PERCEPTUALLY-FRIENDLY H.264/AVC VIDEO CODING

Zhenzhong Chen*  
Nanyang Technological University, Singapore  
Christine Guillemot  
INRIA/IRISA, France

ABSTRACT

This paper presents a perceptual H.264/AVC video coding method based on a foveated just-noticeable-distortion (JND) model. Since the perceptual acuity decreases with the increased eccentricity, a foveation model is adopted to further explore the perceptual redundancy in addition to the spatial and temporal just-noticeable-distortion models. Bit allocation and rate-distortion optimization algorithms based on the foveated model are addressed. The performance of the foveated JND model is assessed with subjective visual tests. Applying the proposed model in H.264/AVC video coding can achieve better visual quality.

Index Terms— Just-noticeable distortion, foveation model, bit allocation, rate-distortion optimization, H.264/AVC.

1. INTRODUCTION

Just-noticeable distortion (JND) provides cues for signal compression algorithms to match the human perception properties [1]. The JND model generally explores the visibility of the minimally perceptible distortion which assumes that the visual acuity is consistent over the image. However, the human visual system 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. When the visual stimulus is projected on the retina, the fixation region projected on the fovea is perceived at the highest resolution. The resolution decreases with increased distance, or eccentricity, from the fixation point \((x_f, y_f)\).

Since the perceptual acuity decreases with increased eccentricity, the visibility threshold of the pixel of the image increases with increased distance from the fixation point. In this paper, a foveated just-noticeable-distortion (FJND) model is presented to exploit the perceptual redundancy in the image. Based on the FJND model, the noise associated with the FJND visibility threshold is randomly added or subtracted on each pixel. The distorted picture (Fig. 1(d)) is perceptually lossless when the fixation point is within the circle in the picture (Fig. 1(b)). When compared to the distorted picture obtained based on the traditional JND model (Fig. 1(c)), the PSNR is decreased from 34.78 dB to 32.85 dB (Fig. 1(d)). More distortion can be tolerated.

This paper extends [2] by FJND based region of interests (ROI) coding and comparisons of conventional JND algorithms. The remainder of this paper is organized as follows. The FJND model is introduced in Section 2. The application of FJND in H.264/AVC is presented in Section 3. The experimental results and the conclusion are given in Sections 4 and 5, respectively.

*This work was carried out during the tenure of an ERCIM “Alain Ben-rousset” Fellowship Programme at INRIA/IRISA, France.

2. FOVEATED JUST-noticeABLE-DISTORTION MODEL

We define the FJND model as a combination of spatial JND (SJND), temporal JND (TJND), and foveation model:

\[
\text{FJND}(x, y, t, v, e) = f(\text{SJND}(x, y, t), \text{TJND}(x, y, t), F(x, y, v, e))
\]

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

2.1. Spatial Just-Noticeable-Distortion Model

The perceptual redundancy in spatial domain is mainly based on the visibility of the minimally perceptible distortion which assumes that the visual acuity is consistent over the image. Usually larger inter-frame difference results in larger temporal masking effect. The temporal JND is defined as:

\[
\text{TJND}(x, y, t, \mu, \gamma, \nu, \epsilon) = \max\{ f_1(bg(x, y), mg(x, y)), f_2(bg(x, y)) \}
\]

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. They are defined as:

\[
f_1(bg(x, y), mg(x, y)) = mg(x, y) \times \alpha(bg(x, y)) + \beta(bg(x, y))
\]

\[
f_2(bg(x, y)) = \left\{ \begin{array}{ll}
T_0 \times \left(1 - \frac{\beta(bg(x, y))}{127}\right)^{1/2} + \epsilon & \text{bg}(x, y) \leq 127 \\
\gamma \times (bg(x, y) - 127) + \epsilon & \text{bg}(x, y) > 127
\end{array} \right.
\]

with \(\alpha(bg(x, y)) = bg(x, y) \times 0.0001 + 0.115\) and \(\beta(bg(x, y)) = \mu - bg(x, y) \times 0.01\) where \(bg(x, y)\) is the average background luminance and \(mg(x, y)\) is the maximum weighted average of luminance differences [3]. In this paper, we assume the view distance \(v\) to be 3 times of the picture width and conduct the same experimental tests as in [3] to obtain the model parameters. The \(T_0, \gamma, \mu, \text{and} \ \epsilon\) are defined as 14, 3/128, 1/4 and 2, respectively.

2.2. Temporal Just-Noticeable-Distortion Model

Usually larger inter-frame difference results in larger temporal masking effect. The temporal JND is defined as:

\[
\text{TJND}(x, y, t, \tau) = \max\{ \tau, \frac{\alpha}{7} \exp\left(-\frac{0.15}{\tau}(\Delta(x, y, t) + 255) + \tau\right) \}
\]

\[
\Delta(x, y, t) \leq 0
\]

\[
\max\{ \tau, \frac{\alpha}{7} \exp\left(-\frac{0.15}{\tau}(255 - \Delta(x, y, t)) + \tau\right) \}
\]

\[
\Delta(x, y, t) > 0
\]

\
\Delta(x, y, t) = \frac{p(x, y, t) - p(x, y, t - 1) + bg(x, y, t) - bg(x, y, t - 1)}{2}
\]
where \( p(x, y, t) \) is luminance value of pixel \((x, y)\). This analytical model is developed based on the experimental tests described in [4].

### 2.3. Foveation Model

We first consider the relationship between the visual acuity and the retinal eccentricity. An analytical model has been developed in [5] to measure the contrast sensitivity as a function of eccentricity where the contrast sensitivity \( CS(f, e) \) is defined as the reciprocal of the contrast threshold \( CT(f, e) \):

\[
CT(f, e) = CT_0 \exp \left( \frac{f e + \varepsilon_2}{\varepsilon_2} \right),
\]

(6)

where \( f \) is the spatial frequency (cycles/degree), \( e \) is the retinal eccentricity (degree), \( CT_0 = \frac{1}{\varepsilon_2} \) is the minimum contrast threshold, \( \varepsilon = 0.106 \) is the spatial frequency decay constant, and \( \varepsilon_2 = 2.3 \) is the half-resolution eccentricity constant.

Wang et al. [6] suggested that the display cutoff frequency \( f_d \) should be half of the display resolution \( v \):

\[
f_d(v) = \frac{v}{2} \approx \frac{1}{2} \times \frac{\pi v}{180}
\]

(8)

Therefore, the cutoff frequency is refined as

\[
f_{m}(v, e) = \min (f_{d}(v), f_{d}(v))
\]

(9)

Based on the experiments [2], the foveation model is defined as

\[
F(x, y, v, e) = W_{f}^{\eta(bg(x,y))}(v, e)
\]

(10)

where \( W_{f}(v, e) \) is the foveated weighting model defined as:

\[
W_{f}(v, e) = 1 + \left( \frac{1 - f_{m}(v, e)}{f_{m}(v, 0)} \right)^{\gamma}
\]

(11)

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

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

with \( a = 7 \) and \( \sigma = 0.8 \). The foveation model reflects that when the eccentricity increases, the visibility threshold increases accordingly.

### 2.4. Foveated Just-Noticeable-Distortion Model

The foveated JND is defined as:

\[
FJND(x, y, t, v, e) = [SJND(x, y)]^{\xi} \times [TJND(x, y, t)]^{\psi} \times [F(x, y, v, e)]^{\eta}
\]

(12)

where we set \( \xi = \psi = \eta = 1 \).

When there are multiple fixation points, we have

\[
F(x, y, v, e) = \min_{i = 1, \ldots, k} F^{i}(x, y, v, e)
\]

(13)

Since we will assume a fixed \( v \) to calculate \( F(x, y, v, e) \) for each pixel, \( F(x, y, v, e) \) can be calculated by only considering the closest fixation point which results in the smallest \( e \) and the minimum \( F(x, y, v, e) \). The fixation points can be obtained from Itti’s attention model [7]. As region of interests (ROI) video coding has been widely discussed, 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.

### 3. APPLICATIONS IN H.264/AVC VIDEO CODING

We use the FJND model for H.264/AVC video coding. In this paper, we consider the macroblock quantization adjustment and rate-distortion optimization according to the FJND model. We use the distortion measure for the macroblock given by [8]:

\[
D = w \frac{Q^2}{\Lambda}
\]

(14)

where \( w \) denotes the noticeable distortion weight and \( \Lambda \) is a constant. We assume equal noticeable distortion for MBs in one video frame, e.g., \( D_i = D_j \) where \( D_i \) and \( D_j \) are the noticeable distortion for MB \( i \) and \( j \), respectively. Let \( Q_i = F^{-1}(R(Q)) \) denotes the reference quantizer determined by the frame-level rate control, we have

\[
Q_i = \frac{w_i}{w_i + Q_r}
\]

(15)

where we define \( w_i \) as

\[
w_i = a + b \frac{1 + m \exp(-c \frac{s_i - s}{\sigma_i})}{1 + n \exp(-c \frac{s_i - s}{\sigma_i})}
\]

with \( a = 0.7, b = 0.6, m = 0, n = 1, c = 4, s_i \) is the average FJND of MB \( i \), and \( s \) is the average FJND of the frame. As H.264/AVC supports flexible block modes, rate-distortion optimization (RDO) minimizes the Lagrangian cost for mode selection:

\[
J_{M}(M|Q, \lambda_{M}) = D_{M}(M|Q) + \lambda_{M} R_{M}(M|Q)
\]

(16)
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 $\lambda_M$ is the Lagrange multiplier.

Since $J_M(M|Q, \lambda_M)$ is convex, the minimization of the Lagrangian cost function. Based on the noticeable distortion in 3 and the conclusion in H.264/AVC [9], we have the Lagrange multiplier for the macroblock $i$ as:

$$\lambda_i = 0.85\mu_i \times 2^{(Q-12)/3} \quad (17)$$

### 4. EXPERIMENTAL RESULTS

To evaluate the performance of the foveated just-noticeable-distortion model, subjective visual tests are conducted. The subjective visual quality assessment and impairment assessment tests are performed in a typical laboratory viewing environment with normal lighting. The display system is a 20" SGI CRT display with resolution of $800 \times 600$. The viewing distance is approximately 3 times of the image width. The test sequences are Akiyo, Stefan, Football, Bus, and Flower, all in CIF format.

#### 4.1. Comparison with State-of-the-Art JND Models

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 noise associated with their JND or FJND visibility threshold has been randomly added or subtracted on each pixel of the original video frame. The observers were asked to follow the simultaneous double stimulus for continuous evaluation (SDSCE) protocol, as in Rec. ITU-R BT.500 [10], to evaluate the impairment of the right video relative to its reference on the left. The impairment results in Table 1 indicate that no noticeable distortion was perceived. We have also calculated the average PSNRs for the non-foveated JND model and foveated JND model. The non-foveated JND model is the spatio-temporal JND (STJND) as 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.

We also compare the proposed FJND model with other two state-of-the-art JND models. The first one is the pixel-domain JND model proposed by Yang [11] and the second one is the DCT-domain based model proposed by Jia [12]. The average PSNRs obtained by adding noise according to the different models are shown in Table 1. 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.

| Table 1. 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-foveated JND | Average PSNR (dB) for Foveated JND | Mean impairment scale for Foveated JND |
| 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         |

#### 4.2. Subjective tests for H.264/AVC coding applications

To evaluate the performance of the proposed FJND model for H.264/AVC coding, the test sequences are coded at bit rates 50 kbps, 300 kbps, 500 kbps, 300 kbps, and 300 kbps, for Akiyo, Stefan, Football, Bus, and Flower, respectively. The frame rate is 30 fps. The FJND based coding method is compared to the original one, i.e., the joint model [13]. The H.264/AVC software platform adopted is the KTA implementation [14].

The algorithm is performed with the frame level rate control as the same as that of joint model for fair comparison. The bit rate mismatch of each test is no larger than 0.2% such that the mismatch is negligible. The average PSNR degradation is 0.28 dB. However, PSNR is not reliable as it does not match HVS very well. We use the Double Stimulus Continuous Quality Scale (DSCQS) protocol [10] which has been widely used for quality assessment. The mean opinion score (MOS) scales for the DSCQS protocol range from 0 to 100 for the quality from bad to excellent. Difference mean option score (DMOS) is 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 of the reconstructed video is closer to the original video. The subjective evaluation demonstrates that the quality of the reconstructed video is closer to the original video and is more pleasant to clients. Our FJND based solution achieves better visual quality.

Several portions from the reconstructed video of Stefan are presented in Fig. 2. Figs. 2(b) and 2(f) show a fixation region. The quality of the reconstructed portion from the FJND based coding method (Fig. 2(f)) is better than the result (Fig. 2(b)) from the joint model based coding method. Since FJND indicates this region is a
4.3. Region-of-Interests (ROI) Applications

We have applied the FJND model to regions-of-interests (ROI) video coding and compared with Yang’s algorithm [15] and Liu’s algorithm [16]. In these comparisons, the ROI is detected by skin color detection. When applying the proposed FJND, the detected ROI is assumed to be the fixation region. The performance of different algorithms has been comparatively assessed. The test sequences, Akiyo, Foreman, Carphone, and Silent, of frame rate 30 fps, were coded using the different algorithms at a bit rate of 128 kbps.

In this paper, we adopt PSPNR (peak signal-to-perceptible-noise ratio) in [4] to evaluate the performance. We directly derive the PSPNR based on the proposed FJND model. As shown in Table 3, our approach outperforms other two ROI approaches by higher PSPNR. Our method exploits the decreasing visual acuity with the eccentricity outside ROI, i.e. with increasing distance of the MB from the ROI as indicated by our FJND model. The saved bits from the MBs which have larger eccentricities from the ROI can be used for coding the ROI. Therefore, the quality of ROI is improved.

5. CONCLUSIONS

In this paper, a foveated JND model is introduced. The usefulness of the FJND model is demonstrated and its applications in H.264/AVC video coding is discussed. With the FJND profile, better subjective quality of the coded video can be achieved.

6. REFERENCES


