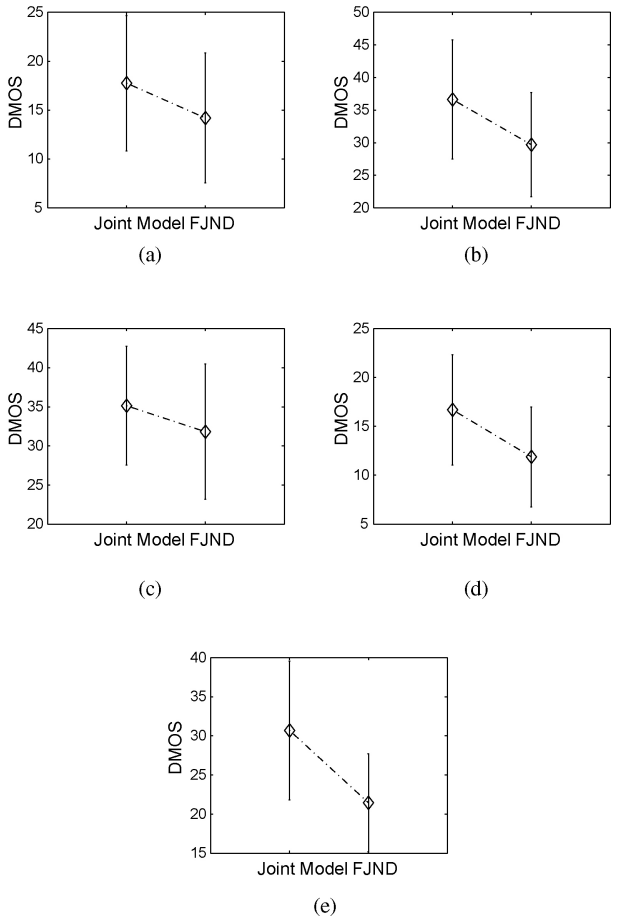


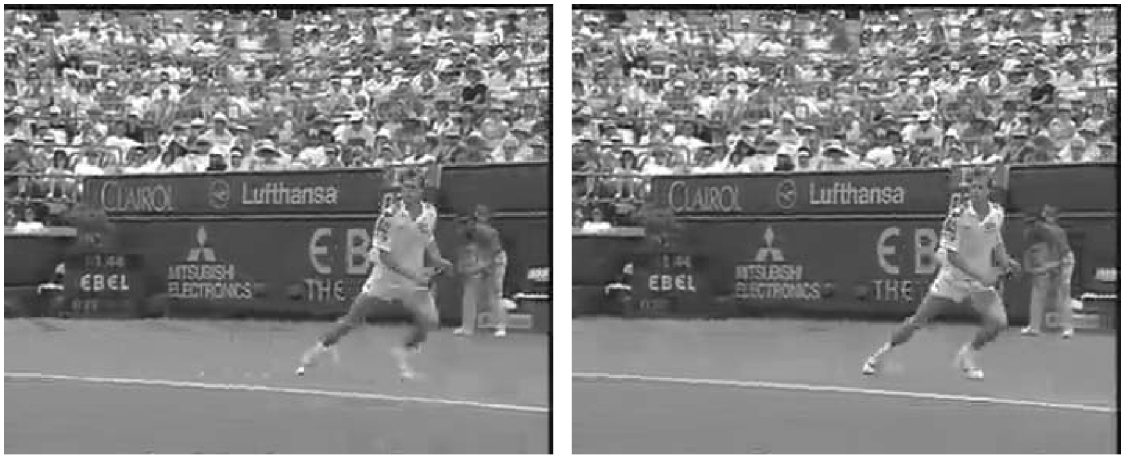
Fig. 17. DMOS comparisons of tests. (a) Test sequence *Akiyo*. (b) Test sequence *Stefan*. (c) Test sequence *Football*. (d) Test sequence *Bus*. (e) Test sequence *Flower*.

D. Subjective Tests for H.264/AVC Coding Applications

Subjective tests have been conducted to evaluate the performance of the proposed approach for H.264/AVC coding applications. The original JM-based method in H.264/AVC is used for comparisons. Two kinds of protocols are used in the subjective tests for coding applications. The first one we used is the SDSCE protocol, as in Rec. ITU-R BT.500 [58] for stimulus comparison. This protocol has been developed in Rec. ITU-R BT.500 to avoid some inadequacies of previous video quality assessment methods. The original video is not needed

TABLE III
RESULTS OF SDSCE TESTS FOR VIDEO CODING

Test Sequence	Joint Model		FJND Method			
	ΔR	PSNR (dB)	ΔR	PSNR (dB)	ΔPSNR (dB)	Mean Comparison Scale
Akiyo	0.2%	37.85	0.2%	37.81	−0.04	0.41
Stefan	0.1%	28.48	0.1%	28.23	−0.25	0.72
Football	0.1%	27.88	0.1%	27.63	−0.25	0.75
Bus	0.1%	27.88	0.1%	27.06	−0.82	0.28
Flower	0.2%	25.24	0.2%	25.28	0.04	1.00



(a) (b)

Fig. 18. Reconstructed video frames of the test sequence *Stefan*. (a) Joint model-based method. (b) FJND-based method.

as the reference in this protocol. The second protocol we used is the double-stimulus continuous quality scale (DSCQS) protocol recommended in Rec. ITU-R BT.500 [58]. It has been widely used in the literature for quality assessment, using the original video sequence as the reference.

The sequences considered in the tests are *Akiyo*, *Stefan*, *Football*, *Bus*, and *Flower* which have been coded at bit rates of 50 kb/s, 300 kb/s, 500 kb/s, 300 kb/s, and 300 kb/s, respectively. The frame rate of these sequences is 30 frames/s. The FJND-based coding method is compared to the original one, i.e., the one used in the H.264 JM implementation [8]. The H.264/AVC software platform adopted is the KTA implementation [60]. Each sequence has been coded with IPPP mode (all P frames except the first I frame). The Simplified UMHexagonS with a search range of ± 32, high-complexity RDO, and CABAC coding have been used. For the test sequences *Football*, *Bus*, and *Flower*, which are shorter than 10 s compositions are made by reversed repetition of the sequence segments to minimize discontinuity at the joints.

1) *SDSCE Tests:* The test follows the SDSCE protocol to evaluate the quality of the right video relative to its reference on the left (Fig. 11). The observers vote during the mid-gray post-exposure period and record their scores according to Fig. 13. Each test video pair is displayed three times. The left and right videos may be exchanged in a pseudo-random fashion during these three presentations whilst the left video is always used as the reference. The first presentation is used to stabilize the observer's opinion and no score should be recorded. They are asked to record their scores for the second and third presentations only. The results are presented in Table III. Compared to the original JM-based method, the PSNRs of the FJND-based method are lower. However, the mean comparison scales from the FJND-based method are positive which means that the observers conclude that the reconstructed videos from the FJND-based method are preferable. Since the reference quantization step Q_r is given by the initial frame-level rate control for both algorithms, the bit rate mismatch, ΔR, is very small.

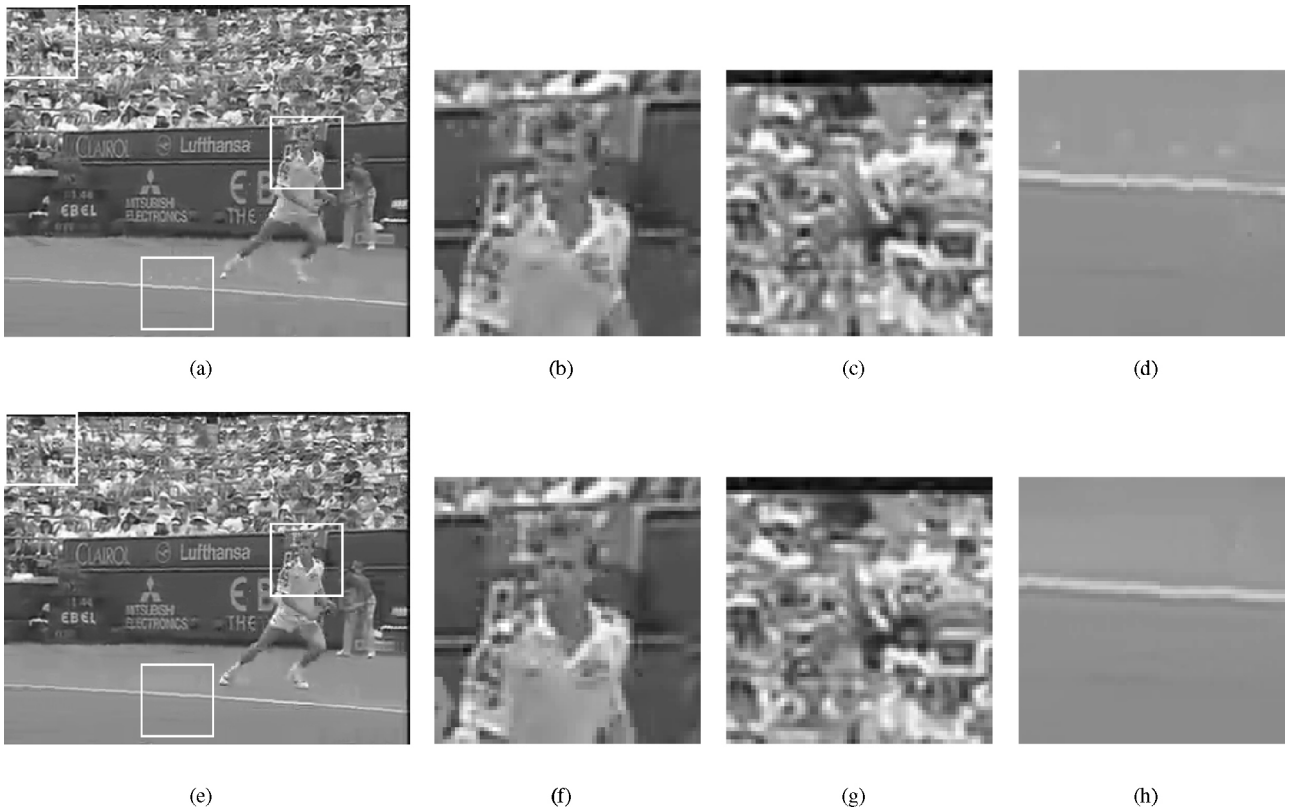
TABLE IV
RESULTS OF DSCQS TESTS FOR VIDEO CODING

Test Sequence	Mean DMOS		
	Joint Model	FJND	ΔDMOS
Akiyo	17.75	14.19	−3.56
Stefan	36.63	29.69	−6.94
Football	35.13	31.81	−3.32
Bus	16.69	11.88	−4.81
Flower	30.69	21.44	−9.25

2) *DSCQS Tests:* The DSCQS protocol has then been used to evaluate the quality of a pair of videos. The procedure is shown in Fig. 15. One stimulus video is the original video sequence, and the other one is the reconstructed video from one of the two coding algorithms. Therefore, ten sequences are randomly presented for five test sequences encoded with two different coding algorithms. The duration between two sequences is 3 s. Each pair of videos is displayed two times. The observers were asked to vote during the voting time as

(a) (b)

Fig. 19. Reconstructed video frames of the test sequence *Flower*. (a) JM-based method. (b) FJND-based method.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 20. Comparisons of regions of the reconstructed frame of the test sequence *Stefan*. (a) Reconstructed frame from the JM-based method. (b) Fixation region in (a). (c) Texture region in (a) away from fixation points. (d) Smooth region in (a) away from fixation points. (e) Reconstructed frame from the FJND-based method. (f) Fixation region in (e). (g) Texture region in (e). (h) Smooth region in (e).

shown in Fig. 15. The mean opinion score (MOS) scales for the DSCQS protocol range from 0 to 100 for the quality from bad to excellent, as shown in Fig. 16.

The difference MOS (DMOS) is then calculated as the difference between the MOSs of the original video and the reconstructed video for each representation. The smaller DMOS indicates that the subjective quality is closer to the original video. Fig. 17 compares the subjective test results. The confidence interval is 95%. As recommended in Rec. ITU-R BT.500 [58], the kurtosis coefficient, β_2, is calculated to determine whether an observer should be rejected. According to the results in the screening process, no observer has been rejected.

Figs. 18 and 19 show sample pictures of videos coded and decoded with the H.264 JM on one hand and with the FJND-based rate control on the other hand. The results obtained with the FJND-based method present better visual quality. Several regions of these two sample images are shown in Fig. 20. Fig. 20(b) shows a fixation region of the reconstructed frame when coded/decoded with the JM-based algorithm. Its visual quality is worse than when coded with the FJND-based method as shown in Fig. 20(f). Since this region is the fixation region, it will be projected on the fovea area and perceived at higher resolution. Therefore, higher bit rate has been given in the FJND-based method and the perceived quality is improved. Fig. 20(c) shows a region away from the fixation region, which has been coded/decoded with the JM-based method. The FJND model establishes that this region can accept higher distortion, since, first it is perceived at lower visual acuity due to its large distance from the fixation point, and second, it can accept higher distortion due to luminance contrast and masking effects. Lower bit rate has thus been allocated to this region but the distortion is imperceptible. Fig. 20(g) shows the result obtained with the FJND-based method. The FJND model also indicates that although the region [Fig. 20(h)] is not a fixation region, it should not be coarsely coded since distortion in smooth regions is easily perceived. Higher distortion in such a region is annoying and degrades the subjective quality, as shown in Fig. 20(d).

V. Conclusion and Discussions

In this paper, a FJND model was proposed. The conventional STJND model is improved by taking into account the foveation properties of the HVS. Since the visual acuity is space variant, the visibility threshold given by the foveation model depends first on eccentricity. However, it has been observed that the visibility threshold is also background luminance dependent, leading to a foveation model defined as a function of both eccentricity and background luminance. The STJND model and the foveation model have then been combined, leading to the FJND model which measures the visibility of the human vision system according to the distance from the fixation point. The subjective tests demonstrate the validity of the FJND model. By applying the proposed FJND model in H.264/AVC video coding system, better perceptual quality of the reconstructed video can be achieved.

The developed FJND model has many potential applications. The application in video coding has been demonstrated in this paper. In addition, the FJND model is useful in interactive video communication, especially for video game and eye-tracking applications. In an interactive video communication system with the FJND information, the most essential information, i.e., the visual information on the fovea, can be delivered to the client with higher priority and fidelity. This is effective when the network bandwidth drops significantly. The proposed FJND model can also be used in other foveated scalable video coding systems such as [32] to achieve perceptually lossless video communication. We can also develop the FJND model in DCT or wavelet domain by using JND and foveation functions in DCT or wavelet domain [26], [31], [33], [59]. This paper focus on spatial domain FJND, so developing FJND model in DCT or wavelet domain is out of the scope of this paper. But it is worth future investigating on these topics.

Based on the JND threshold, the lowest bit rate for representing an image/video with imperceptible distortion (not larger than the JND visibility threshold), correspond to the perceptual entropy [10]. When compared with other JND models, the proposed FJND model leads to higher visibility thresholds and thus leads to lower perceptual entropy. It is also useful to design perceptual coding algorithm which will aim at minimizing the perceptible distortion subject to bit rate constraint. In most image/video compression applications, the coding distortion is much larger than the threshold of visibility. To quantify the perceptible distortion when the distortion is above the visibility threshold is not an easy problem [65], [66]. At low rates, quantization steps need to be tuned to have minimum perceptible distortion [67]. One possible way to use the model at low rates, where the distortion is likely to be above the visibility threshold, would be to use the FJND value at each pixel as a weighting factor to build weighted squared-error metrics [68], [69]. However, we are aware that a simple weighting of the MSE may not lead to the optimal trade-off between visual quality and rate. To have the optimal trade-off, the definition of a series of minimum suprathresholds visual distortion step sizes for the quantization process in the transform domain would probably be more appropriate, but this requires further work.

Acknowledgment

The authors would like to thank members from TEMICS, VISTA, TEXMEX, ASAP, ESPRESSO, BUNRAKU, ARMOR, and SOSSO, in INRIA/IRISA, for their help.

References

[1] *Information Technology: Digital Compression and Coding of Continuous Tone Still Images*, ISO/IEC IS 10918-1 | ITU-T Rec. T.81, 1994.

[2] *Information Technology: JPEG 2000 Image Coding System*, ISO/IEC 15444-1, 2000.

[3] *Video Codec for Audiovisual Services at* $p \times 64$ *kb/s*, ITU-T Rec. H.261, version 1, 1990; version 2, 1993.

[4] *Video Coding for Low-Bitrate Communication*, ITU-T Rec. H.263, version 1, 1995; version 2, 1998; version 3, 2000.

[5] *Coding of Moving Pictures and Associated Audio Fordigital Storage Media at up to About 1.5 Mb/s, Part 2: Visual*, ISO/IEC 11172-2:1991, 1991.

[6] *Information Technology: Generic Coding of Moving Pictures and Associated Audio, Part 2: Visual*, ISO/IEC 13818-2:1994, 1994.

[7] *Information Technology: Coding of Audio/Visual Objects, Part 2: Visual*, ISO/IEC 14496-2:1999, version 1, 1999; version 2, 2000; version 3, 2004.

[8] *Draft ITU-T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264|ISO/IEC 14496-10 AVC)*, document JVT-G050.doc, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, 2003.

[9] A. N. Netravali and B. Prasada, "Adaptive quantization of picture signals using spatial masking," *Proc. IEEE*, vol. 65, no. 4, pp. 536–548, Apr. 1977.

[10] 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.

[11] K. N. Ngan, K. S. Leong, and H. Singh, "Adaptive cosine transform coding of images in perceptual domain," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1743–1749, Nov. 1989.

[12] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Pacific Grove, CA, May 1989, pp. 1945–1948.

[13] 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.

[14] S. S. Hemami, "Visual sensitivity considerations for sub-band coding," in *Proc. Asilomar Conf. Signals, Sytems, Comput.*, Pacific Grove, CA, Nov. 1997.

[15] I. S. Höntsch and L. J. Karam, "Locally adaptive perceptual image coding," *IEEE Trans. Image Process.*, vol. 9, no. 9, pp. 1472–1483, Sep. 2000.

[16] J. Malo, J. Gutierrez, I. Epifanio, F. Ferri, and J. M. Artigas, "Perceptual feedback in multigrid motion estimation using an improved DCT quantization," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1411–1427, Oct. 2001.

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

[18] M. Gaubatz, S. Kwan, B. Chern, D. M. Chandler, and S. S. Hemami, "Spatially-adaptive wavelet image compression via structural masking," in *Proc. IEEE Int. Conf. Image Process.*, Atlanta, GA, Oct. 2005, pp. 1897–1900.

[19] Z. Wang, Q. Li, and X. Shang, "Perceptual image coding based on a maximum of minimal structural similarity criterion," in *Proc. IEEE Int. Conf. Image Process.*, San Antonio, TX, Sep. 2007, pp. 121–124.

[20] C.-W. Tang, C.-H. Chen, Y.-H. Yu, and C.-J. Tsai, "Visual sensitivity guided bit allocation for video coding," *IEEE Trans. Multimedia*, vol. 8, no. 1, pp. 11–18, Feb. 2006.

[21] C.-W. Tang, "Spatial temporal visual considerations for efficient video coding," *IEEE Trans. Multimedia*, vol. 9, no. 2, pp. 231–238, Jan. 2007.

[22] D. H. Kelly, "Motion and vision II: Stabilized spatio-temporal surface," *J. Opt. Soc. Am.*, vol. 69, pp. 1340–1349, Oct. 1979.

[23] C.-H. Chou and Y.-C. Li, "A perceptually tuned sub-band 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.

[24] C.-H. Chou and C.-W. Chen, "A perceptually optimized 3-D sub-band codec for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 2, pp. 143–156, Apr. 1996.

[25] X. Yang, W. Lin, Z. Lu, E. P. 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.

[26] W. Lin, "Computational models for just-noticeable difference," in *Digital Video Image Quality and Perceptual Coding*, H. R. Wu and K. R. Rao, Eds. Boca Raton, FL: CRC Press, 2006.

[27] X. Zhang, W. Lin, and P. Xue, "Just-noticeable difference estimation with pixels in images," *J. Visual Commun. Image Represent.*, vol. 19, pp. 30–41, Jan. 2008.

[28] B. Wandell, *Foundations of Vision*, Sunderland, MA: Sinauer Associates, Inc., 1995.

[29] R. S. Wallace, P.-W. Ong, B. B. Bederson, and E. L. Schwartz, "Spacevariant image processing," *Int. J. Comput. Vis.*, vol. 13, pp. 71–90, Jan. 1994.

[30] T. Kuyel, W. Geisler, and J. Ghosh, "Retinally reconstructed images: Digital images having a resolution match with the human eyes," *IEEE Trans. Syst, Man, Cybern. A*, vol. 29, no. 2, pp. 235–243, Mar. 1999.

[31] Z. Wang and A. C. Bovik, "Embeded foveation image coding," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1397–1410, Oct. 2001.

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

[33] C.-C. Ho, J.-L. Wu, and W.-H. Cheng, "A practical foveation-based rate-shaping mechanism for MPEG videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 11, pp. 1365–1372, Nov. 2005.

[34] Z. Chen, J. Han, and K. N. Ngan, "Dynamic bit allocation for multiple video objects coding," *IEEE Trans. Multimedia*, vol. 8, no. 6, pp. 1117–1124, Dec. 2006.

[35] P. L. Silsbee, A. C. Bovik, and D. Chen, "Visual pattern image sequence coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 3, no. 4, pp. 291–301, Aug. 1993.

[36] E. Nguyen, C. Labit, and J. M. Odobez, "A ROI approach for hybrid image sequence coding," in *Proc. IEEE Int. Conf. Image Process.*, Austin, TX, Nov. 1994, pp. 245–249.

[37] P. Cicconi and H. Nicolas, "Efficient region-based motion estimation and symmetry oriented segmentation for image sequence coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 4, no. 3, pp. 357–364, Jun. 1994.

[38] P. Salembier, L. Torres, F. G. Meyer, and C. Gu, "Region-based video coding using mathematical morphology," *Proc. IEEE*, vol. 83, no. 6, pp. 843–857, Jun. 1995.

[39] H. Sanderson and G. Crebbin, "Image segmentation for compression of images and image sequences," in *Proc. IEE-Vision, Image Signal Process.*, vol. 142. Feb. 1995, pp. 15–21.

[40] X. Yang and K. Ramchandran, "A low-complexity region-based video coder using backward morphological motion field segmentation," *IEEE Trans. Image Process.*, vol. 8, no. 3, pp. 332–345, Mar. 1999.

[41] H. P. Zhang and F. Bossen, "Region-based coding of motion fields for low-bitrate video compression," in *Proc. IEEE Int. Conf. Image Process.*, Singapore, Oct. 2004, pp. 1117–1120.

[42] X. Yang, W. Lin, Z. Lu, X. Lin, S. Rahardja, E. P. Ong, and S. Yao, "Rate control for videophone using local perceptual cues," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 4, pp. 496–507, Apr. 2005.

[43] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[44] L. Itti and C. Koch, "Computational modeling of visual attention," *Nature Rev. Neurosci.*, vol. 2, pp. 194–203, Mar. 2001.

[45] O. Le Meur, P. Le Callet, and D. Barba, "A coherent computational approach to model the bottom-up visual attention," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 5, pp. 802–817, May 2006.

[46] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[47] A. P. Bradley and F. W. M. Stentiford, "Visual attention for region of interest coding in JPEG 2000," *J. Visual Commun. Image Represent.*, vol. 14, pp. 232–250, Sep. 2003.

[48] A. N. Netravali and B. G. Haskell, *Digital Pictures: Representation and Compression*, New York: Plenum, 1988.

[49] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images," *Soc. for Info. Display Dig. Tech. Papers*, vol. XXIV, pp. 946–949, 1993.

[50] J. Lubin, "A visual system discrimination model for imaging system design and evaluation," in *Vision Models for Target Detection and Recognition*, E. Peli, Ed. River Edge, NJ: World Scientific, 1995, pp. 245–283.

[51] W. S. Geisler and J. S. Perry, "A real-time foveated multisolution system for low-bandwidth video communication," in *Proc. SPIE*, vol. 3299. Jul. 1998, pp. 294–305.

[52] Y. Liu, Z. G. Li, and Y. C. Soh, "Region-of-interest-based resource allocation for conversational video communication of H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 18, no. 1, pp. 134–139, Jan. 2008.

[53] J. Ribas-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 172–185, Feb. 1999.

[54] Z. G. Li, W. Gao, F. Pan, S. W. Ma, K. P. Lim, G. N. Feng, X. Lin, S. Rahardja, H. Q. Lu, and Y. Lu, "Adaptive rate control for H.264," *J. Visual Commun. Image Represent.*, vol. 17, pp. 376–406, Apr. 2006.

[55] N. Kamaci, Y. Altunbasak, and R. M. Mersereau, "Frame bit allocation for the H.264/AVC video coder via Cauchy-density-based rate and distortion models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 8, pp. 994–1006, Aug. 2005.

[56] Z. Chen and K. N. Ngan, "Toward rate-distortion tradeoff in real-time color video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 2, pp. 158–167, Feb. 2007.

[57] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 688–703, Jul. 2003.
[58] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R BT.500-11, 2002.
[59] Y. Jia, W. Lin, and A. A. Kassim, "Estimating just-noticeable distortion for video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 7, pp. 820–829, Jul. 2006.
[60] *KTA Software 1.4* [Online]. Available: http://iphome.hhi.de/suehring/tml/download/KTA/
[61] A. B. Watson, J. Hu, and J. F. McGowan, III, "Digital video quality metric based on human vision," *J. Electron. Imaging*, vol. 10, pp. 20–29, Jan. 2001.
[62] A. B. Watson and J. Malo, "Video quality measures based on the standard spatial observer," in *Proc. IEEE Int. Conf. Image Process.*, Rochester, NY, Sep. 2002, pp. III-41–III-44.
[63] Z. Wang, A. C. Bovik, and E. P. Simoncelli, "Objective video quality assessment," in *The Handbook of Video Databases Design and Applications*, B. Furht and O. Marqure, Eds. Boca Raton, FL: CRC Press, Sep. 2003, pp. 1041–1078.
[64] Z. Wang, A. C. 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, pp. 600–612, Apr. 2004.
[65] M. G. Ramos and S. S. Hemami, "Suprathreshold wavelet coefficient quantization in complex stimuli: Psychophysical evaluation and analysis," *J. Opt. Soc. Am. A*, vol. 18, pp. 2385–2397, Oct. 2001.
[66] D. M. Chandler and S. S. Hemami, "Suprathreshold image compression based on contrast allocation and global precedence," in *Proc. 8th Soc. Photo-Optic. Instrum. Eng. Human Vision Electron. Imaging*, 2003, pp. 73–86.
[67] T. N. Pappas, T. A. Michel, and R. O. Hinds, "Supra-threshold perceptual image coding," in *Proc. IEEE Int. Conf. Image Process.*, Lausanne, Switzerland, Sep. 1996, pp. 237–240.
[68] C. F. Harris and J. W. Modestino, "Entropy-constrained sub-band coding of images using a perceptual distortion criteria," in *Proc. IEEE Int. Symp. Informat. Theory*, San Antonio, TX, Jan. 1993, p. 279.
[69] A. B. Watson, "DCT quantization matrics visually optimized for individual images," in *Proc. 4th Human Vision, Visual Process., Digital Display*, Bellingham, WA, 1993.

Zhenzhong Chen (S'02–M'07) received the B.Eng. degree from Huazhong University of Science and Technology, Wuhan, China, and the Ph.D. degree from Chinese University of Hong Kong (CUHK), Shatin, Hong Kong, both in electrical engineering.

He was an ERCIM Fellow with the Institut National de Recherche en Informatique et Automatique, Rennes Cedex, France. He held visiting positions with Universite Catholique de Louvain, Belgium, and Microsoft Research Asia, Beijing, China. He is currently a Lee Kuan Yew Research Fellow with Nanyang Technological University, Singapore. His current research interests include visual signal processing and compression, visual perception, and multimedia communications.

Dr. Chen received the CUHK Faculty Outstanding Ph.D. Thesis Award, the Microsoft Fellowship, and the ERCIM "Alain Bensoussan" Fellowship. He serves as a Voting Member of the IEEE Multimedia Communications Technical Committee, and the Technical Program Committee Member of the GLOBECOM and the CCNC.

Christine Guillemot (SM'04) received the Ph.D. degree from Ecole Nationale Supérieure des Telecommunications, Paris, France.

From 1985 to 1997, she was with France Telecom in the areas of image and video compression for multimedia, and digital television. From 1990 to 1991, she was a Visiting Scientist with Bellcore (Bell Communication Research) in the U.S. She is currently the Director of Research with the Institut National de Recherche en Informatique et Automatique, Rennes Cedex, France. She is the co-inventor of 14 patents. She has co-authored eight book chapters, around 45 international journal publications and more than 120 conference publications. Her current research interests include signal and image processing for image and video compression and communication applications.

Dr. Guillemot was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 2000 to 2004, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY from 2004 to 2006, and the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2007 to 2009. She has been a Member of the IEEE IMDSP and MMSP technical committees.