Scientific
Research
Publishing

A Perceptual Video Coding Based on JND Model

Qingming Yi, Wenhui Fan, Min Shi

College of Information Science and Technology, Jinan University, Guangdong, China
Email: tyqm@jnu.edu.cn, hbyc_fwh@163.com, shimin@jnu.edu.cn

Abstract

In view of the fact that the current high efficiency video coding standard does not consider the characteristics of human vision, this paper proposes a perceptual video coding algorithm based on the just noticeable distortion model (JND). The adjusted JND model is combined into the transformation quantization process in high efficiency video coding (HEVC) to remove more visual redundancy and maintain compatibility. First of all, we design the JND model based on pixel domain and transform domain respectively, and the pixel domain model can give the JND threshold more intuitively on the pixel. The transform domain model introduces the contrast sensitive function into the model, making the threshold estimation more precise. Secondly, the proposed JND model is embedded in the HEVC video coding framework. For the transformation skip mode (TSM) in HEVC, we adopt the existing pixel domain called nonlinear additively model (NAMM). For the non-transformation skip mode (non-TSM) in HEVC, we use transform domain JND model to further reduce visual redundancy. The simulation results show that in the case of the same visual subjective quality, the algorithm can save more bitrates.

Keywords

High Efficiency Video Coding, Just Noticeable Distortion, Nonlinear Additively Model, Contrast Sensitivity Function

1. Introduction

Nowadays, high definition video is becoming more and more popular. However, the growth of storage capacity and network bandwidth cannot meet the demands for high resolution for storage and transmission. Therefore, ITU-T and ISO/IEC worked together to release a new generation of efficient video coding standard—HEVC [1]. HEVC still follows the traditional hybrid coding framework and uses statistical correlation to remove space and time redundancy in order to achieve the highest possible compression effect. However, as the ulti-

mate receiver of video, Human Visual System [2] has some visual redundancy due to its own characteristics. In order to get the perceptual redundancy, researchers have done a lot of work, of which the widely accepted model is the just noticeable distortion model. Video encoding based on perceptible distortion is mainly to use the human eye's visual masking mechanism. When the distortion is less than the human sensitivity threshold, the human eye is imperceptible [3]. In recent years, the JND model has received wide attention in the aspects of video image encoding [4] [5], digital watermarking [6], image quality evaluation [7] and so on. At present, several JND models have been proposed: the JND model based on pixel domain and the JND model based on transform domain.

For the JND model based on pixel domain, it usually considers two main factors including luminance adaptive masking and contrast masking effect. C. H. Chou and Y. C. Li [8] proposed the pixel domain JND model for the first time. The lager one of the calculated luminance adaptive masking value and contrast masking effect value was used as the final JND threshold. Yang [9] and others proposed the classical nonlinear additively masking model. The two kinds of masking effects were added together to get the corresponding JND values. To some extent, the interaction between the two masking effects was considered. To solve the problem of lack of precision in the calculation of the contrast masking value for the above methods, Liu [10] assigned different weights to texture region and edge region in the image through texture decomposition on the basis of NAMM model, which made the JND model have better calculation accuracy. Wu [11] proposed a JND model based on luminance adaptive and structural similarity, which further considered the sensitivity of human eyes to different regular and irregular regions when computing texture masking.

The JND model based on transform domain could easily introduce the contrast sensitivity function into the model with high accuracy. Since most image coding standards adopt DCT transform, the JND model based on DCT domain has attracted much attention of researchers. Ahumada *et al.* [12] obtained a JND model of a grayscale image by calculating the spatial CSF function. Based on this, Waston [13] proposed the DCTune method, further considering the features of luminance adaptation and contrast masking. Zhang [14] made the JND model more accurate by adding a luminance adaptive factor and a contrast masking factor. Wei *et al.* [15] introduced gamma correction to the JND model and proposed a more accurate video image JND model.

2. Nonlinear Additively Masking Model

The NAMM model is simulated in pixel domain from the aspects of luminance adaptation and texture masking to obtain the JND threshold of pixel domain. The JND estimation based on the pixel domain can be written as the nonlinear additively of the luminance adaptation and the contrast masking, as shown in Equation (1):

$$\text{JND}_{\text{pixel}}(x, y) = T^l(x, y) + T^t(x, y) - C^{lt} \cdot \min\left\{T^l(x, y), T^t(x, y)\right\} \tag{1}$$

where, $T^l(x,y)$ and $T^t(x,y)$ denote the basic threshold of adaptive background luminance and texture masking; C^{lt} represents the overlapping part of two kinds of effects, and it is used to adjust the two factors. The larger the C^{lt} value is, the stronger superposition between the adaptive background luminance and texture masking is. When C^{lt} is 1, the superposition effect between the two factors is the greatest; when C^{lt} is 0, there is no superposition effect between the two effects. In fact, the superposition is between the maximum and the minimum, where C^{lt} is equal to 0.3.

Figure 1 shows the curve of the background luminance and the visual threshold obtained from the experimental results. It simulates the background luminance model and shows the distortion threshold that the human eye can tolerate under a certain background luminance.

$T^l(x,y)$ can be determined according to the visual threshold curve in Figure 1.

$$T^l(x,y) = \begin{cases} 17\left(1 - \sqrt{\dfrac{\overline{I_Y(x,y)}}{127}} + 3\right), & \overline{I}_Y(x,y) \leq 127 \\ \dfrac{3}{128}\left(\overline{I}_Y(x,y) - 127\right) + 3, & \text{others} \end{cases} \tag{2}$$

where $\overline{I}_Y(x,y)$ is the average background luminance value.

Due to the characteristics of HVS itself, distortion that occurs in plain and edge areas is more noticeable than texture areas. In order to estimate the JND threshold more accurately, it is necessary to distinguish the edge and non-edge regions. Therefore, considering the edge information, the calculation method of the texture masking threshold $T^t(x,y)$ is:

$$T^t(x,y) = \beta G_\theta(x,y) W_\theta(x,y) \tag{3}$$

where β is the control parameter and its value is set as 0.117. $G_\theta(x,y)$ donates the maximal weighted average of gradients around the pixel at (x, y); $W_\theta(x,y)$ is an edge-related weights of the pixel at (x, y), and its corresponding matrix W_θ is detected by the Gaussian low-pass filter.
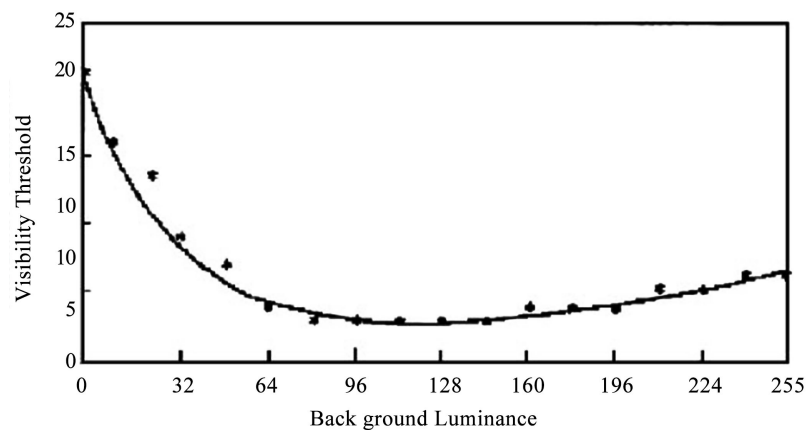

Figure 1. Background luminance and visual threshold.

$G_\theta(x, y)$ is defined as:

$$G_\theta(x, y) = \max_{k=1,2,3,4} \left\{ grad_{\theta,k}(x, y) \right\} \tag{4}$$

with

$$grad_{\theta,k}(x, y) = \frac{1}{16} \sum_{i=1}^{5} \sum_{j=1}^{5} I_\theta(x-3+i, y-3+j) \times g_k(i, j) \tag{5}$$

where, $g_k(i, j)$ are four directional high-pass filters for texture detection, as shown in Figure 2.

3. Improved JND Model Based on DCT Domain

A typical JND model based on DCT domain is expressed as a product of a base threshold and some modulation factors. Assume that t is expressed as the frame index in the video sequence, n is the block index in the tth frame, and (i, j) is the DCT coefficient index. Then the corresponding JND threshold can be expressed as:

$$\text{JND}_{\text{DCT}}(n, i, j, t) = T(n, i, j, t) \times a_{\text{Lum}}(n, t) \times a_{\text{Contrast}}(n, i, j, t) \tag{6}$$

where $T(n, i, j, t)$ is the spatial-temporal base distortion threshold, which is calculated from the spatial-temporal contrast sensitivity function; $a_{\text{Lum}}(n, t)$ denotes the luminance adaptation factor; $a_{\text{Contrast}}(n, i, j, t)$ is expressed as a contrast masking factor.

3.1. Spatial-Temporal Contrast Sensitivity Function

In psychophysics experiments, the visual sensitivity of the human eye is related to the spatial frequency and time frequency of the input signal. The contrast sensitive function is usually used to quantify the relationship between these factors. It is defined as the inverse of the distortion perceived by human eye, when the contrast changes. The spatial-temporal contrast sensitivity function curve is shown in Figure 3. If we consider the (i, j) th in the nth DCT block in the tth frame, then the corresponding CSF function can be written as:

$$G(n, i, j, t) = c_0 \left(k_1 + k_2 \left| \log\left(\varepsilon \cdot v(n, t)/3 \right) \right|^3 \right) \cdot v(n, t) \cdot \left(2\pi \rho_{i,j} \right)^2$$
$$\cdot \exp\left(-2\pi \rho_{i,j} \cdot c_1 \cdot \left(\varepsilon \cdot v(n, t) + 2 \right) / k_3 \right) \tag{7}$$

where $v(n, t)$ depicts the associated retinal image velocity; the empirical constant k_1, k_2 and k_3 are set as 6.1, 7.3 and 23. c_0 and c_1 control the magnitude and the bandwidth of a CSF curve; $\rho_{i,j}$ is the spatial subband frequency:

$$\rho_{i,j} = \frac{1}{2N} \sqrt{\left(i/\varpi_x \right)^2 + \left(j\varpi_y \right)^2} \tag{8}$$

where, ϖ_x and ϖ_y are the horizontal and vertical sizes of a pixel in degrees of visual angle, respectively. They are related to the viewing distance l and the display width Λ of a pixel on the monitor, as follows:

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)

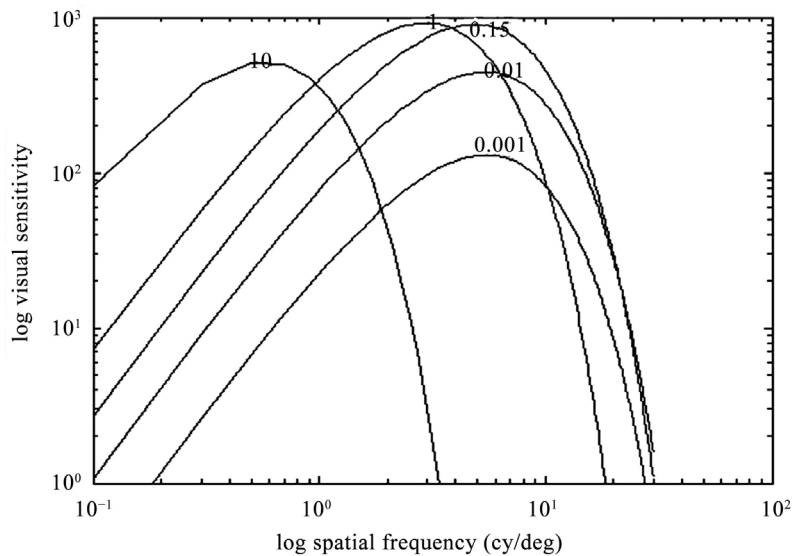**Figure 2.** Directional high-pass filters for texture detection.

Figure 3. Spatial CSF at different retinal velocities.

$$\varpi_h = 2 \cdot \arctan\left(\frac{\Lambda_h}{2 \cdot l}\right), h = x, y \tag{9}$$

when Equation (7) is used for predicting distortion threshold due to spatial-temporal CSF, several factors needs to be considered: 1) the sensitivity modeled by Equation (7) represents the inverse of distortion threshold; 2) the CSF threshold represented in the luminance needs to be scaled into the gray levels for digital image; 3) since Equation (7) comes from experimental data of one-dimensional spatial frequency, for any subband, the threshold is actually higher than the one given by Equation (7), and therefore a compensating needs to be introduced for a DCT sub-band. With all consideration mentioned above, the base threshold for a DCT sub-band is determined as:

$$T(n,i,j,t) = \frac{1}{G(n,i,j,t)} \times \frac{M}{\Phi_i \Phi_j (L_{\max} - L_{\min})} \times \frac{1}{r + (1-r)\cos^2 \theta_{i,j}} \tag{10}$$

where, $L_{\max}$ and $L_{\min}$ represent the display luminance values corresponding to the maximum and minimum gray levels, respectively; M is the number of gray levels, which is generally valued at 256; Φ_i and Φ_j belong to the DCT normalization factor; $\theta_{i,j}$ accounts for the effect of an arbitrary subband; r is set to 0.6.

3.2. Luminance Adaptive Factor and Contrast Masking Factor

The luminance masking mechanism is related to the brightness change in the image. According to Weber-Fechner's law, the minimum perceptible luminance of human eye shows a higher threshold in the areas with brighter or darker background brightness, which is called luminance adaptive effect. The calculation formula of the luminance adaptive factor is:

$$a_{\mathrm{Lum}}(n,t) = \begin{cases} (60-\bar{I})/150 + 1, & \bar{I} \leq 60 \\ 1, & 60 \prec \bar{I} \prec 170 \\ (\bar{I}-170)/425 + 1, & \bar{I} \geq 170 \end{cases} \tag{11}$$

where $\bar{I}$ represents the average brightness.

The contrast masking effect is an important perceptual property in the HVS, usually related to the awareness of a signal in the presence of another signal. When the contrast sensitivity factor is calculated, the image is first detected by Canny edge, and the image blocks are divided into three types: plain, edge and texture region. Since the human eye is more sensitive to distortions that occur in plain areas and in edge areas, different weights need to be assigned to different areas. Based on the above considerations, the weighted factor for each classification block is determined by the following equation:

$$\psi = \begin{cases} 1, & \text{in plain and edge region} \\ 2.25, & \text{in texture region and } (i^2 + j^2) \leq 16 \\ 1.25, & \text{in texture region and } (i^2 + j^2) > 16 \end{cases} \tag{12}$$

where i and j are the DCT coefficient indices.

Taking the masking effect in the intra frame into account, the final contrast masking factor is:

$$a_{\mathrm{contrast}}(n,i,j,t) = \begin{cases} \psi, & \text{in plain and edge region } (i^2 + j^2) \leq 16 \\ \psi \cdot \min\left(4, \max\left(1, \frac{C(n,i,j,t)}{T(n,i,j,t) \cdot a_{\mathrm{Lum}}(n,t)}^{0.36}\right)\right), & \text{others} \end{cases} \tag{13}$$

4. Simulation Results

4.1. Evaluation of the Improved JND Model Based on Transform Domain

In order to verify the effectiveness of our proposed JND model based on DCT domain, we selected eight test images of different contents and complexities as shown in Figure 4 to carry out simulation experiments. Theoretical analysis shows that under a certain visual quality, the larger the threshold of the JND model is, the more visual redundancy will be excavated. Under the same injected noise energy, a more accurate JND model leads to better perceived quality. In order to verify the validity of the model, the thresholds calculated by the corresponding JND models are introduced as noise into the DCT coefficients:
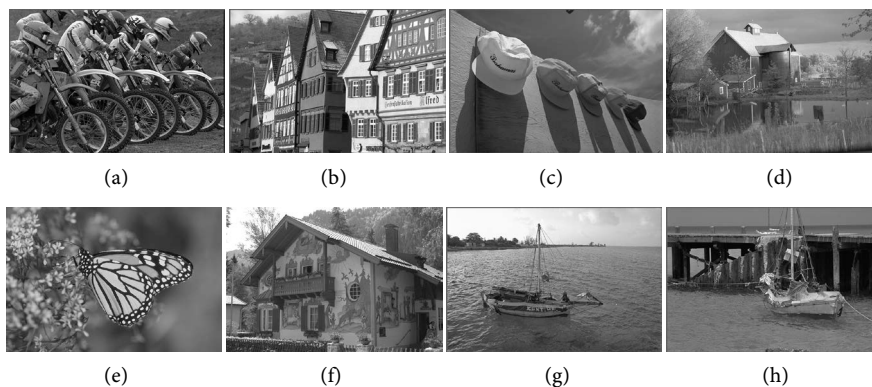
Figure 4. Eight test images. (a) Bikes; (b) Buildings; (c) Caps; (d) House; (e) Monarch; (f) Painted house; (g) Sailing 1; (h) Sailing 4.

$$C_{\text{noise}}\left(n,i,j,t\right) = C\left(n,i,j,t\right) + M_{n,i,j}^{\text{random}} \cdot \text{JND}\left(n,i,j,t\right) \qquad (14)$$

where, $C\left(n,i,j,t\right)$ and $C_{\text{noise}}\left(n,i,j,t\right)$ represent DCT coefficients and DCT coefficients after noise injection; $M_{n,i,j}^{\text{random}}$ random takes +1 and −1.

The JND model presented in this paper is compared with the three models shown in Table 1 respectively. As can be seen from the table, the PSNR measured by this model is the smallest. Under the same visual quality, the smaller the PSNR value of the image is, the greater the energy is introduced into the noise and the larger the corresponding JND threshold is. This means that the larger JND threshold obtained by this model can tolerate more distortion, and the accuracy of the model has been further improved.

4.2. The Overall Performance of the Perceptual Video Coding Scheme

In order to make full use of the JND characteristics of the human visual system to reduce the perceived redundancy of the input video, we integrated the designed JND model into the HEVC coding framework. For the transform skip mode, we chose the existing pixel domain JND model; and the proposed JND model based on DCT domain is utilized for the transform non-skip mode. Figure 5 shows the overall framework of the perceptual video coding scheme.

In order to verify the effectiveness of the algorithm proposed in this paper, the algorithm will be implemented on HM11.0, using the full I-frame encoding configuration environment. The initial quantization parameters are set to 22, 27, 32 and 37, respectively. The test sequences used in the experiment include Kimono, Cactus with a resolution of 1920 × 1080, BQMall, PartyScene with a resolution of 832 × 480, and Basketball Drill Text and China Speed for screen content encoding. We will evaluate the performance of the algorithm in terms of bit number reduction and encoding time. Compared to HM11.0, the bit rate reduction of the perceptual video coding scheme and the encoding time are calculated by the following formula:

Table 1. PSNR between different models.

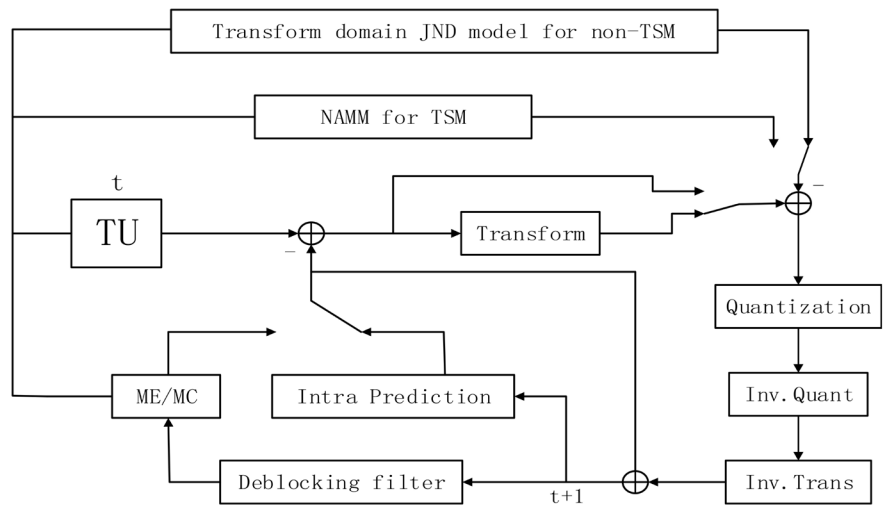| Test Image	PSNR			
	[8]	[14]	[16]	Prop
Bikes	28.73	28.71	27.36	25.85
Buildings	29.65	28.89	28.09	24.87
Caps	31.71	29.70	31.23	26.12
House	31.96	29.09	29.48	25.81
Monarch	31.47	29.80	31.43	26.15
Painted House	30.84	29.08	28.35	25.74
Sailing1	31.62	28.88	29.20	25.44
Sailing4	30.18	29.05	28.75	25.91
Average	30.770	29.150	29.236	25.736

Figure 5. Overall coding framework.

$$\Delta \text{Bitrate} = \frac{\text{Bitrate}_{\text{Pro}} - \text{Bitrate}_{\text{ref}}}{\text{Bitrate}_{\text{Pro}}} \times 100\% \qquad (15)$$

$$\Delta \text{Time} = \frac{\text{Time}_{\text{Pro}} - \text{Time}_{\text{ref}}}{\text{Time}_{\text{Pro}}} \times 100\% \qquad (16)$$

Table 2 shows the comparison of the performance of the proposed algorithm and Chen's [17] and Bae's [4] schemes under different quantization parameters. The experimental results show that compared with the algorithm of [4], the algorithm reduces the encoding bit rate by 4.3%, compared with Chen's algorithm, the encoding bit rate decreases by up to 7.58%.

In order to more intuitively show the bitrate reduction of each algorithm, Figure 6 shows the comparison of bitrates at different QP values. It can be ob-

served that compared with other method, more bit saving can be obtained by our method in most cases. It also can be seen that the smaller the QP value is, the more bits are reduced. This is because finer quantification will result in a larger JND threshold.

Table 2. Comparison of the performance of each program.

| Test Sequence	QP	ΔBitrates (%)			ΔTime (%)		
		PVCChen	PVCBae	PVCProp	PVCChen	PVCBae	PVCProp
Kimono	22	−15.97	−6.81	−2.23	6.31	2.16	22.39
	27	−13.22	−5.12	−12.24	8.42	11.29	23.19
	32	−10.92	−0.91	1.13	9.21	14.26	22.63
	37	−10.11	−3.07	−4.44	8.09	15.69	24.09
Cactus	22	−19.15	−49.10	−44.62	7.38	10.17	18.41
	27	−9.95	−19.67	−9.51	6.58	3.78	20.45
	32	−5.80	−5.47	−5.10	8.56	8.68	19.59
	37	−4.51	−3.90	−1.58	8.18	10.73	19.20
BQMall	22	−15.08	−26.49	−32.00	8.41	3.47	18.29
	27	−9.91	−18.35	−5.72	10.66	11.12	20.09
	32	−6.71	−6.25	−4.65	12.06	16.37	21.26
	37	−4.94	−5.01	−3.06	15.16	18.19	21.79
PartyScene	22	−16.50	−31.00	−36.91	6.19	5.09	16.97
	27	−12.16	−25.03	−6.50	6.86	2.96	20.30
	32	−8.62	−11.91	−16.76	8.55	9.79	22.21
	37	−8.01	−6.66	−5.75	12.47	15.27	23.62
BasketballDrill Text	22	−12.51	−23.18	−48.04	6.43	10.63	19.22
	27	−8.63	−17.79	−38.48	5.08	16.86	21.20
	32	−4.93	−7.24	−30.08	6.42	22.14	21.91
	37	−3.58	−5.14	−24.51	7.35	24.54	22.51
ChinaSpeed	22	−16.51	−32.41	−59.22	5.00	8.73	21.70
	27	−11.46	−23.65	−42.25	6.59	12.63	22.34
	32	−7.61	−15.99	−20.62	7.29	16.00	21.48
	37	−7.41	−9.12	−9.46	8.58	19.14	23.76
Average		−11.69	−14.97	−19.27	8.16	12.07	21.19
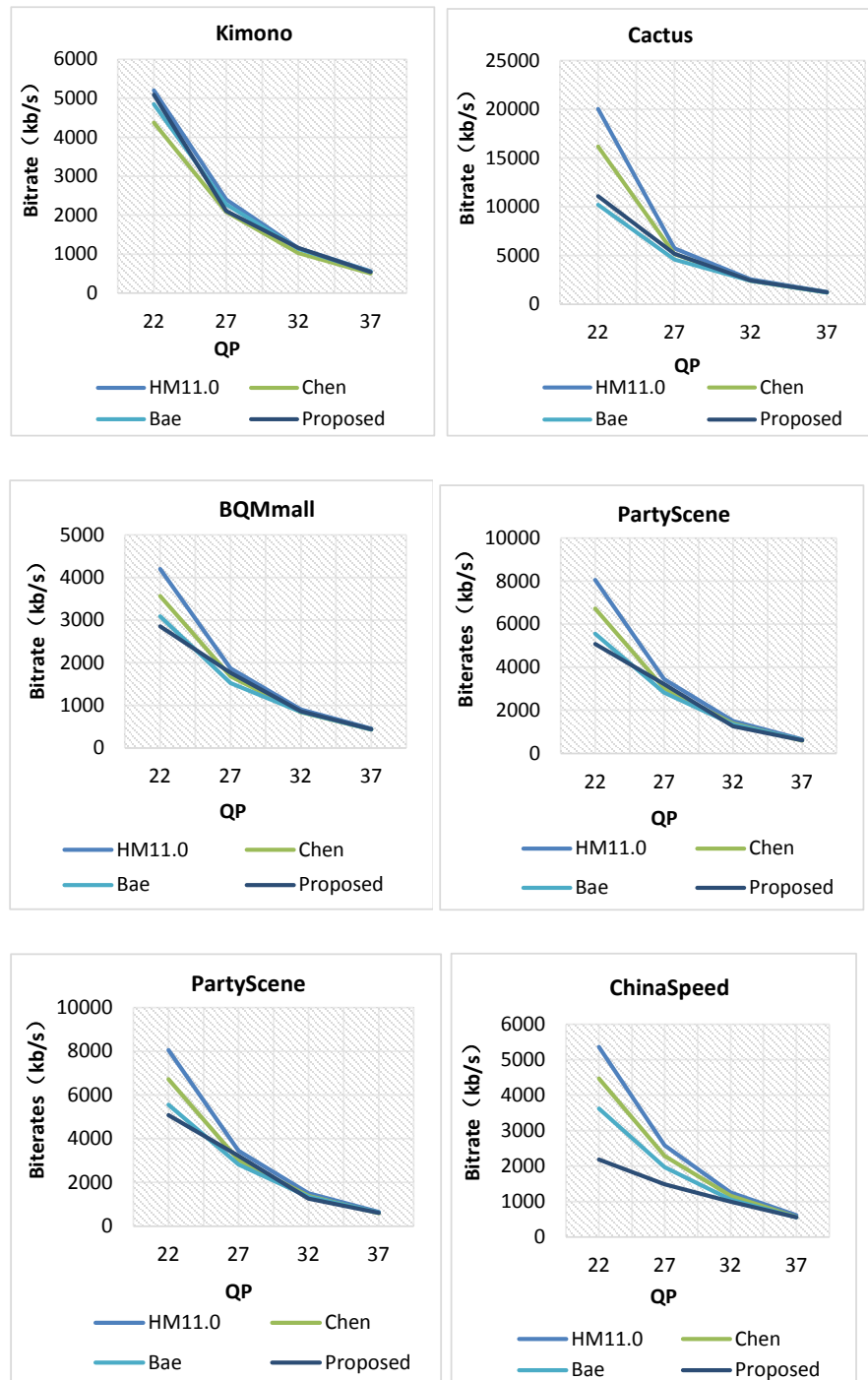
Figure 6. Comparisons of the bitrates.

Acknowledgements

In this paper, we introduce JND model based on pixel domain and JND model based on DCT domain into HEVC framework. Both models have their own advantages: the JND model based on pixel domain can directly give the JND threshold in the pixel domain, and the calculation is easier. The JND model based on the DCT domain integrates the CSF function and the estimated value is more

accurate. Based on the above analysis and combining the advantages of the two models, we choose the pixel domain JND model for the transform skip mode, and choose the more accurate DCT domain JND for the transform non-skip mode. Simulation experimental results show that compared with other models, this algorithm can save up to 7.58% of the coding rate.

References

[1] Sullivan, G.J., Ohm, J., Han, W.J. and Wiegand, T. (2012) Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Transaction on Circuits & Systems for Video Technology*, **21**, 1649-1668. https://doi.org/10.1109/TCSVT.2012.2221191

[2] Wu, H.R. and Rao, K.P. (2005) Digital Video Image Quality and Perceptual Coding. CRC Press, Boca Raton, FL, USA. https://doi.org/10.1201/9781420027822

[3] Jayant, N., Johnston, J. and Safranek, R. (1993) Signal Compression Based on Models of Human Perception. *Proceedings of the IEEE*, **81**, 1385-1422. https://doi.org/10.1109/5.241504

[4] Kim, J., Bae, S.H. and Kim, M. (2015) An HEVC-Compliant Perceptual Video Coding Scheme Based on JND Models for Variable Block-Sized Transform Kernels. *IEEE Transactions on Circuits and Systems for Video Technology*, **25**, 1786-1800. https://doi.org/10.1109/TCSVT.2015.2389491

[5] Ki, S., Bae, S.H., Kim, M. and Ko, H. (2018) Learning-Based Just Noticeable Quantization Distortion Modeling for Perceptual Video Coding. *IEEE Transactions on Image Processing*, **27**, 3178-3193. https://doi.org/10.1109/TIP.2018.2818439

[6] Wan, W., Liu, J., Sun, J., Ge, C. and Nie, X. (2015) Logarithmic STDM Watermarking Using Visual Saliency-Based JND Model. *Electronics Letters*, **51**, 758-760. https://doi.org/10.1049/el.2014.4329

[7] Wang, H. (2016) MCL-JCV: A JND-Based H.264/AVC Video Quality Assessment Dataset. 2016 *IEEE International Conference on Image Processing*, 25-28 September 2016, Phoenix, AZ, 1509-1513. https://doi.org/10.1109/ICIP.2016.7532610

[8] Chou, C.H. and Li, Y.C. (1995) A Perceptual Tuned Sub-Band Image Coder Based on the Measure of Just-Noticeable-Distortion Profile. *IEEE Transaction on Circuits & Systems for Video Technology*, **5**, 467-476. https://doi.org/10.1109/76.475889

[9] Yang, X.K., Ling, W.S., Lu, Z.K., Ong, E.P. and Yao, S.S. (2005) Just Noticeable Distortion Model and Its Applications in Video Coding. *Signal Processing Image Communication*, **20**, 662-680. https://doi.org/10.1016/j.image.2005.04.001

[10] Liu, A., Lin, W., Paul, M., Deng, C. and Zhang, F. (2010) Just Noticeable Difference for Images with Decomposition Model for Separating Edge and Textured Regions. *IEEE Transactions on Circuits & Systems for Video Technology*, **20**, 1648-1652. https://doi.org/10.1109/TCSVT.2010.2087432

[11] Wu, J.J., Qi, F. and Shi, M. (2012) Self-Similarity Based Structural Regularity for Just Noticeable Difference Estimation. *Journal of Visual Communication and Image Representation*, **23**, 845-852. https://doi.org/10.1016/j.jvcir.2012.04.010

[12] Ahumada Jr., A.J. and Peterson, H.A. (1992) Luminance-Model-Based DCT Quantization for Color Image Compression. *Human Vision*, *Visual Processing*, *& Digital Display III*, **1666**. https://doi.org/10.1117/12.135982

[13] Watson, A.B. (1993) A Technique for Visual Optimization of DCT Quantization Matrices for Individual Images. *Society for Display Digest of Technical Papers XXIV*, **9**, 946-949. https://doi.org/10.2514/6.1993-4512

[14] Zhang, X.H., Lin, W.S. and Xue, P. (2005) Improved Estimation for Just-Noticeable Visual Distortion. *Signal Processing*, **85**, 795-808. https://doi.org/10.1016/j.sigpro.2004.12.002

[15] Wei, Z. and Ngan, K.N. (2009) Spatio-Temporal Just Noticeable Distortion Profile for Grey Scale Image/Video in DCT Domain. *IEEE Transaction on Circuits and Systems for Video Technology*, **19**, 337-346. https://doi.org/10.1109/TCSVT.2009.2013518

[16] Wu, J., Shi, G., Lin, W., Liu, A. and Qi, F. (2013) Just Noticeable Difference Estimation for Images with Free-Energy Principle. *IEEE Transactions on Multimedia*, **15**, 1705-1710. https://doi.org/10.1109/TMM.2013.2268053

[17] Chen, Z. and Guillemot, C. (2010) Perceptually-Friendly H.264/AVC Video Coding Based on Just-Noticeable-Distortion Model. *IEEE Transactions on Circuits and Systems for Video Technology*, **20**, 806-819. https://doi.org/10.1109/TCSVT.2010.2045912