Adaptive Block-Size Transform Based Just-Noticeable Difference Profile for Videos

Lin Ma and King N. Ngan
Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong
Email: {lma, knngan}@.ee.cuhk.edu.hk

***Abstract*—In this paper, we propose a novel adaptive block-size transform (ABT) based just-noticeable difference (JND) model for videos. Firstly, the ABT-based spatial JND profile is extended to spatial-temporal JND model for videos by considering temporal contrast sensitivity function (TCSF), eye movement, and the motion information of the objects in video sequence. Furthermore, a metric named motion characteristics distance (MCD) is proposed to depict the motion characteristics similarity between a macroblock and its corresponding sub-blocks. Based on the proposed MCD and the obtained spatial image content information, a novel balanced strategy is proposed to determine which transform size is employed to generate the resulting JND model. Experimental results have demonstrated that our proposed scheme could tolerate more distortions while preserving better perceptual quality than other JND profiles, which means that the proposed model consists well with human vision system (HVS). Moreover, for the balanced strategy, experiments have shown that temporal motion characteristics accord very well with the spatial image content information, which has demonstrated the efficiency of our proposed balanced strategy.**

I. INTRODUCTION

JND accounts for the smallest detectable difference between a staring and secondary level of a particular sensory stimulus in psychophysics [1], which is also known as the difference limen or differential threshold. The JND model can be employed for depicting and modeling the property of HVS efficiently in many multimedia processing research areas, such as perceptual image/video compression [4]-[6], image/video perceptual quality evaluation [2] [3] etc.

Generally, automatic JND model can be determined in the image domain [7] [8], or transform domain [6], or even their combination [11]. JND model generated in image domain, which is also denoted as pixel-based JND, mainly focuses on the background luminance adaptation and the spatial contrast masking. In [9], Yang et al. deduce the overlapping effect of luminance adaptation and spatial contrast masking to refine the JND profile in [7]. However, pixel-based JND model has not considered the HVS sensitivity for different frequency components. Therefore it could not describe the HVS property accurately. JND model generated in the transform domain named the subband-based JND usually incorporates all the major affecting factors, namely, CSF, luminance adaptation, and contrast masking. In [4], the DCT JND thresholds are developed based on spatial CSF. Then the basic JND model is improved by Watson [5] in DCTune model after considering contrast masking. More recently, Wei et al. [10] incorporate new formulae of luminance adaptation, contrast masking and Gamma correction to estimate the JND threshold in DCT domain. Moreover, in order to exploit the HVS property over different transform sizes, Ma et al. [12] proposed to combine ABT together with the DCT-based JND profile for images. In addition, in order to generate a JND model for videos, temporal HVS property should be taken into account. An empirical function based on the luminance difference between adjacent frames is proposed in [7] [9] to model the temporal masking property. And Kelly [13] proposed to measure the spatio-temporal CSF model at a constant retinal velocity, which is tuned to a particular spatial frequency. Based on Kelly's model, Jia et al. [14] estimated JND for videos by considering spatio-temporal CSF, eye movement. Furthermore, Wei et al. [10] take the directionality of the motion into consideration to produce the resulting temporal modulation factor.

Recently, ABT has attracted researcher's attention for developing the next generation video coding standard. Specifically, larger block-size based transforms could provide better energy compaction and better preservation of details, while smaller ones could prevent more ringing artifacts during compression [15]. Also we have shown its potential in exploiting HVS property during the development of spatial JND model for images [12]. Therefore, ABT based JND model for videos is promising, which could be easily applied in many multimedia perceptual researches, such as video compression, video quality assessment, and so on.

In this paper, inspired by recent progresses on ABT based video coding strategy, a novel ABT based JND profile for vid eos is proposed. Firstly, temporal HVS property is modeled by considering TCSF, eye movement, and the motion informatio n of the objects in the video sequence. Secondly, a novel balan ced strategy for adaptively adjusting the temporal motion char acteristics and the spatial image content is proposed to determi ne which transform size is employed to yield the resulting JN D profile. The rest of the paper is organized as follows. In Sect ion II, the proposed ABT based JND profile for video is introd uced. Experimental results are shown in Section III. Finally, S ection IV concludes the paper.

II. THE PROPOSED ABT BASED JND MODEL FOR VIDEOS

JND profile for videos in the DCT domain could be deter mined by considering both spatial HVS property T_{spatio} and te mporal modulation factor T_{tempo} , which could be expressed as:

This work was partially supported by a grant from the Chinese University of Hong Kong under the Focused Investment Scheme (Project 1903003).

$$T(k,m,n,i,j) = T_{spatio}(m,n,i,j) \times T_{tempo}(k,m,n,i,j), \quad (1)$$

where k denotes the frame index of the video sequence, (m,n) is the position of DCT block in the current frame, (i,j) indicates the DCT subband, and T is the final obtained JND threshold for each corresponding frame. Since the spatial ABT based JND model for images has been discussed in [12], we will focus on the HVS temporal property.

A. Temporal Modulation Factor

Based on the temporal contrast sensitivity experiment, Robson [16] gave the TCSF results as shown in Figure 1. It has shown that the form of the fall-off in the sensitivity at high spatial frequencies is independent of the temporal frequency and vice versa, while a fall-off in sensitivity at low spatial frequencies occurs only when the temporal frequency is also low and vice versa. Therefore, we would like to fit the similar shape at high spatial frequencies and the low spatial frequency with high temporal frequencies (larger than 10Hz). In [10], it has been revealed that the logarithm of the temporal contrast sensitivity values nearly follows the same slope for different spatial frequencies. And an empirical slope -0.03 is adopted. TCSF could be further modeled as:

$$\log(TCSF(\omega_t)) = -0.03 \cdot \omega_t + TCSF(\omega_0), \quad (2)$$

where ω_t is the temporal frequency, $TCSF(\omega_t)$ denotes the contrast sensitivity value when temporal frequency is 0, which means that only spatial CSF is considered to generate the final contrast sensitivity value. As JND is the reciprocal of sensitivity value, the temporal modulation factor could be modeled as $10^{-0.03\omega_t}$, and Equation (1) could be expressed as:

$$T(k,m,n,i,j) = T_{spatio}(m,n,i,j) \times 10^{-0.03\omega_t}. \quad (3)$$

Also the TCSF curves in Figure 1 have shown the characteristic of a bandpass filter at the lower spatial frequencies. According to Kelly's model [13], the contrast sensitivity is nearly constant for the temporal frequencies less than 10Hz. Therefore, an empirical formula for calculating temporal modulation factor, by accounting both spatial ω_s and temporal frequency ω_t, could be further derived as:

$$T_{tempo} = \begin{cases} 1 & \omega_s < 5cpd \ \& \ \omega_t < 10Hz \\ 10^{-0.03(\omega_t - 10)} & \omega_s < 5cpd \ \& \ \omega_t \geq 10Hz \\ 10^{-0.03\omega_t} & \omega_s \geq 5cpd \end{cases}. \quad (4)$$

B. Temporal Frequency Calculation

Actually, the temporal frequency of a video signal depends on not only the motion information, but also the spatial frequency of the object [17], which is demonstrated as:

$$\omega_t = \omega_{sx} \cdot v_x + \omega_{sy} \cdot v_y, \quad (5)$$

where ω_{sx} and ω_{sy} are the horizontal and vertical component of the spatial frequency, respectively, which are determined for different transform sizes. Detailed information of the DCT spatial frequencies calculated for different transform sizes could be referred to [12]. v_x and v_y are retina velocities for depicting the object motion. By considering eye movement v_{EM} and the object move information v_I in image plane, the retina velocity [18] could be calculated according to:

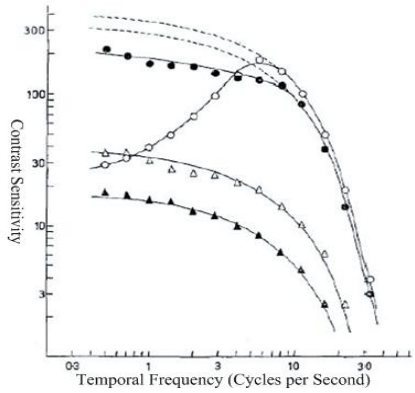

Figure 1.Temporal contrast sensitivity (reciprocal of threshold contrast) over different spatial frequency: 22 cpd (filled triangle), 16cpd (open triangle), 4cpd (filled circle), 0.5cpd (open circle).

$$v_\Delta = v_{I\_\Delta} - v_{EM\_\Delta} \quad (\Delta = x, y). \quad (6)$$

And the eye movement velocity could be determined by:

$$v_{EM\_\Delta} = \min[G_s \times v_{I\_\Delta} + v_{\min}, v_{\max}] \quad (\Delta = x, y), \quad (7)$$

where G_s denotes the gain of the smooth pursuit eye movements, v_{min} indicates the minimum eye velocity due to the drift movement, and v_{max} is the maximum eye velocity related to the saccadic eye movement, which are empirically set as 0.98, 0.15 deg/s, 80 deg/s, respectively. And the velocity on image plane v_I is generated by:

$$v_{I\_\Delta} = f_r \times MV_\Delta \times \theta_\Delta \quad (\Delta = x, y), \quad (8)$$

where f_r denotes the frame rate of the video sequence, MV_Δ is the motion vector of the object, which could be approximated by the block-based motion estimation (BME). For different transform sizes, different size BMEs are employed to generate the velocity on image plane. θ_Δ is the visual angle, which could be generated by viewing distance d and the display width/height of a pixel Γ on the monitor:

$$\theta_\Delta = 2 \cdot \arctan(\Gamma_\Delta / (2 \cdot d)) \quad (\Delta = x, y). \quad (9)$$

C. Balanced Strategy for JND Profile between Different Block-size Transforms

In [12], A balanced strategy is proposed to adaptively adjust the image content to determine which transform block-size is utilized to generate the final spatial JND profile. In order to extend the balanced strategy from spatial to temporal, we should consider not only spatial image content, but also the temporal video motion characteristics, which is approximately described by motion vectors of the objects, especially of different block sizes.

The balanced strategy for video JND is illustrated in Figure 2, where two macroblocks ***A*** and ***C*** are taken as examples. Firstly, the spatial image content type of the macroblock and its corresponding sub-blocks should be determined and categorized into three types [12], namely PLANE, EDGE, and TEXTURE, respectively. From Figure 2, we can see that spatial content types of the macroblock ***A*** and its sub-blocks appear the same, which are indicated by the same color. However, the sub-block content types of ***C*** are different with each

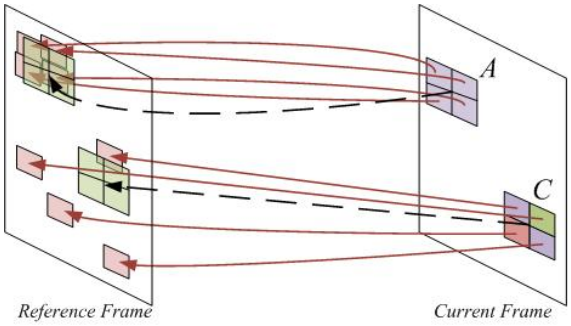

Figure 2. Balance Strategy between 16×16 and 8×8 JND profiles for video (Dash arrow line denotes the motion information of macroblock (16×16), solid arrow line represents the motion information of sub-blocks (8×8).)

other denoted in different colors. Therefore, ***A*** and ***C*** needs to be processed separately. Subsequently, BME is employed to depict the temporal motion characteristics of the objects. We further define a metric, named motion characteristics distance (MCD), to evaluate the similarity of the motion information between the macroblock and its corresponding sub-blocks:

$$MCD=\sqrt{\sum_{i=0}^{N-1}[(MV_{8\_i\_x}-MV_{16\_x})^2+(MV_{8\_i\_y}-MV_{16\_y})^2]}\Big/N, \quad (10)$$

where N denotes the number of sub-blocks of each macroblock (in the proposed scheme, N=4 is employed, which means that four 8×8 sub-blocks compose the 16×16 macroblock), MV_{8_i} and MV_{16_i} indicate the motion vectors for 8×8 and 16×16 block, respectively. Based on MCD, we could make decision on which size of transforms is employed for yielding the resulting JND profile. If the calculated MCD appears very small, which means that the motion characteristics of the macroblock and its corresponding sub-blocks, such as ***A*** in Figure 2, appear nearly the same as each other. Therefore, by considering the same spatial content type and similar temporal motion information, 16×16 DCT based JND profile will be utilized to generate the JND threshold for ***A***. However, if the obtained MCD appears very large, which indicates that motion vectors of the macroblock and its sub-blocks appear diversely, like ***C*** in Figure 2. The 4 sub-blocks of ***C*** are divided separately and move independently. Consequently, we could not regard C as a unit for the different spatial image contents and diverse motion characteristics. 8×8 DCT based JND profile for each sub-block will be thereby employed to depict the resulting JND model. And we employ a hard thresholding scheme to describe the property of MCD. If MCD is smaller than a threshold T_{thr}, the macroblock is temporally regarded as a unit. On the contrary, if MCD is larger than T_{thr}, the macroblock should be divided into sub-blocks which are individually temporally considered. Actually T_{thr} should be set adaptively according to the content and motion information of the input video sequence. However, T_{thr} is empirically set as 1.25 in our implementation for simplicity, which means that the average tolerance of motion vector difference between the macroblock and its sub-blocks is no larger than 1.25 pixels.

Actually, there also exist macroblocks of which the image content appears the same, while temporal motion will describe the sub-blocks individually, or the image content presents differently, and the motion information regard it as a whole part. However, for the natural video sequences, if a macroblock is recognized as a unit with the same spatial image content, it will be very likely to appear as a whole part while considering the temporal motion characteristics. Also the sub-blocks with different spatial image contents will move separately and independently while accounting the temporal motion characteristics Therefore, the macroblocks, of which the spatial and temporal information does not agree with each other, will occur in very small probability. Experimental results will illustrate this situation in Section III.

III. Experimental Results

In order to demonstrate the efficiency of the proposed ABT-based JND model for videos, distortion is injected into each DCT coefficient of each video frame to evaluate the error tolerance ability of HVS according to:

$$\hat{I}_{typ}(k,m,n,i,j)=I_{typ}(k,m,n,i,j)+Ran_{(k,m,n,i,j)}\cdot T_{JND\_typ}(k,m,n,i,j), \quad (11)$$

where $\hat{I}_{typ}$ is the noise-contaminated DCT coefficient which located on the $(i,j)th$ subband of $(m,n)th$ block in k frame, $Ran_{(k,m,n,i,j)}$ takes +1 or -1 randomly to avoid introducing a fixed pattern of changes, T_{JND_typ} is the JND threshold obtained by the proposed ABT-based scheme, typ denotes the final transform block-size to generate the resulting JND profile.

The proposed JND model is tested on several typical CIF (352×288) video sequences, with the frame rate as 30 fps. In our experiments, 250 frames of each sequence are tested, with the first frame as INTRA and the rest as INTER frames. We compare our method with Yang et al.'s method [9], which estimates the JND profile in pixel domain, and Wei et al.'s JND model [10] generated in DCT domain. Comparisons in terms of PNSR are listed in Table I. As we have evaluated the efficiency of ABT based JND model for images in [12], only the average PSNR of INTER frames is calculated to evaluate the efficiency of different JND profiles for videos. From Table I, it has been clearly shown that the proposed JND model could yield smaller PSNR values than other JND profiles, which means that our JND profile could tolerate more distortions.

Table I. PSNR Comparison between Different JND Profiles.

Video Sequence	Yang	Wei	Proposed Profile
TEMPETE	31.68dB	27.42dB	27.04dB
FOOTBALL	34.43dB	28.39dB	28.17dB
FOREMAN	35.29dB	28.29dB	28.02dB
MOBILE	33.10dB	27.48dB	26.93dB
SILENCE	34.43dB	28.26dB	27.93dB

In order to provide a more convincing evaluation of the proposed JND model, subjective test is conducted to assess the perceptual quality of the noise-contaminated videos. Double stimulus continuous quality scale (DSCQS) method, as specified in ITU-R BT.500 [19], is employed to evaluate the perceptual quality. Two sequences are presented to viewers, of which one is original and the other is processed. Ten viewers (half of them are experts in image/video processing and the other half are not) are asked to offer their opinions. Mean opinion score (MOS) is scaled for testers to vote: Bad (0-20), Poor (20-40), Fair (40-60), Good (60-80), and Excellent (80-100). Then the difference between MOSes of original and noise-injected video sequence is calculated as the differential mean opinion score (DMOS). Therefore, the smaller the DMOS, the higher quality the noise-contaminated video

presents. In this experiment, the viewing monitor is a Viewsonic Professional series P225fb CRT display. The viewing distance is set as 4 times the image height. Detailed information of the subjective test results is depicted in Table II. From the test results, we can see that average DMOS of the proposed scheme is only 6.98, which means that the noise-injected video sequences by our method have similar quality with the original videos. Therefore, our methods could effectively exploit the HVS property.

Table II. Subjective Evaluation Results (DMOS for Noise-contaminated video Sequences at 30fps).

Video Sequence	Yang	Wei	Proposed Profile
TEMPETE	7.3	6.6	6.4
FOOTBALL	7.6	6.2	5.6
FOREMAN	13.2	9.2	8.3
MOBILE	9.7	7.0	7.1
SILENCE	13.9	8.7	7.5

In order to further test the consistency between the temporal balanced strategy and spatial balanced strategy, firstly we need to examine the macroblock spatial content type according to the balanced strategy proposed by [12]. Furthermore, we would like to record the motion characteristic type for each macroblock of INTER frames by the balanced strategy presented in Section II C. Based on the obtained spatial and temporal macroblock type, a hit ratio (HR) curve is utilized to demonstrate the hit rate for each frame of the test video sequences. For each test INTER frame, the hit rate h indicates the percentage of the macroblocks whose types are consistent between spatial image content and temporal motion characteristic, which means that both spatial and temporal balanced strategy select the same size DCT (8×8 or 16×16) for a macroblock to generate the resulting JND model. The HR curves for each video sequence are illustrated in Figure **3**. We can see that the hit rates of FOOTBALL and FOREMAN is a bit lower than the other sequences, with the average hit rate as 77%. That is because the two sequences both contain high motion characteristics. Therefore, the consistency between spatial and temporal characteristics seems a little loose. However, as the motion appears slightly, the hit rates of the other sequences are much higher, with the average hit rate of the other sequences as 93%. Furthermore, the hit rates h of different sequences mostly appear higher than 70%, which means that the proposed temporal balanced strategy accords very well with former proposed spatial balanced strategy [12]. Also it means that the balanced strategy is efficient and meaningful for depicting both spatial image content and temporal video motion characteristics.

IV. CONCLUSION

In this paper, a novel ABT-based JND profile for videos is proposed by exploiting the HVS properties over different transform sizes. A new balanced strategy is proposed for each macroblock to decide which transform block-size is to be employed by considering not only spatial image content but also temporal video motion characteristics. Based on the proposed model, our JND profile could tolerate more distortions with the same visual quality compared with other JND models, which means that our model is more effective in exploiting the HVS properties.

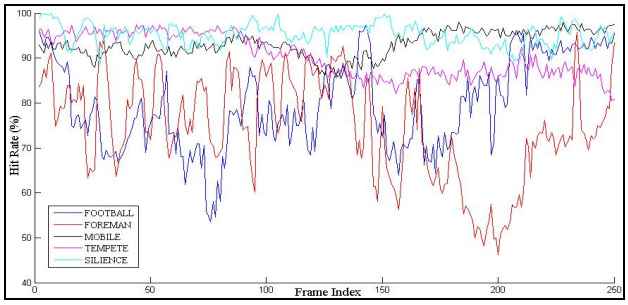

Figure 3. The HR curves of the macroblocks for each INTER frame of different test video sequences.

REFERENCES

[1] Weber's Law of Just Noticeable Differences, http://www.usd.edu/psyc301/WebersLaw.htm

[2] W. Lin, L. Dong, P. Xue, "Visual Distortion Gauge Based on Discrimination of Noticeable Contrast Changes", IEEE Transactions on Circuits and Systems for Video Technology. Vol. 15, no. 7, pp. 900-909, 2005.

[3] Z. Lu, W. Lin, X. Yang, E. Ong, S. Yao, "Modeling Visual Attention's Modulatory Aftereffects on Visual Sensitivity and Quality Evaluation", IEEE Transactions on Image Processing. Vol. 14, no. 11, pp. 1928-1942, 2005.

[4] A. J. Ahumada, H. A. Peterson, "Luminance-model-based DCT Quantization for Color Image Compression", Proceedings of the SPIE, Human Vision, Visual Porcessing, and Digital Display III 1666, pp. 365-374, 1992.

[5] A. B. Watson, "DCTune: A Technique for Visual Optimization of DCT Quantization Matrices for Individual Images", Society for Information Display (SID) Digest 24, pp. 946-949, 1993.

[6] I. Hontsch, L. J. Karam, "Adaptive Image Coding with Perceptual Distortion Control", IEEE Transactions on Image Processing. Vol. 11, no. 3, pp. 213-222, 2002.

[7] C. Chou, C. Chen, "A Perceptual Optimized 3-D Subband Codec for Video Communication Over Wireless Channels", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 6, pp. 143-156, 1996.

[8] Y. Chin, T. Berger, "A Software-only Videocodec Using Pixelwise Conditional Differentialreplenishment and Perceptual Enhancements", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, pp. 438-450, 1999.

[9] X. Yang, W. Lin, Z. Lu, E. Ong, S. Yao, "Motion-Compensated Residue Preprocessing in Video Coding Based on Just-Noticeable-Distortion Profile", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, pp. 742-750, 2005

[10] Z. Wei, N. K. Ngan, "Spatial-Temporal Just Noticeable Distortion Profile for Grey Scale Image/Video in DCT Domain", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 19, pp. 337-346, 2009.

[11] X. Zhang, W. Lin, P. Xue, "Just-Noticeable Difference Estimation with Pixels in Images", Journal of Visual Communication and Image Representation, Vol. 19, pp. 30-41, 2008.

[12] L. Ma, King N. Ngan, "Adaptive Block-Size Transform Based Just-Noticeable Difference Profile for Images", Accepted by IEEE Pacific-Rim Conference on Multimedia, 2009.

[13] D. H. Kelly, "Motion and Vision II. Stabilized Spatio-temporal threshold Surface", J. Opt. Soc. Am. Vol. 69, pp. 1340-1349, 1979.

[14] Y. Jia, W. Lin and A. A. Kassim, "Estimating Just-Noticeable Distortion for Video", IEEE Transactions on Circuits and Systems for Video Technology, Vol. 16, pp. 820-829, 2006.

[15] J. Dong, J. Lou, C. Zhang, L. Yu, "A New Approach to Compatible Adaptive Block-Size Transforms", VCIP , 2005.

[16] J. G. Robson, "Spatial and Temporal Contrast Sensitivity Functions of the Visual Seystem", J. Opt, Soc. Am. Vol. 56, pp. 1141-1142, 1966.

[17] Y. Wang, J. Ostermann, Y. Zhang, "Video Processing and Communications", Prentice Hall, 2002.

[18] S. Daly, "Engineering Observations from Sptaiovelocity and Spatiotemporal visual models", Proc. SPEI, Vol. 3299, pp. 180-191, 1998.

[19] "Methodology for the Subjective Assessment of the Quality of Television Pictures", *ITU-R BT.500.11*, 2002.