Motion-Compensated Residue Preprocessing in Video Coding Based on Just-Noticeable-Distortion Profile

Xiaokang Yang, *Senior Member, IEEE*, Weisi Lin, *Senior Member, IEEE*, Zhongkhang Lu, *Senior Member, IEEE*, EePing Ong, *Member, IEEE*, and Susu Yao, *Member, IEEE*

***Abstract*—We present a motion-compensated residue signal preprocessing scheme in video coding scheme based on just-noticeable-distortion (JND) profile. Human eyes cannot sense any changes below the JND threshold around a pixel due to their underlying spatial/temporal masking properties. An appropriate (even imperfect) JND model can significantly help to improve the performance of video coding algorithms. From the viewpoint of signal compression, smaller variance of signal results in less objective distortion of the reconstructed signal for a given bit rate. In this paper, a new JND estimator for color video is devised in image-domain with the nonlinear additivity model for masking (NAMM) and is incorporated into a motion-compensated residue signal preprocessor for variance reduction toward coding quality enhancement. As the result, both perceptual quality and objective quality are enhanced in coded video at a given bit rate. A solution of adaptively determining the parameter for the residue preprocessor is also proposed. The devised technique can be applied to any standardized video coding scheme based on motion compensated prediction. It provides an extra design option for quality control, besides quantization, in contrast with most of the existing perceptually adaptive schemes which have so far focused on determination of proper quantization steps. As an example for demonstration, the proposed scheme has been implemented in the MPEG-2 TM5 coder, and achieved an average peak signal-to-noise (PSNR) increment of 0.505 dB over the twenty video sequences which have been tested. The perceptual quality improvement has been confirmed by the subjective viewing tests conducted.**

***Index Terms*—Human visual system (HVS), just-noticeable-distortion (JND), rate-distortion, video coding.**

I. Introduction

SINCE the ultimate receiver of most decompressed video signal is the human visual system (HVS), the goal of video compression and coding should aim at the lowest bit rate for signal representation at certain level of perceptual quality, or more often than not, the highest perceptual quality with a given bit rate. In the latter case, it is imperative for us to design a coding algorithm that minimizes perceptual distortion between the original and the decoded visual signal.

Human eyes cannot sense any changes below the just noticeable distortion (JND) threshold around a pixel due to their underlying spatial/temporal sensitivity and masking properties [1]. An appropriate (even imperfect) JND model can significantly help to improve the performance of video coding algorithms. Several methods for finding JND have been proposed, based upon intensive research in subbands, such as discrete cosine transform (DCT) and wavelet domain [2]–[12], as well as some work in image-domain [12]–[14].

Perceptual coding has been so far focused mainly on determination of proper quantization steps for image coding with subband JND [3]–[5], [7], [6], [8], [12], [13]. A few attempts have been made to nonstandard video coding [14], [9]. In [14], image-domain JND has been used as the threshold for inter-frame replenishment with low-motion (like head-and-shoulder) scenes in low bit-rate videophony. In [9], subband JND has been used in the quantization process and also in controlling the block splitting process in variable-size motion search.

This paper aims at proposing a motion-compensated residue signal preprocessing scheme in video coding based on JND profile. It can be applied to any prevalent standardized hybrid video coding (cascading of motion estimation and DCT), such as H.261/263 [15], [16], MPEG-1/2/4 [17]–[19]. If visual signal to be coded has smaller variance, less objective distortion is resulted in the reconstructed frame for a given bit rate [20], [21]. We attempt to explore the methodology that reduces the variance of motion compensated residues in an intercoded frame under the JND guidance. A JND-adaptive preprocessing module is proposed to achieve this by adding or subtracting an appropriate quantity regulated by the JND profile. A parameter used to adjust the said quantity is determined by minimizing the overall objective distortion for the current frame, i.e., the sum of the quantization distortion for the preprocessed residue signal and the distortion introduced by the residue signal preprocessor. The preprocessing module provides an extra design option for coding quality control, besides quantization.

Because of the importance of accurate JND profile to this work, a new JND estimator for color video will be proposed based on our initial work [22]. The proposed preprocessing is applied to motion-compensated residues in image domain and, therefore, image-domain JNDs are more convenient to be used than subband JNDs. Furthermore, handling of perceptual sensitivity on edge regions is easier in image domain than in

Manuscript received April 2, 2003; revised November 26, 2003. This paper was recommended by Associate Editor J.-R. Ohm.

X. K. Yang was with the Media Division, Institute for Infocomm Research, Agency for Science, Technology, and Research (A-STAR), 119613 Singapore. He is currently with the Institute of Image Communication and Information Processing, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: xkyang@ieee.org; xkyang@sjtu.edu.cn).

W. Lin, Z. Lu, E. P. Ong, and S. Yao are with Media Division, Institute for Infocomm Research, Agency for Science, Technology and Research (A-STAR), 119613 Singapore (e-mail: wslin@i2r.a-star.edu.sg; zklu@i2r.a-star.edu.sg; epong@i2r.a-star.edu.sg; ssyao@i2r.a-star.edu.sg).

Digital Object Identifier 10.1109/TCSVT.2005.848313

subbands (due to the shift variance in a transform domain). Luminance masking and texture masking are the major considerations in image-domain JNDs. In Chou's work [12], [13], texture masking was determined with the average background luminance and the average luminance difference around the pixel. The spatial JND is then decided by the dominant effect of texture masking and luminance masking. Temporal masking was accounted for by multiplying the spatial JND with the average interframe luminance difference. In Chiu's system [14], JND was formulated as the weighted sum of luminance masking threshold and the relative magnitude of a spatial/temporal activity measure. In both Chou's and Chiu's methods, only JND for the luminance component is considered, and the edge regions are not distinguished from the nonedge ones.

In this work, a new formula nonlinear additivity model for masking (NAMM) for spatial JND in image-domain has been devised to generalize the existing approaches [12], [14], in an attempt to match the HVS characteristics better. In the NAMM, effects of luminance adaptation and texture masking are added with provision to deduct their overlapping, in analogy with the saliency effect from different stimuli in the recent vision research results [23]. The new model accounts for the difference between edge regions and nonedge regions, since masking in edge regions is not as significant as in nonedge regions [24]. The formula is also applied to color components.

The rest of the paper is organized as follows. In Section II, we present the NAMM model for image-domain JND profile for color video. In Section III, the basic framework of the proposed JND-based preprocessing scheme for motion-compensated residue is presented. In Section IV, a solution for determining the model parameter is presented based on distortion minimization. In Section V, the experimental results on overall performance of the scheme is given. Finally, we conclude the paper in Section VI.

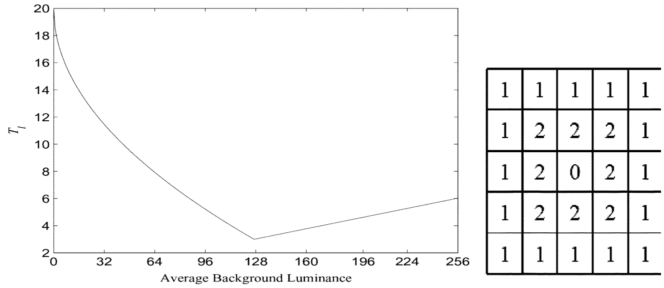
II. Image-Domain JND Profile for Color Video

In this section, the spatial part JND, $\mathrm{JND}_s(x,y)$, is to be firstly considered with visual information within the frame $I(x,y)$. Spatiotemporal JND, $\mathrm{JND}(x,y)$, is then obtained by integrating temporal (interframe) masking with $\mathrm{JND}_s(x,y)$.

A. Spatial JND With NAMM

There are primarily two factors affecting spatial luminance JND in image domain: a) background luminance masking, because the HVS is sensitive to luminance contrast rather than the absolute luminance value; b) texture masking, because the reduction of visibility for changes is caused by the texture (nonuniformity) in the neighborhood, and, therefore, textured regions can hide more error than smooth or edge areas.

Since these two types of masking co-exist in most images, how to effectively integrate them is an important issue in obtaining accurate spatial JND profile. We believe that: i) combinative effect of multiple maskings should take some form of addition (not linear addition though) of individual factors, because simultaneous existence of multiple masking factors in a neighborhood makes targets (e.g., coding artifacts in a decoded image) more difficult to be noticed when compared with the case of one source of masking alone; ii) masking effect in chrominance channels could be also exploited to improve compression performance; iii) distinction of edge regions from smooth and textured regions avoid over-estimation of masking effect around the edge.



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

(a) (b)

Fig. 1. Illustration of modeling T_l: (a) T_l as a function of average background luminance, (b) the weighted low-pass filter for calculating average background luminance.

The spatial JND of each pixel can be described in the following NAMM

$$\mathrm{JND}_s(x,y) = T_l(x,y) + T_t(x,y) - C_{l,t} \cdot \min\{T_l(x,y), T_t(x,y)\} \quad (1)$$

where $T_l(x,y)$ and $T_t(x,y)$ are the visibility thresholds for the two primary masking factors, background luminance masking and texture masking; and $C_{l,t}(0 < C_{l,t} < 1)$ accounts for the overlapping effect in masking. The parameter selection of $C_{l,t}$ allows the compound effect for co-existence of luminance masking and texture masking to be reflected fully in (1).

The JND estimator in [12] is a special case of the proposed NAMM, because if $C_{l,t} = 1$, (1) becomes

$$\mathrm{JND}_s(x,y)^{\mathrm{simplified-I}} = \max\{T_l(x,y), T_t(x,y)\} \quad (2)$$

The JND estimator in [14] is also a special case of the proposed NAMM when $T_l(x,y)$ is considered as the major masking factor, i.e., $\min\{T_l(x,y), T_t(x,y)\} \equiv T_t(x,y)$. In this case, (1) becomes

$$\mathrm{JND}_s(x,y)^{\mathrm{simplified-II}} = T_l(x,y) + C' \cdot T_t(x,y) \quad (3)$$

where $C' = 1 - C_{l,t}$. In [14], C' is determined according to the magnitude of $T_l(x,y)$.

According to the experimental results of the prior works in [3] and [12], the relationship between $T_l(x,y)$ and the average background luminance is modeled by a root equation for low background luminance (below 127) while the other part (over 127) is approximated by a linear function, as illustrated in Fig. 1(a) or equivalently described as follows:

$$T_l(x,y) = \begin{cases} 17\left(1 - \sqrt{\frac{\overline{I(x,y)}}{127}}\right) + 3, & \text{if } \overline{I(x,y)} \le 127 \\ \frac{3}{128}(\overline{I(x,y)} - 127) + 3, & \text{otherwise} \end{cases} \quad (4)$$

with

$$\overline{I(x,y)} = \frac{1}{32}\sum_{i=1}^{5}\sum_{j=1}^{5} I(x-3+i, y-3+j) \cdot B(i,j) \quad (5)$$

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

g_1

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

g_2

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

g_3

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

g_4

Fig. 2. Directional high-pass filters for texture detection.

where $B(i,j)$ are a weighted low-pass filter [12] as illustrated in Fig. 1(b).

Texture masking can be determined with local spatial activities (e.g., gradients around the pixel). For more accurate JND estimation, texture masking in edge and nonedge regions has to be distinguished. Edge is directly related to the image content that demarcates object boundaries, surface crease, reflectance change and other significant visual events. Distortion around edge is easier to be noticed than that in smooth and textured regions due to the fact that edge structure attracts more attention from a typical HVS [24], and there is a substantial body of literature attesting to the importance of edges to primate perception (e.g., [25], [26]). The proposed $T_t(x,y)$, therefore, takes the difference for edge into account

$$T_t(x,y) = \eta \cdot G(x,y) \cdot W_e(x,y) \tag{6}$$

where η is a control parameter; $G(x,y)$ denotes the maximal weighted average of gradients around the pixel at (x,y) [12]

$$G(x,y) = \max_{k=1,2,3,4}\{|\mathrm{grad}_k(x,y)|\} \tag{7}$$

with

$$\mathrm{grad}_k(x,y) = \frac{1}{16}\sum_{i=1}^{5}\sum_{j=1}^{5} I(x-3+i, y-3+j)\cdot g_k(i,j) \tag{8}$$

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

$W_e(x,y)$ is an edge-related weight of the pixel at (x,y), and its corresponding matrix $\mathbf{W_e}$ is computed by edge detection followed with a Gaussian low-pass filter

$$\mathbf{W_e} = \boldsymbol{\mathcal{L}} * \mathbf{h} \tag{9}$$

where $\boldsymbol{\mathcal{L}}$ is the edge map of the original video frame, with element values of 0.1 and 1 for edge and nonedge pixels, respectively; $\mathbf{h}$ is a $k \times k$ Gaussian low-pass filter with standard deviation σ ($k = 7$ and $\sigma = 0.8$ in this work). For edge detection, the Canny detector [27] is used with the sensitivity thresholds of 0.5.

The NAMM model is also applied to all color components by considering the overlapping effects between luminance masking and texture masking in chrominance channels. The performance of NAMM has been evaluated on color images in our previous work [22]. As a summary of the performance, the proposed NAMM scheme provides a more accurate JND profile toward the actual JND bound in the HVS, since it is capable of exploiting larger JND values without jeopardizing the visual quality; it showed to outperform the approach in [12] by more than 2 dB (in terms of PSNR) of permitted data redundancy on an average for a same level of visual quality.

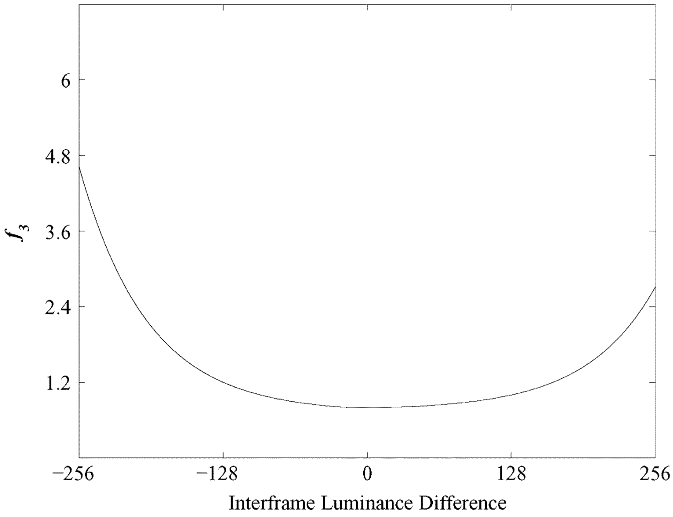


Fig. 3. Temporal masking effect.

B. Spatio-Temporal JND Profile

Temporal effect has been added to form the final JND. Usually bigger interframe difference (caused by motion) leads to larger temporal masking, as the empirical curve (approximation for the results in [13]) shown in Fig. 3. The overall JND can be, therefore, denoted as

$$\mathrm{JND}(x,y,t) = f_3(\mathrm{ild}(x,y,t)) \cdot \mathrm{JND}_s(x,y) \tag{10}$$

where $\mathrm{ild}(x,y,t)$ represents the average interframe luminance difference between the tth and $(t-1)$th frame [13]

$$\begin{aligned}\mathrm{ild}(x,y,t) = \frac{1}{2}(&I(x,y,t) - I(x,y,t-1)\\ &+ \overline{I(x,y,t)} - \overline{I(x,y,t-1)})\end{aligned} \tag{11}$$

where $\overline{I(x,y,t)}$ is the average intensity and $f_3()$ is the empirical function defined in Fig. 3. In the case of small interframe changes (i.e., $|\mathrm{ild}(x,y,t)| < 5$), the scaling factor is around its minimum ($f_3(\mathrm{ild}(x,y,t) \simeq 0.8$). The scaling factor increases with the increase of interframe luminance difference.

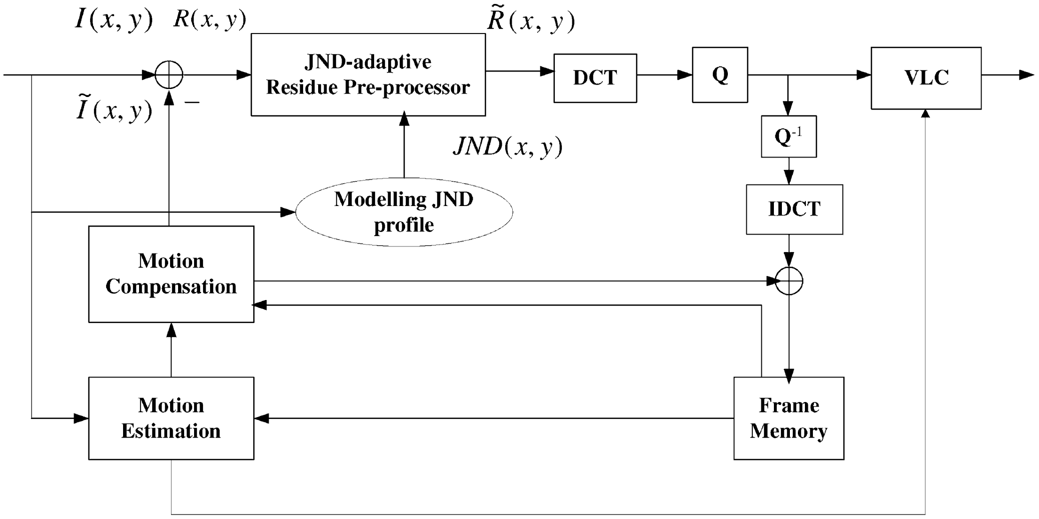


Fig. 4. The framework for the proposed perceptually adaptive video coding scheme based on JND profile.

III. JND-Based Preprocessing of Motion-Compensated Residues

The framework for the proposed perceptually adaptive JND-based preprocessing scheme for motion-compensated residue is given in Fig. 4, where $I(x,y), \tilde{I}(x,y), R(x,y)$, $\tilde{R}(x,y)$, and $\mathrm{JND}(x,y)$, respectively, denote the original intensity, the motion compensated intensity, the residue signal before JND-adaptive preprocessing, the residue signal after JND-adaptive preprocessing and the JND value for a pixel located at (x,y) of the current frame; $\mathrm{JND}(x,y)$ can be obtained using the JND estimator presented in Section II.

The addition of the proposed scheme to a typical hybrid video coding scheme is: before the DCT stage, each motion compensated residue is adjusted by a JND-adaptive preprocessor. The proposed scheme can be completely compatible with all current video coding standards, since it is unnecessary to transmit extra side-information and change the coding syntax.

The JND-adaptive residue preprocessor can be described as (12), shown at the bottom of the page, where $\overline{R_B}$ is the average of the 64 residues in the 8×8 block around $(\mathrm{x,y})$; $\lambda(\in [0,1])$ is a parameter to be determined. The constraint of $\lambda \in [0,1]$ is to avoid introducing perceptual distortion into motion compensated residues in the JND-adaptive preprocessing process.

Signal distortion of a conventional video coder is introduced by the quantizer for DCT coefficients. With the proposed scheme, additional consideration about distortion is the JND-adaptive residue preprocessing. Let D_1 and D_2, respectively, denote the distortion of coding the preprocessed residues and the distortion due to JND-adaptive preprocessing.

When the quantized data are entropy-coded, the MSE-based signal distortion is proportional to the variances of DCT coefficients [20], [21] except for very low bit rates (large distortion). It is expected that reduction of the variance of residue signal leads to reduction of the variances of DCT coefficients, and consequently leads to reduction of D_1, since the DCT coefficients are the decomposition of residue signal in frequency-domain.

As can be seen from (12), the larger λ, the closer $\tilde{R}(x,y)$ will be to $\overline{R_B}$. Therefore, D_1 can be reduced by compressing the residues preprocessed with a larger λ. On the other hand, a larger λ will result in larger D_2. As the result, a tradeoff exists between D_1 and D_2, and an optimal value for λ exists for minimization of the overall distortion $D = D_1 + D_2$.

IV. Parameter Determination

D_1 and D_2 are derived as in Appendixes I and II. To illustrate the relationship between λ and the distortion (D_1, D_2, and D), 20 video sequences are used in experiments. These sequences are from video quality expert group (VQEG) [28] (listed in Table I), and span a spectrum of various motion, zooming, color, and texture. The first ten sequences are with frame rate of 25 f/s and resolution of 720 × 576 pixels, and the rest are with frame rate of 30 f/s and resolution of 720 × 480 pixels. Fig. 5 shows D_1, D_2, and D as the functions of λ for these 20 video sequences compressed at bit rate of 5 Mb/s. There exists a minimum for each of these convex curves, and as will be shown in Section V, the positions of the minima determine how well the proposed preprocessing can perform.

With (21), (26), and (28) in the Appendixes, the overall distortion D is computed as

$$\begin{aligned} D &= D_1 + D_2 \\ &= \epsilon^2 \cdot e^{-\alpha \cdot b} \cdot \sigma_{\mathbf{R}}^2 \left[1 - \psi \cdot \left(\sigma_{\mathbf{R}}^2\right)^{-\phi} \cdot \ln\left(\lambda^2 \cdot P_{\mathrm{JND}}^2 + 1\right)\right] \\ &\quad + \omega \cdot \lambda^2 \cdot P_{\mathbf{JND}}^2 \cdot \ln\left(\sigma_{\mathbf{R}}^2 + 1\right) \end{aligned} \tag{13}$$

$$\tilde{R}(x,y) = \begin{cases} R(x,y) + \lambda \cdot \mathrm{JND}(x,y), & \text{if } R(x,y) - \overline{R_B} < -\lambda \cdot \mathrm{JND}(x,y) \\ \overline{R_B}, & \text{else if } |R(x,y) - \overline{R_B}| \leq \lambda \cdot \mathrm{JND}(x,y) \\ R(x,y) - \lambda \cdot \mathrm{JND}(x,y), & \text{otherwise} \end{cases} \tag{12}$$

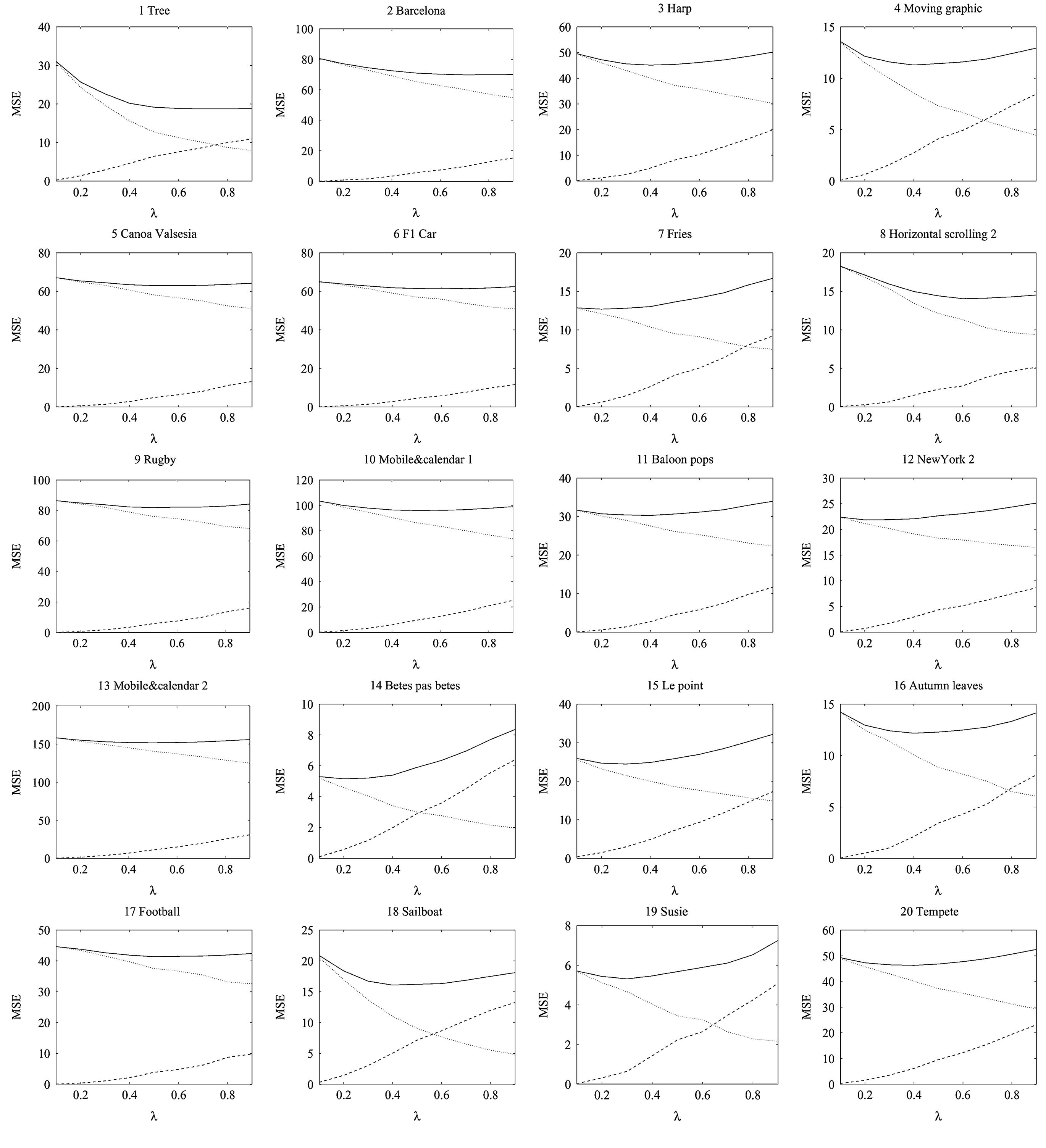


Fig. 5. Tradeoff between D_1 and D_2 with the relationship of λ for 20 sequences compressed at 5 Mb/s (D_1, D_2, and $D = D_1 + D_2$ are illustrated by the doted, the dashed, and the solid lines, respectively).

where b is the number of bits assigned for coding the residue signal, $\sigma_{\mathbf{R}}^2$ is the variance of residue signal $\mathbf{R}$, $P_{\mathbf{JND}}^2$ is defined in (23), $\epsilon^2 = 1.2, \alpha = 1.386, \psi = 0.4, \phi = 0.3$, and $\omega = 0.1$.

A minimum exists with D versus λ curves (as illustrated in the examples of Fig. 5). Let $(\partial D/\partial \lambda) = 0$, we have

$$(\lambda^*)^2 = \Lambda \tag{14}$$

with

$$\Lambda = \frac{1}{P_{\mathbf{JND}}^2} \cdot \left(\frac{\psi \cdot \epsilon^2 \cdot e^{-\alpha \cdot \bar{b}} \cdot \left(\sigma_{\mathbf{R}}^2\right)^{1-\phi}}{\omega \cdot \ln\left(\sigma_{\mathbf{R}}^2 + 1\right)} - 1 \right). \tag{15}$$

Considering $\lambda \in [0, 1]$, we finally get

$$\lambda^* = \begin{cases} 0, & \Lambda < 0 \\ \sqrt{\Lambda}, & 0 \leq \Lambda \leq 1 \\ 1, & \Lambda > 1. \end{cases} \tag{16}$$

TABLE I
THE OVERALL SIGNAL DISTORTION (PSNR) AND PERCEPTUAL DISTORTION RATING SCORES (DMOS) OF THE RECONSTRUCTED FRAMES ENCODED BY THE ORIGINAL MPEG-2 TM5 CODER AND THE CODER WITH THE PROPOSED JND-ADAPTIVE PRE-PROCESSOR, RESPECTIVELY.

Video sequence	PSNR (dB)			DMOS		
	TM5	the proposed scheme	Gain	TM5	the proposed scheme	Quality Gain
1.Tree	35.124	36.961	1.837	6.4	4.3	2.1
2.Barcelona	30.281	30.862	0.581	38.5	24.3	14.2
3.Harp	31.994	32.411	0.417	31.0	21.4	9.6
4.Moving graphic	37.645	38.650	1.005	31.5	22.4	9.1
5.Canoa Valsesia	29.261	29.481	0.221	64.8	56.5	8.3
6.F1 Car	30.637	31.005	0.368	49.5	41.9	7.6
7.Fries	37.279	37.192	-0.087	30.0	20.3	9.7
8.Horizontal scrolling2	30.857	31.812	0.955	41.0	29.9	11.1
9.Rugby	28.401	28.667	0.266	37.9	34.3	3.6
10.Mobile&calendar 1	29.453	29.880	0.428	35.6	29.7	5.9
11.Baloon-pops	33.586	33.796	0.210	34.2	26.7	7.5
12.NewYork 2	39.597	39.930	0.334	22.9	9.0	13.9
13.Mobile&Calendar 2	27.479	27.779	0.299	39.5	26.4	13.1
14.Betes pas betes	42.451	42.667	0.216	13.6	6.7	6.9
15.Le point	34.247	34.536	0.289	46.7	36.3	10.4
16.Autumn leaves	37.677	38.422	0.745	13.3	5.6	7.7
17.Football	31.664	31.952	0.288	37.3	32.5	4.8
18.Sailboat	36.222	37.371	1.149	10.2	6.4	3.8
19.Susie	41.212	41.446	0.234	11.1	6.2	4.9
20.Tempete	32.118	32.471	0.354	13.3	9.7	3.6
Average over 20 Sequences	-	-	0.505	-	-	7.9

V. OVERALL PERFORMANCE

The proposed JND-based preprocessing scheme has been implemented by incorporating the JND estimator and the JND-adaptive residue preprocessor into the MPEG-2 Test Model 5 (TM5) [29] coder. The 20 test sequences listed in Table I are used for experiments, since they represent a good diversity of visual signal. The target bit rate is set as 5 Mb/s. The GOP setting is $N = 12$ and $M = 3$ (both P and B frames are present).

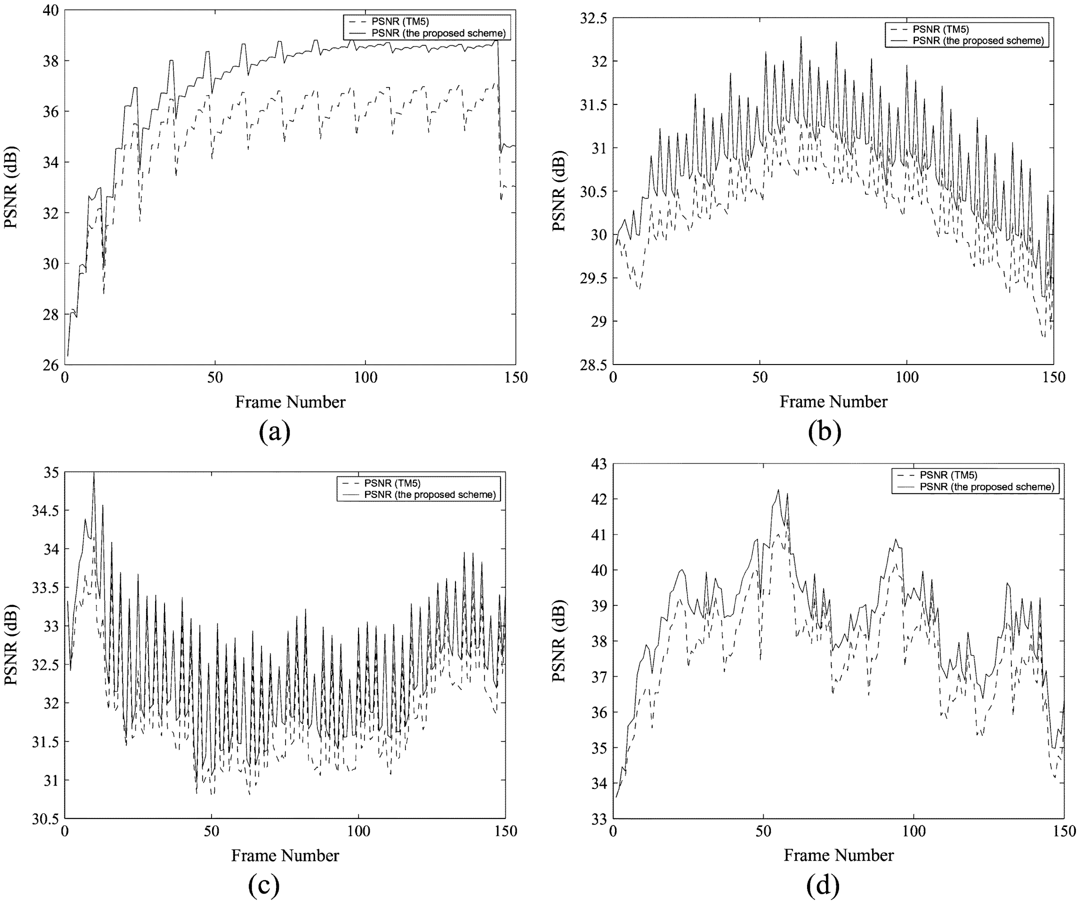
A. PSNR Comparison

Fig. 6 illustrates the overall PSNR comparison of the first four video sequences under test, between the original TM5 coder and the improved method with the proposed scheme. It can be seen that PSNR are improved with the proposed scheme. The results for all 20 video sequences are listed in Table I. The PSNR is significantly improved for all other sequences except for the "Fries" sequence, in which a slight PSNR loss of 0.087 dB occurs; an average PSNR gain of 0.505 dB is achieved by the proposed scheme in comparison with the original TM5 coder.

There is close correlation between the PSNR gain and the location of the minimum in D-versus-λ curve in Fig. 5. If the minimum occurs around $\lambda = 1$ for a sequence, the resultant PSNR gain is substantial; if the minimum occurs close to $\lambda = 0$ for a sequence, the resultant PSNR gain is small; and there are a number of cases in which λ and the resultant PSNR gain lie in between. This demonstrates that the proposed scheme is able to determine λ^* adaptively with different visual signal contents.

B. Subjective Quality Evaluation

In order to further confirm the coding quality improvement by the proposed scheme, we performed subjective quality evaluation. Double stimulus continuous quality scale (DSCQS) method, as in Rec. ITU-R BT.500 [30], was used to evaluate the subjective quality of a decoded sequence relative to its original sequence. A decoded sequence is obtained with the MPEG-2 TM5 coder or the improved coder with the proposed JND-adaptive scheme. Each display session for an original sequence and an associated decoded sequence is: *Video Sequence 1, two seconds of grey screen, Video Sequence 2, two seconds of grey screen*. The display repeats twice before the viewers are requested to vote for the quality of each sequence. Both the display order of the sequences in a session and the order of the 20 test sequences were randomized for viewers. The Mean Opinion Score (MOS) scales for viewers to vote for the quality after viewing are: Excellent (100–80), Good (80–60), Fair (60–40), Poor (40–20), and Bad (20–0).



(a)

(b)

(c)

(d)

Fig. 6. PSNR comparison of the sequences compressed by the MPEG-2 TM5 coder and the coder with the proposed JND-adaptive preprocessor individually: (a) *Tree*, (b) *Barcelona*, (c) *Harp*, (d) *Moving graphic*.

Eleven observers (five of them are with average image processing knowledge and the rest are naive) were involved in the experiments. Their eyesight is either normal or has been corrected to be normal with spectacles. The subjective visual quality assessment was performed in a typical laboratory environment with normal fluorescent ceiling light, using a 21″ EIZO T965 professional color monitor with resolution of 1600 × 1200. The viewing distance is approximately six times of the image height.

Difference mean opinion scores (DMOS) are calculated as the difference of MOSs between the original video and the processed video. The smaller the DMOS is, the higher perceptual quality of the processed video has when compared with the original video. Table I compares the averaged DMOSs over the all 11 observers for the 20 sequences. From the table, we can see that the subjective rating is consistently better for the decoded sequences with the proposed scheme, and an average subjective quality gain of 7.9 measured in DMOS is achieved by the proposed scheme.

VI. Conclusion

Based on an image-domain JND profile, a perceptually adaptive preprocessor for motion-compensated residues in video encoder has been proposed, demonstrated with a wide variety of test video sequences and subjective viewing. The proposed technique can be applied to any standardized video coding scheme,

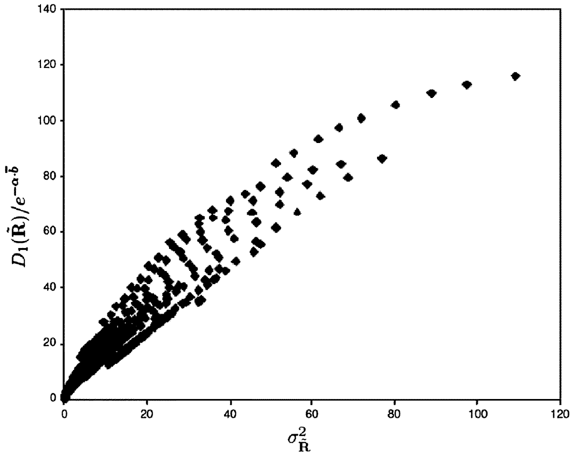


Fig. 7. $D_1/e^{-\alpha\cdot\bar{b}}$ as a function of $\sigma^2_{\tilde{\mathbf{R}}}$.

such as H.261/263 and MPEG-1/2/4. It improves both objective coding quality (PSNR) and perceptual quality of the decoded images for a given bit rate. The major contributions of this paper include: 1) a new image-domain JND estimator has been devised with our NAMM for color images/video, allowing a more accurate visibility threshold; 2) the JND profile has been incorporated into a residue signal preprocessor to achieve effective reduction of the variance for the residue signal after motion compensation. The resultant video coding process has an

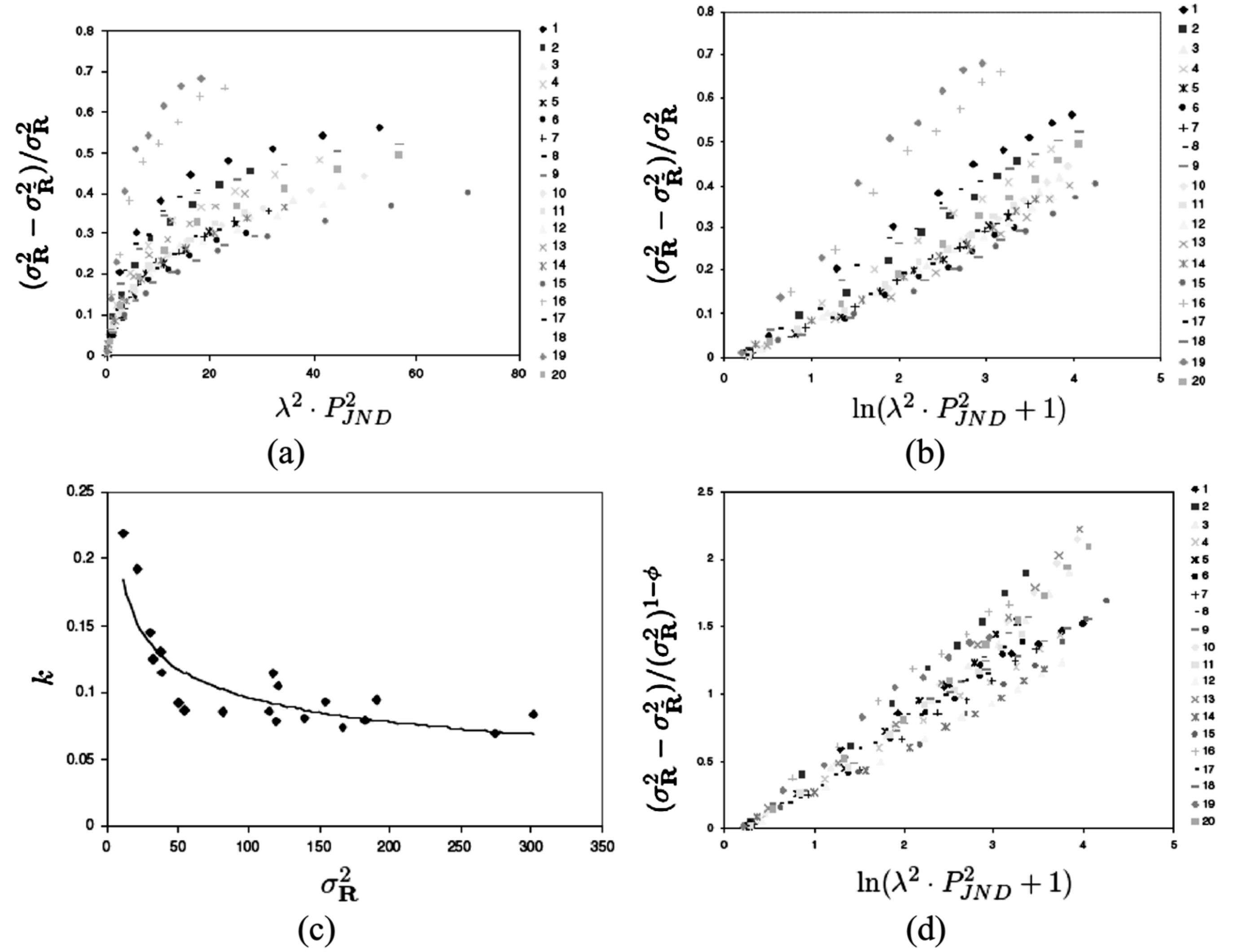


Fig. 8. Modeling $\sigma^2_{\tilde{\mathbf{R}}}$ (the indexes in the legend correspond to the sequence number in Table I: (a) $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/\sigma^2_{\mathbf{R}}$ as a function of $\lambda^2 \cdot P^2_{\mathrm{JND}}$, (b) $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/\sigma^2_{\mathbf{R}}$ as a function of $\ln(\lambda^2 \cdot P^2_{\mathrm{JND}} + 1)$, (c) Data fitting for determining k, (d) $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/(\sigma^2_{\mathbf{R}})^{1-\phi}$ as a function of $\ln(\lambda^2 \cdot P^2_{\mathrm{JND}} + 1)$.

extra designing option of quality improvement, besides quantization; and 3) a method of determining the optimum parameter for the preprocessor has been also devised for improvement of both PSNR and perceptual quality.

Appendix I
Modelling D_1

It is known that, except for very low bit rate (large distortion), the number of bits (b) versus MSE-based signal distortion (D_0) relation for a zero-mean independent and identically distributed (i.i.d.) source (i.e., the DCT coefficients, $\mathbf{Z}$) can be approximated by the following formula when the quantized data are entropy-coded [31], [20], [32], [21]:

$$b(\Delta) = \frac{1}{\alpha} \log_e \left(\epsilon^2 \cdot \beta \cdot \frac{\sigma^2_{\mathbf{Z}}}{\Delta^2} \right) \tag{17}$$

$$D_0(\Delta) = \frac{\Delta^2}{\beta}. \tag{18}$$

Combining the two equations above

$$D_0(b) = \epsilon^2 e^{-\alpha \cdot b} \cdot \sigma^2_{\mathbf{Z}} \tag{19}$$

where Δ is the quantization step size; $\beta = 12$ and $\alpha = 1.386 (= 2/\log_2 e)$ for uniform, Gaussian, and Laplacian distributions; $\sigma^2_{\mathbf{Z}}$ is the signal variance; ϵ^2 is source-dependent and

$$\epsilon^2 \simeq \begin{cases} 1, & \text{for uniform distribution} \\ 1.4, & \text{for Gaussian distribution} \\ 1.2, & \text{for Laplacian distribution.} \end{cases} \tag{20}$$

Since the residue signal is the synthesis of DCT coefficients, the variance of the residue signal affects signal coding distortion in a similar way, i.e., the relationship between D_1 and the signal variance $\sigma^2_{\tilde{\mathbf{R}}}$ can be expressed as below, in analogy with (19)

$$D_1 = \epsilon^2 \cdot e^{-\alpha \cdot b} \cdot \sigma^2_{\tilde{\mathbf{R}}} \tag{21}$$

where $\sigma^2_{\tilde{\mathbf{R}}}$ represents the variance of the preprocessed residue signal $\tilde{\mathbf{R}}$; $\epsilon^2 = 1.2$ and $\alpha = 1.386$ since $\tilde{R}_{x,y}$ follows Laplacian distribution.

The statistical studies for a large number of video sequences have shown that (21) is valid and $\tilde{R}_{x,y}$ follows Laplacian distribution. Fig. 7 plots $D_1/e^{-\alpha \cdot b}$ versus $\sigma^2_{\tilde{\mathbf{R}}}$ for the 20 sequences

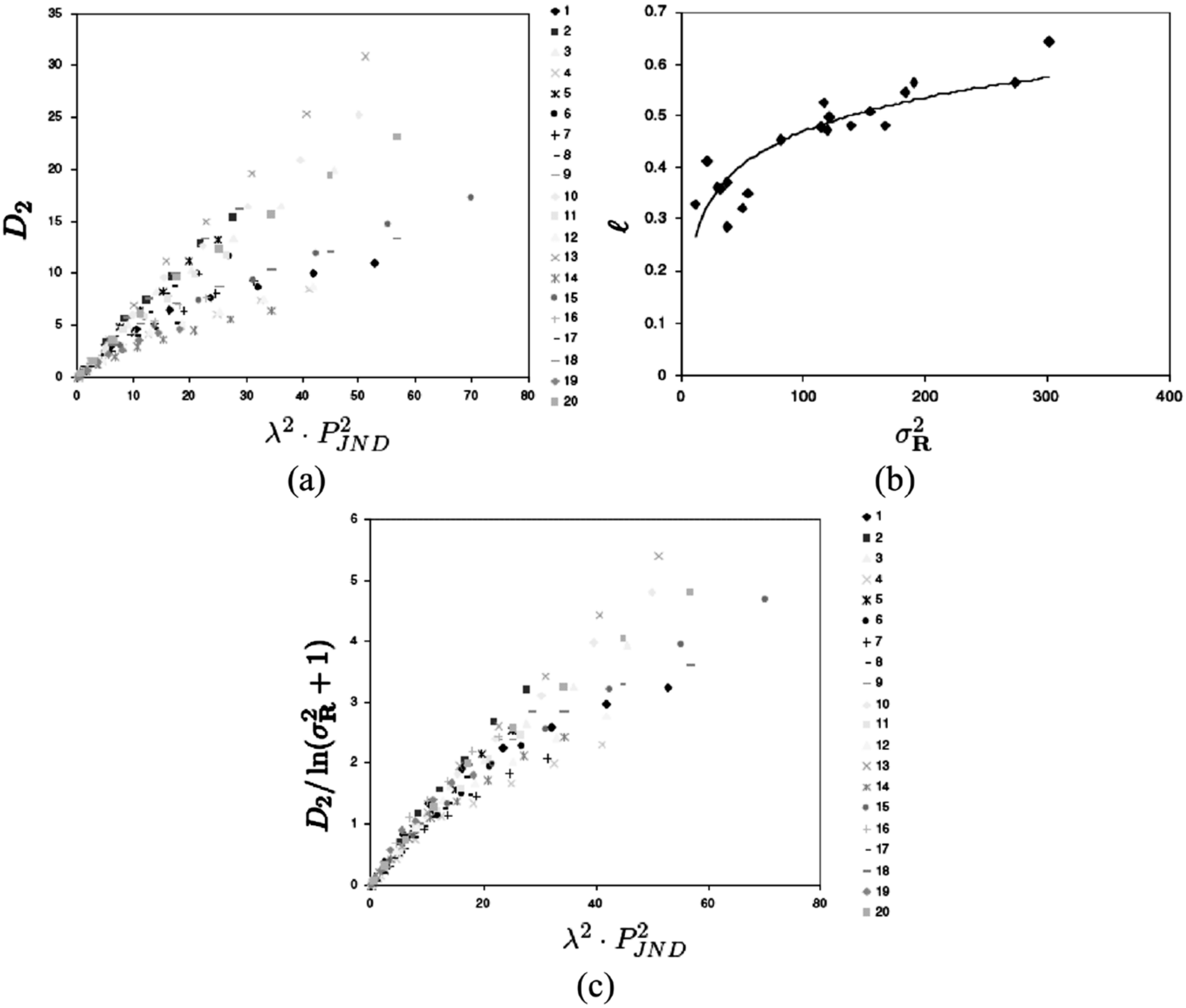


Fig. 9. Modeling D_2 (the indexes in the legend correspond to the sequence number in Table I: (a) D_2 as a function of $\lambda^2 \cdot P^2_{\text{JND}}$, (b) Data fitting for determining ℓ, (c) $D_2/\ln(\sigma^2_{\mathbf{R}} + 1)$ as a function of $\lambda^2 \cdot P^2_{\text{JND}}$.

compressed over a wide bit rate range from 0.02 to 0.4 bit per residue. By linearly regressing the data in Fig. 7, we found that (21) fits the data well with correlation coefficient of 0.93.

To determine D_1 in (21), we need to estimate $\sigma^2_{\tilde{\mathbf{R}}}$, which can be mathematically expressed by

$$\sigma^2_{\tilde{\mathbf{R}}} = \int_{-\infty}^{+\infty} (\tilde{R} - \overline{\tilde{R}})^2 p(\tilde{R}) d\tilde{R} \tag{22}$$

where $p(\tilde{R})$ is probability density function of $\tilde{R}$.

Due to the difficulty in deriving the closed-form expression for $\sigma^2_{\tilde{\mathbf{R}}}$ from (12) and (22), a statistical model is, therefore, to be derived. Let P^2_{JND} be the average power of JND profile for the frame

$$P^2_{\text{JND}} = \frac{1}{X \cdot Y} \sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} (\text{JND}(x,y))^2. \tag{23}$$

Since $\tilde{R}(x,y)$ depends on $\lambda \cdot \text{JND}(x,y)$ and $R(x,y)$ according to (12), $\sigma^2_{\tilde{\mathbf{R}}}$ is related to $\lambda^2 \cdot P^2_{\text{JND}}$ and $\sigma^2_{\mathbf{R}}$. Two additional points have been observed toward the relationship between $\sigma^2_{\tilde{\mathbf{R}}}$ and $\sigma^2_{\mathbf{R}}$: 1) $\sigma^2_{\tilde{\mathbf{R}}} = \sigma^2_{\mathbf{R}}$ if $\lambda = 0$ when the JND-adaptive residue preprocessor is passed by; and 2) the variance reduction, $\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}}$, is proportional to $\sigma^2_{\mathbf{R}}$, i.e., residue signal with larger $\sigma^2_{\mathbf{R}}$ leads to more significant variance reduction for a given λ. Therefore, the relationship between $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/\sigma^2_{\mathbf{R}}$ and $\lambda^2 \cdot P^2_{\text{JND}}$ is to be exploited.

$(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/\sigma^2_{\mathbf{R}}$ represents the normalized variance reduction due to JND-adaptive residue preprocessing. We processed each sequence in the VQEG data set with $\lambda = 0.1, 0.2, \ldots, 0.9$ at 5 Mb/s. Fig. 8(a) plots $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/\sigma^2_{\mathbf{R}}$ as a function of $\lambda^2 \cdot P^2_{\text{JND}}$, and there are nine points for each sequence. It is observed that there is approximately logarithmic relationship between $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/\sigma^2_{\mathbf{R}}$ and $\lambda^2 \cdot P^2_{\text{JND}}$. Considering the fact that $\sigma^2_{\tilde{\mathbf{R}}} = \sigma^2_{\mathbf{R}}$ when $\lambda = 0$, we use $\ln(\lambda^2 \cdot P^2_{\text{JND}} + 1)$ to replace $\lambda^2 \cdot P^2_{\text{JND}}$ as a variable for modeling $\sigma^2_{\tilde{\mathbf{R}}}$. Fig. 8(b) shows that $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/\sigma^2_{\mathbf{R}}$

are linear functions of $\ln(\lambda^2 \cdot P^2_{\mathrm{JND}} + 1)$ with different slopes (k) for different sequences. We further find that the slope mainly depends on $\sigma^2_{\mathbf{R}}$ of the sequences. Fig. 8(c) depicts the relationship between k and $\sigma^2_{\mathbf{R}}$ for the 20 sequences. By fitting the 20 points in Fig. 8(c), k is determined as

$$k = \psi \cdot \left(\sigma^2_{\mathbf{R}}\right)^{-\phi} \tag{24}$$

where ψ and ϕ are constants ($\psi = 0.4$ and $\phi = 0.3$ in this work).

Fig. 8(d) plots $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/(\sigma^2_{\mathbf{R}})^{1-\phi}$ against $\ln(\lambda^2 \cdot P^2_{\mathrm{JND}} + 1)$, and in comparison with Fig. 8(b), it can be seen that $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/(\sigma^2_{\mathbf{R}})^{1-\phi}$ has much less dependency with source than $(\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}})/\sigma^2_{\mathbf{R}}$. $\sigma^2_{\tilde{\mathbf{R}}}$ is, therefore, modeled as

$$\frac{\sigma^2_{\mathbf{R}} - \sigma^2_{\tilde{\mathbf{R}}}}{\sigma^2_{\mathbf{R}}} = \psi \cdot \left(\sigma^2_{\mathbf{R}}\right)^{-\phi} \cdot \ln\left(\lambda^2 \cdot P^2_{\mathrm{JND}} + 1\right) \tag{25}$$

or

$$\sigma^2_{\tilde{\mathbf{R}}} = \sigma^2_{\mathbf{R}} \left[1 - \psi \cdot \left(\sigma^2_{\mathbf{R}}\right)^{-\phi} \cdot \ln\left(\lambda^2 \cdot P^2_{\mathrm{JND}} + 1\right)\right] \tag{26}$$

APPENDIX II
MODELLING D_2

From (12), it can be judged that D_2 should be proportional to $\lambda^2 \cdot P^2_{\mathrm{JND}}$, except for the range of $-\lambda \cdot \mathrm{JND}(x,y) \leq R(x,y) - \overline{R_B} \leq \lambda \cdot \mathrm{JND}(x,y)$, in which nonlinear operation is involved. Fig. 9(a) illustrates the correlation between D_2 and $\lambda^2 \cdot P^2_{\mathrm{JND}}$ with different slopes (ℓ) for different sequences. It has been found that ℓ mainly depends on $\sigma^2_{\mathbf{R}}$ of each sequence. Therefore a fitting process has been performed in Fig. 9(b) for $(\ell, \sigma^2_{\mathbf{R}})$. With the additional consideration that $D_2 = 0$ when $\lambda = 0$, we get

$$\ell = \omega \cdot \ln\left(\sigma^2_{\mathbf{R}} + 1\right) \tag{27}$$

where ω is a constant ($\omega = 0.1$ in this work).

In Fig. 9(c), the correlation between $D_2/\ln(\sigma^2_{\mathbf{R}} + 1)$ and $\lambda^2 \cdot P^2_{\mathrm{JND}}$ is shown for all 20 sequences in the VQEG data set. D_2 is, therefore, modeled as follows:

$$D_2 = \omega \cdot \lambda^2 \cdot P^2_{\mathbf{JND}} \cdot \ln\left(\sigma^2_{\mathbf{R}} + 1\right). \tag{28}$$

REFERENCES

[1] N. S. Jayant, J. D. Johnston, and R. J. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, no. 10, pp. 1385–1422, Oct. 1993.

[2] A. J. Ahumada and H. A. Peterson, "Luminance-model-based dct quantization for color image compression," in *Proc. SPIE Int. Conf. Human Vision, Visual Processing and Digital Display—III*, 1992, pp. 365–374.

[3] R. J. Safrenek and J. D. Johnson, "A perceptually tuned sub-band image coder with image dependent quantization and postquantization data compression," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP89)*, 1989, pp. 1945–1948.

[4] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proc. SPIE Int. Conf. Human Vision, Visual Processing and Digital Display IV*, vol. 1913, 1993, pp. 202–216.

[5] 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.

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

[7] ——, "Adaptive image coding with perceptual distortion control," *IEEE Trans. Image Process.*, vol. 11, no. 3, pp. 213–222, Mar. 2002.

[8] A. P. Bradley, "A wavelet visible difference predictor," *IEEE Trans. Image Process.*, vol. 8, no. 5, pp. 717–730, May 1999.

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

[10] T. D. Tran and R. Safranek, "A locally adaptive perceptual masking threshold for image coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP06)*, 1996, pp. 1882–1885.

[11] H. Y. Tong and A. N. Venetsanopoulos, "A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking," in *Proc. IEEE Int. Conf. Image Processing (ICIP98)*, 1998, pp. 428–431.

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

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

[14] Y. J. Chiu and T. Berger, "A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancement," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 4, pp. 438–450, Apr. 1999.

[15] *Video codec for audiovisual services at p × 64 kbits/s*, 1993. ITU-T Recommendation H.261, ITU.

[16] *Video Coding for Low Bitrate Communication*, 1998. ITU-T Recommendation H.263 Version 2, ITU.

[17] *Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5 Mb/s (MPEG-1), Part 2: Video*, 1993. ISO/IEC JTC 1/SC 29/WG 11/11 172-2, ISO.

[18] *Generic Coding of Moving Pictures and Associated Audio (MPEG-2), Part 2: Video*, 1995. ISO/IEC JTC 1/SC 29/WG 11/13 818-2, ISO.

[19] *Information Technology-Coding of Audio-Visual Objects (MPEG-4). Part 2: Visual*, 2001. ISO/IEC JTC 1/SC 29/WG 11/14496, ISO.

[20] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[21] H.-M. Hang and J.-J. Chen, "Source model for transform video coder and its application—Part I: Fundamental theory," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 287–298, Feb. 1997.

[22] X. Yang, W. Lin, Z. Lu, E. Ong, and S. Yao, "Just-noticeable-distortion profile with nonlinear additivity model for perceptual masking in color images," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP2003)*, vol. 3, Apr. 2003, pp. 609–612.

[23] H. C. Nothdurft, "Salience from feature contrast: Additivity across dimensions," *Vis. Res.*, vol. 40, pp. 1183–1201, 2000.

[24] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Process.*, vol. 70, pp. 177–200, 1998.

[25] D. Marr, *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. San Francisco, CA: Freeman, 1982.

[26] J. H. Elder and R. M. Goldberg, "Local scale control for edge detection and blur estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 7, pp. 699–716, Jul. 1998.

[27] C. John, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, no. 6, pp. 679–698, Jun. 1986.

[28] (2000, Mar.) Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Qualtiy Assessment. [Online]. Available: www.its.bldrdoc.gov/vqeg/

[29] *MPEG-2 Test Model 5*, Apr. 1993. ISO/IEC JTC1/SC29 WG11, ISO.

[30] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, 2000. ITU-R Rec. BT. 500-10, ITU-R, ITU.

[31] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Theory*, vol. IT-14, no. 9, pp. 676–683, Sep. 1968.

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

[33] S. R. Smoot and L. A. Rowe, "Study of dct coefficient distributions," in *Proc. SPIE Human Vision and Electronic Imaging*, vol. 2657, 1996, pp. 403–411.

Xiaokang Yang (M'00) received the B.Sci. degree from Xiamen University, Xiamen, China, in 1994, the M.Eng. degree from Chinese Academy of Sciences, Beijing, China, in 1997, and the Ph.D. degree from Shanghai Jiaotong University, Shanghai, China, in 2000.

He is currently an Associate Professor in the Institute of Image Communication and Information Processing, Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. From April 2002 to October 2004, he was a Research Scientist in the Institute for Infocomm Research, Singapore. From September 2000 to March 2002, he worked as a Research Fellow in the Centre for Signal Processing, Nanyang Technological University, Singapore. He has published over 50 refereed papers. His current research interests include scalable video coding, video transmission over networks, video quality assessment, digital television, and pattern recognition.

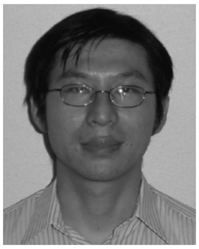
Dr. Yang received the Best Young Investigator Paper Award at IS&T/SPIE International Conference on Video Communication and Image Processing (VCIP2003). He is currently a Member of Visual Signal Processing and Communications Technical Committee of the IEEE Circuits and Systems Society.

Weisi Lin (M'92-SM'98) received the B.Sc. and M.Sc. degrees from Zhongshan University, Guangzhou, China in 1982 and 1985, respectively. He received the Ph.D. degree from King's College, London University, London, U.K., in 1992.

He had taught and/or researched in Zhongshan University (China), Shantou University (China), Bath University (U.K.), and the National University of Singapore and Institute of Microelectronics (Singapore). He has been the Project Leader of a number of successfully delivered projects in development of digital multimedia related technologies since 1997. He is currently an Associate Lead Scientist in the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His current research interests include perceptual visual distortion metrics, perceptual video coding, and multimedia signal processing.

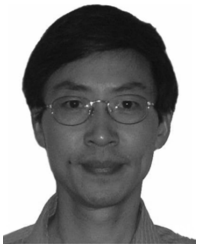
Zhongkang Lu (S'95-M'99) received the B.Eng. degree in biomedical engineering from Southeast University, Nanjing, China, in 1993, and the M.Eng. and Ph.D. degrees in electrical engineering from Shanghai Jiaotong University, Shanghai, China, in 1996 and 1999, respectively.

Between 1996 and 1998, he was an exchange student in the Department of Electronic and Information Engineering, Hong Kong Polytechnic University, Hong Kong. From 1999 to 2001, he was a Research Fellow in the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore. Currently, he is a Research Scientist in Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore. His research interests include perceptual visual signal processing, pattern recognition, and computer vision.



EePing Ong (M'02) received the B.Eng. and Ph.D. degrees in electronics and electrical engineering from the University of Birmingham, Birminhgam, U.K., in 1993 and 1997, respectively.

From 1997 to 2001, he was with the Institute of Microelectronics, Singapore. Thereafter, he joined the Centre for Signal Processing, Nanyang Technological University, Singapore. Since 2002, he has been with the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore, where he is currently a Research Scientist. His research interests include optical flow, motion estimation, video object segmentation and tracking, perceptual image/video quality metrics, perceptual image/video coding, and multimedia signal processing.



Susu Yao (M'97) received the Ph.D. degree from the National University of Defense Technology, Changsha, China in 1993.

He was a Visiting Scholar in Heriot-Watt University, Edinburgh, U.K., from 1991 to 1993. From 1993 to 1995, he was a Postdoctoral Fellow and Associate Professor in Southeast University, Nanjing, China. Since 1996, he has been an Associate Professor and Full Professor in Nanjing Institute for Communication Engineering, Nanjing, China. In 2000, he joined the Centre for Signal Processing in Nanyang Technological University, Singapore. His main areas of research interest are image and video compression, wavelet transform, soft computing, image, and video postprocessing, perceptual image quality metrics, for which he has published more than 40 papers. He is currently an Associate Lead Scientist in the Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore.