Available online at www.sciencedirect.com
SCIENCE DIRECT®
Signal Processing: *Image Communication* 20 (2005) 662–680
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Just noticeable distortion model and its applications in video coding

X.K. Yang[a,*], W.S. Ling[b], Z.K. Lu[b], E.P. Ong[b], S.S. Yao[b]

[a]*Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200030, China*
[b]*Institute for Infocomm Research, Agency for Science, Technology and Research, Singapore 119613, Singapore*

Received 21 September 2004; accepted 25 April 2005

Abstract

We explore a new perceptually-adaptive video coding (PVC) scheme for hybrid video compression, in order to achieve better perceptual coding quality and operational efficiency. A new just noticeable distortion (JND) estimator for color video is first devised in the image domain. How to efficiently integrate masking effects together is a key issue of JND modelling. We integrate spatial masking factors with the nonlinear additivity model for masking (NAMM). The JND estimator applies to all color components and accounts for the compound impact of luminance masking, texture masking and temporal masking. Extensive subjective viewing confirms that it is capable of determining a more accurate visibility threshold that is close to the actual JND bound in human eyes. Secondly, the image-domain JND profile is incorporated into hybrid video encoding via the JND-adaptive motion estimation and residue filtering process. The scheme works with any prevalent video coding standards and various motion estimation strategies. To demonstrate the effectiveness of the proposed scheme, it has been implemented in the MPEG-2 TM5 coder and demonstrated to achieve average improvement of over 18% in motion estimation efficiency, 0.6 dB in average peak signal-to perceptual-noise ratio (PSPNR) and most remarkably, 0.17 dB in the objective coding quality measure (PSNR) on average. Theoretical explanation is presented for the improvement on the objective coding quality measure. With the JND-based motion estimation and residue filtering process, hybrid video encoding can be more efficient and the use of bits is optimized for visual quality.

Keywords: Video coding; Perception; Human visual system; Just noticeable distortion

1. Introduction

Since the ultimate receiver of most decompressed video signal is the human visual system (HVS), the goal of video compression and coding should therefore be to achieve the lowest bit rate

*Corresponding author.
E-mail addresses: xkyang@sjtu.edu.cn, xkyang@ieee.org (X.K. Yang).

0923-5965/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2005.04.001

for signal representation at a 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. However, the traditional fidelity criteria, such as mean square error (MSE) and the related peak signal-to-noise ratio (PSNR), do not reflect perceptual distortion well [11].

It is known that 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 [22]. Obviously, any un-noticeable signal difference need not to be coded in the bitstream and reflected in the distortion measure. An appropriate (even imperfect) JND model can significantly help to improve the performance of video coding algorithms. Several methods for finding a JND model have been proposed, in subbands (i.e., discrete cosine transform (DCT) or wavelet domain) [1,2,31,33–36,14,15,26,8] and image-domain [6–8].

Perceptual coding has so far been focused mainly on determination of proper quantization steps for image coding with subband JND [31,35,36,15,14,2,8,7]. A few attempts have been made to non-standard video coding [6,26]. In [6], an image-domain JND profile has been used as the threshold for inter-frame replenishment with low-motion (like head-and-shoulder) scenes in low bit-rate videophony. In [26], a subband JND model 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 new perceptually-adaptive video coding (PVC) scheme in which the image-domain JND profile is used in the motion prediction loop for the standardized hybrid video coding (cascading of motion estimation and DCT), such as H.261/263 [20,21], MPEG-1/2/4 [17–19] and the emerging JVT video coding [23].

Because of the importance of accurate JND profile to this work, a new model for JND determination will be firstly exploited. Due to the intended application in motion estimation, image-domain JND is of our interest in this paper. Many algorithms for subband JND exist (e.g., [1,31,33–35,14]), but there are relatively fewer reports for JND in image domain. In principle, JND in image-domain can be viewed as the compound effect of all subbands. However, in a practical point of view, it is better to estimate the image-domain JND directly, for the sake of operating efficiency. Luminance adaptation and texture masking are the major consideration in spatial JND. In Chou's work [7,8], texture masking was determined with the average background luminance and the average luminance difference around the pixel. It was assumed that the resultant spatial JND is the dominant effect of texture masking or luminance adaptation (the Max(.) operation is used). Temporal masking (i.e., motion masking) was accounted for by evaluating the average inter-frame luminance difference. In Chiu's system for conditional video replenishing [6], the JND was formulated as the weighted sum of luminance adaptation threshold and the relative magnitude of a spatial/temporal activity measure. In both Chou's and Chiu's methods, only the JND for the luminance is considered. As the first part of this paper, a new formulae (the non-linear additivity model for masking—NAMM for short) for spatial JND in image-domain will be devised 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 effect, in analogy with the saliency effect from different stimuli in the recent vision research results [30]. The new model also accounts for the difference between edge regions and non-edge regions, since error is more visible in edge regions than in non-edge regions [9,11]; the formulae also applies to color components. Temporal masking is also incorporated in the case of a video sequence.

In the proposed PVC scheme, the JND profile calculated with the NAMM will be used in motion estimation and in deciding whether residues after motion compensation are to be coded. The basic ideas are: (1) If a pixel's difference in motion estimation is below the associated JND, it should be excluded in the sum of absolute difference (SAD) evaluation because of the invisibility of such difference. This leads to improvement of

perceptual coding quality and immediate savings of motion estimation effort. (2) For the same reason, after motion compensation, it is unnecessary to DCT-transform those residues below the JND so that some bits can be saved for better DCT coding of the residues above the JND under a chosen bit rate. The DCT coefficients of larger residues are more crucial to impact an objective fidelity measure, so the PVC scheme not only yields higher perceptual quality for the coded video but also has the tendency to increase the objective quality measure (like MSE or PSNR).

The rest of the paper is organized as follows. In Section 2, we present the NAMM model and its corresponding image-domain JND profile for color video. Relevant subjective test results for the NAMM are also presented. In Section 3, the proposed PVC scheme is presented, based on the JND-guided motion estimation and residue filtering. Theoretical explanation is also given for the potential increase in objective quality measure as a result of the residue filtering. The proposed scheme can be adopted in the encoding process in the prevalent compression standards, such as JPEG, MPEG-x and H.26x. In Section 4 the experimental results illustrating the performance of the PVC is given. Finally, we present the conclusions in Section 5.

2. Image-domain JND profile for color video

Let $I_\theta(x,y,t)$ denote the intensity of a pixel at (x,y) in the tth image frame for a color channel θ, and $1\leqslant x\leqslant M$, $1\leqslant y\leqslant N$, $\theta = Y,\ C_b,\ C_r$. The objective of this section is to determine $\mathrm{JND}_\theta(x,y,t)$, the JND value for the pixel at (x,y) of color channel θ. The spatial part, $\mathrm{JND}^S_\theta(x,y)$, is to be firstly considered with visual information within the frame (denoted as $I_\theta(x,y)$ hereinafter). $\mathrm{JND}_\theta(x,y,t)$ is then obtained by integrating temporal (interframe) masking with $\mathrm{JND}^S_\theta(x,y)$.

2.1. Spatial JND based on a nonlinear additivity model of masking (NAMM)

There are primarily two factors affecting the spatial JND in the image domain: (a) background luminance adaptation—the HVS is sensitive to luminance contrast rather than absolute luminance value, and an approximate curve on visibility threshold versus background luminance based on the experiments [8] is illustrated in Fig. 1 for digital images; (b) texture masking—the reduction of visibility of changes is caused by an increase in the texture non-uniformity in the neighborhood and, therefore, textured regions can hide more error than smooth areas.

Since these two types of masking co-exist in most images, how to effectively integrate them is an important issue in obtaining an accurate spatial JND profile. Although a useful method had been reported in [8], there are three drawbacks in the approach: the compound spatial masking effect is simplified as the maximum value between the visibility thresholds for luminance masking and texture masking; only the JND threshold for the luminance component in an image is considered; edge regions are not distinguished from the non-edge ones. We believe that: (i) the 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) the masking effect in chrominance channels could also be exploited to improve compression performance; (iii) a distinction of edge regions from smooth and textured regions avoid over-estimation of masking effect around the edge.

In analogy with the saliency effect from different stimuli in the recent vision research results [30], the

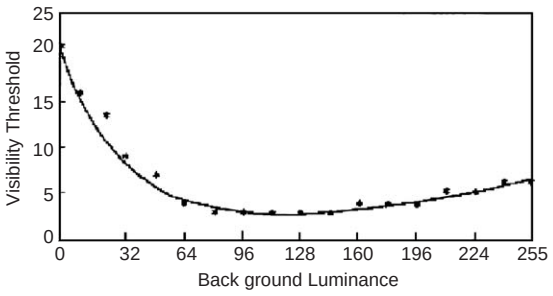

Fig. 1. Visibility threshold vs. background luminance.

nonlinear additivity model for masking (NAMM) is proposed to determine the visibility threshold for the overall masking effect:

$$T = \sum_{\iota=1}^{\aleph} T^{\iota} - \sum_{\iota=1}^{\aleph} \sum_{j=\iota+1}^{\aleph} C^{\iota,j}\gamma(T^{\iota}, T^{j}), \tag{1}$$

where T^{ι} is the visibility threshold due to the ιth masking factor; $C^{\iota,j}$ is the gain reduction factor due to overlapping between two masking factors; $\gamma(\cdot,\cdot)$ is an appropriate nonlinear function for evaluating the overlapping effect of two factors.

The spatial JND of each pixel can be given as an approximation of the nonlinear model described in (1):

$$\mathrm{JND}_{\theta}^{S}(x,y) = T^{l}(x,y) + T_{\theta}^{t}(x,y) - C_{\theta}^{lt} \times \min\{T^{l}(x,y), T_{\theta}^{t}(x,y)\}, \tag{2}$$

where $T^{l}(x,y)$ and $T_{\theta}^{t}(x,y)$ are the visibility thresholds for the two primary masking factors, background luminance adaptation and texture masking, respectively, at a color channel θ; and C_{θ}^{lt} accounts for the overlapping effect in masking.

The bigger value C_{θ}^{lt} takes, the more significant overlapping effect it represents. When $C_{\theta}^{lt} = 1$, there is maximum overlapping effect between $T^{l}(x,y)$ and $T_{\theta}^{t}(x,y)$. When $C_{\theta}^{lt} = 0$, there is not overlapping effect between $T^{l}(x,y)$ and $T_{\theta}^{t}(x,y)$. The real-world situations lie in between, so $0 < C_{\theta}^{lt} < 1$. In addition, the overlapping of $T^{l}(x,y)$ with $T_{Y}^{t}(x,y)$ is bigger than that with $T_{C_b}^{t}(x,y)$ (or $T_{C_r}^{t}(x,y)$), because both $T^{l}(x,y)$ and $T_{Y}^{t}(x,y)$ are derived from the Y component; i.e., $C_{Y}^{lt} > \max(C_{C_b}^{lt}, C_{C_r}^{lt})$. The valuation of C_{θ}^{lt} is also related to the viewing conditions (like lighting, display device, viewing distance, etc.). For this study, experiments have been conducted in a room illuminated by fluorescent ceiling lights (this is the typical conditions under which people would view digital image), and with a 21" EIZO T965 professional color monitor of resolution of 1600×1200. The viewing distance is approximately six times that of the image height. We valuate C_{θ}^{lt} as: $C_{Y}^{lt} = 0.3$, $C_{C_b}^{lt} = 0.25$ and $C_{C_r}^{lt} = 0.2$, for the experiments.

The JND estimator in [8] is a special case of the proposed NAMM, because if only the Y component is considered and $C_{Y}^{lt} = 1$, (2) becomes

$$\mathrm{JND}_{Y}^{S}(x,y)_{\text{simplified-I}} = \max\{T^{l}(x,y), T_{Y}^{t}(x,y)\}. \tag{3}$$

The parameter selection of $C_{Y}^{lt} < 1$ allows the compound effect for co-existence of luminance masking and texture masking to be reflected fully in (2).

The JND estimator in [6] is the special case of the proposed NAMM when $T^{l}(x,y)$ is considered as the major masking factor, i.e., $\min\{T^{l}(x,y), T_{Y}^{t}(x,y)\} \equiv T_{Y}^{t}(x,y)$. In this case, (2) becomes

$$\mathrm{JND}_{Y}^{S}(x,y)_{\text{simplified-II}} = T^{l}(x,y) + C' T_{Y}^{t}(x,y) \tag{4}$$

where $C' = 1 - C_{Y}^{lt}$. In [6], C′ is determined according to the magnitude of $T^{l}(x,y)$.

$T^{l}(x,y)$ can be determined with the visibility threshold curve in Fig. 1 [8], i.e.:

$$T^{l}(x,y) = \begin{cases} 17\left(1 - \sqrt{\frac{\overline{I}_{Y}(x,y)}{127}}\right) + 3 & \text{if } \overline{I}_{Y}(x,y) \leqslant 127, \\ \frac{3}{128}(\overline{I}_{Y}(x,y) - 127) + 3 & \text{otherwise,} \end{cases} \tag{5}$$

where $\overline{I}_{Y}(x,y)$ is the average background luminance at (x,y). The computation of $T_{\theta}^{t}(x,y)$ will be addressed in the following subsection.

2.2. Edge-adaptive visibility threshold of texture masking

For more accurate JND estimation, texture masking in edge and non-edge 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 textured regions due to the fact that edge structure attracts more attention from a typical HVS [9,11], and there is a substantial body of literature attesting to the importance of edges to primate perception (e.g., [27,10]). The proposed $T_{\theta}^{t}(x,y)$ therefore takes the edge information into account:

$$T_{\theta}^{t}(x,y) = \beta_{\theta} G_{\theta}(x,y) W_{\theta}(x,y), \tag{6}$$

where $G_{\theta}(x,y)$ denotes the maximal weighted average of gradients around the pixel at (x,y); β_{θ}

is a control parameter for each color channel; $\beta_Y < \min(\beta_{C_b}, \beta_{C_r})$, since the HVS is more sensitive for a difference in Y space than in C_b or C_r space. Under the viewing conditions described in the previous subsection, β_Y, β_{C_b} and β_{C_r} are set as 0.117, 0.65 and 0.45, respectively.

$G_\theta(x,y)$ is determined as

$$G_\theta(x,y) = \max_{k=1,2,3,4}\{\text{grad}_{\theta,k}(x,y)\} \tag{7}$$

with

$$\text{grad}_{\theta,k}(x,y) = \frac{1}{16}\sum_{i=1}^{5}\sum_{j=1}^{5} I_\theta(x-3+i, y-3+j) \times g_k(i,j), \tag{8}$$

where $g_k(i,j)$ are four directional high-pass filters for texture detection [7], as shown in Fig. 2.

In (6), $W_\theta(x,y)$ is an edge-related weight of the pixel at (x,y), and its corresponding matrix $\mathbf{W}_\theta$ is computed by edge detection followed with a Gaussian low-pass filter:

$$\mathbf{W}_\theta = Ł_\theta * \mathbf{h}, \tag{9}$$

where $Ł_Y$ is the edge map of Y component, detected by Canny detector [4] with threshold of 0.5, and $Ł_{C_b}$ and $Ł_{C_r}$ are identical, being the down-scaled version of $Ł_Y$, with element values of 0.1 and 1 for edge and non-edge pixels respectively; $\mathbf{h}$ is a $k \times k$ Gaussian low pass filter with standard deviation σ, to smooth $Ł_\theta$ and therefore avoid too dramatic changes for $\mathbf{W}_\theta$ in a small neighborhood. The standard deviation σ should be bigger than 0.5 to have smoothing effect, and it is chosen as 0.8 since only a modest degree of smoothing is needed in (9). A kernel size $k = 7$ is appropriate for the chosen σ.

2.3. Performance evaluation of NAMM

JND models can be tested by comparing $I_\theta(x,y)$ with its variation, $I_\theta^{\text{JND}}(x,y)$, which is formed via:

$$I_\theta^{\text{JND}}(x,y) = I_\theta(x,y) + s_{\text{rand}}(x,y,\theta)\text{JND}_\theta^S(x,y), \tag{10}$$

where $\text{JND}_\theta^S(x,y)$ is as given by (2), $s_{\text{rand}}(x,y,\theta)$ takes a value of either $+1$ or -1 at random in order to avoid fixed artifact patterns introduced to the image regarding x, y and θ. If $\text{JND}_\theta^S(x,y)$ is close to the JND bound in the HVS, it should take the largest possible value while perceptual distortion in the image constructed by (10) is minimized.

In the extreme case, (10) becomes (11) if a JND model simply generates random noise:

$$I_\theta^{\text{Non-JND}}(x,y) = I_\theta(x,y) + \alpha s_{\text{rand}}(x,y,\theta)\text{rand}(x,y,\theta), \tag{11}$$

where $\text{rand}(x,y,\theta)$ takes a random value in (0.0, 1.0) and α is a magnitude control factor.

In Fig. 3, the 512×512 image "Lena", part of which is shown in Fig. 3a, is processed by (10) using NAMM and using random noise injection (11), respectively, and the resulting images are shown in Fig. 3b and c. In comparison with the original image (Fig. 3a), a human viewer can discern significantly less quality difference in the JND-modified image (Fig. 3b) than in the random-noise injected version (Fig. 3c), although their PSNRs are similar (Fig. 3c actually being 0.08 dB higher); this is because the proposed JND profile effectively shapes the added noise to the perceptually insensitive regions. JND_Y^S in (2) and the JND derived in [8] for grey level images are shown in Fig. 3d and e, respectively, and the human viewers rate their visual quality to be similar; this

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.

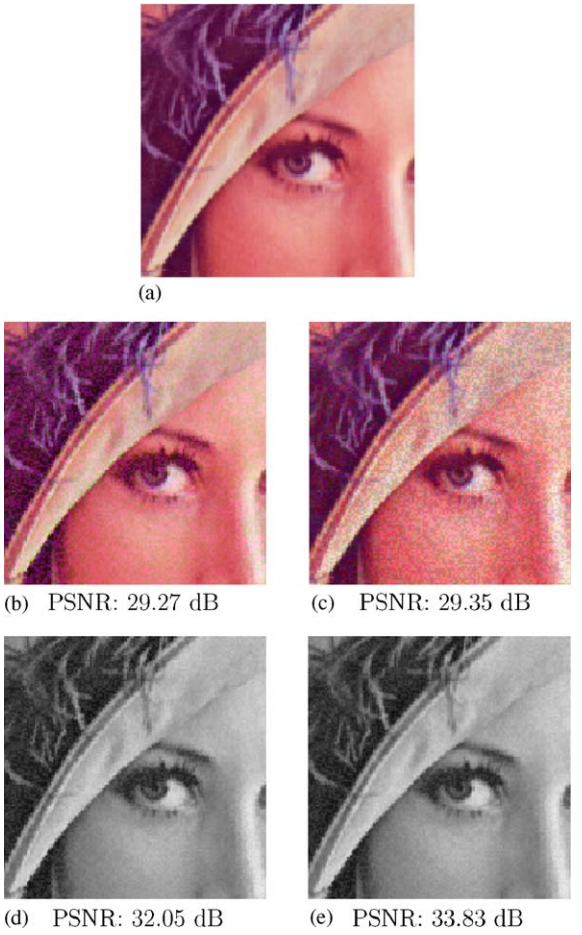

Fig. 3. Tests and comparison for the proposed NAMM in "Lenna" image: (a) original image; (b) noise injection with NAMM; (c) random noise injection; (d) noise injection with grey-level NAMM, i.e., the NAMM for grey-level image (JND_Y^S is computed by Eq. (2) without considering JND_{Cb}^S and JND_{Cr}^S); (e) noise injection with the model in [8].

therefore indicates that the NAMM model allows more data redundancy (of 1.78 dB in this grey level image case) for a similar perceptual picture quality level in the grey level image. Although the whole full-color image has processed, only the central part of the image has been displayed in Fig. 3 for clearer comparison of the printed versions.

To evaluate the performance of the proposed NAMM in comparison to the method in [8], the same experiment as above has been performed for 30 standard test color images as shown in Fig. 4. The results obtained are consistent and the average PSNR is shown in Table 1. In addition, the comparative subjective quality assessment of the noised images has also been performed. For fairness of comparison, noise is injected into the Y component according to the JND model in [8] while noise is injected into all YC_bC_r components according to the proposed NAMM.

The subjective assessment setup is similar to that in [25] and the quality assessment was performed by eight subjects (five of them are with average image processing knowledge and the rest are naive) under the above-mentioned subjective viewing conditions. On each trial of the experiment, subjects viewed two images of the same scene (see Fig. 5). Subjects were then given time to vote on the comparative quality of two images, using the continuous quality comparison scale shown in Table 2. The subjects were not allowed to respond until after they had viewed the images for at least 2 s. The order of the presentation of the 30 possible trials was randomized in each session. At each trial, the noised image appears randomly on the left- or right-hand side of the display (Fig. 5).

The results of the comparative subjective quality are listed in Table 3, where the mean and the standard deviation are computed over all 30 possible trials. From Table 3, we can see that the overall comparative subjective quality tends to a near zero mean of -0.055 with its associated standard deviation of 1.159. Thus, the subjective quality for the images noised with the proposed JND profile is very close to that with the JND profile in [8].

In summary, the proposed NAMM scheme provides a more accurate JND profile towards the actual JND bound in the HVS, since it is capable of exploiting larger JND values without jeopardizing the visual quality. As indicated in Table 1 (the first and the last rows), it is expected to outperform the approach in [8] by more than 2 dB (in terms of PSNR) of the permitted data redundancy on average for the same level of visual quality, mainly because of the due consideration of the compound masking effect and full color impacts. Consequently, it enables better visual compression and data hiding.

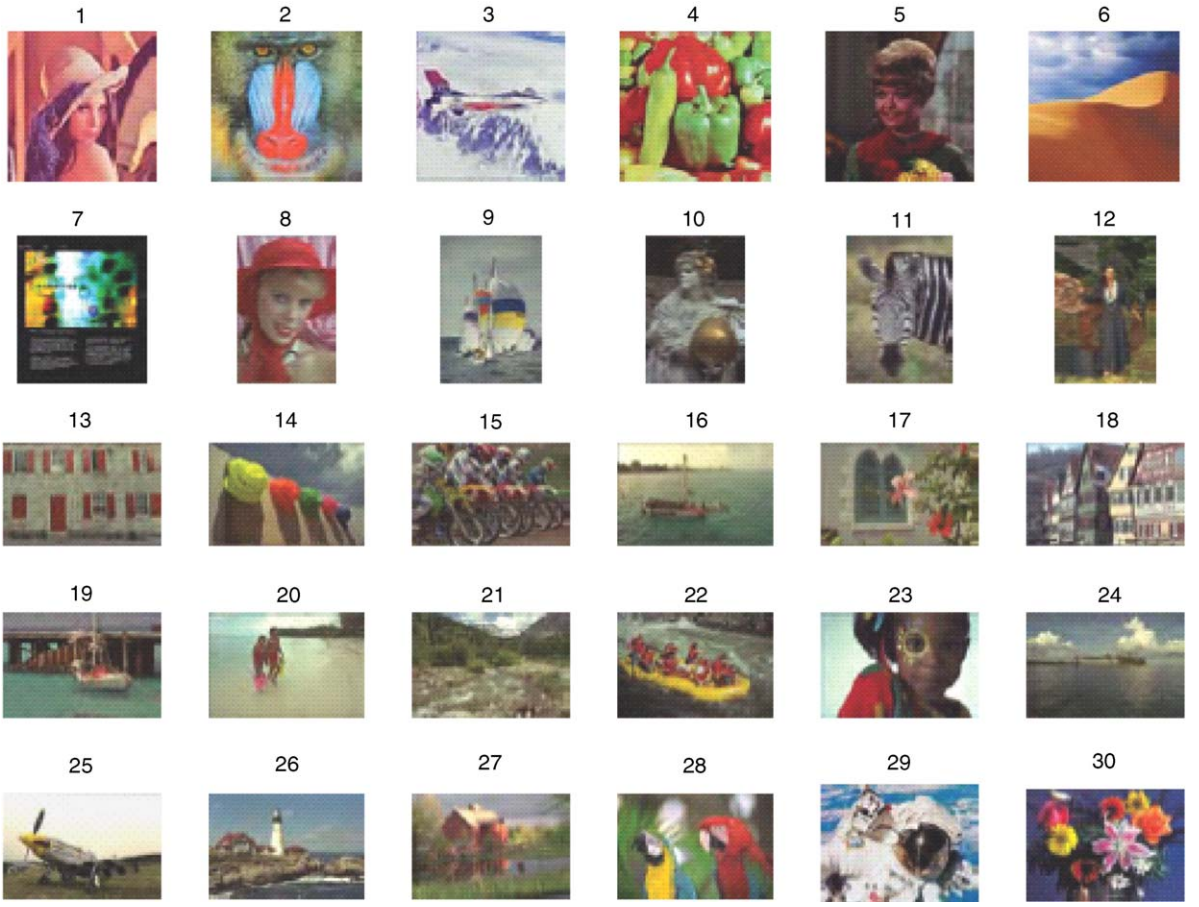

Fig. 4. Thumbnail of the color image set for testing.

Table 1
Comparison of average PSNR (dB) for 30 images

Method	Average PSNR (dB)
Noise injection with NAMM (full color)	29.04
Random noise injection	29.13
Noise injection with grey level NAMM	30.37
Noise injection with model in [8]	31.44

Fig. 5. Experimental setup for the comparative subjective quality assessment of JND profiles.

2.4. Spatio-temporal JND profile for color video

Temporal effect on JND can be incorporated as the scaled amplitude of the spatial JND [7]. Usually bigger interframe difference (caused by motion) leads to larger temporal masking, as the empirical curve [7] shown in Fig. 6. The overall JND can therefore be denoted as

$$\mathrm{JND}_\theta(x,y,t) = f(ild_\theta(x,y,t))\mathrm{JND}^S_\theta(x,y), \qquad (12)$$

where $ild_\theta(x,y,t)$ represents the average inter-frame luminance difference between the tth and

Table 2
Comparison scale for subjective quality evaluation

−3	The left one much worse than the right one
−2	The left one worse than the right one
−1	The left one slightly worse than the right one
0	The same
+1	The left one slightly better than the right one
+2	The left one better than the right one
+3	The left one much better than the right one

Table 3
The comparative subjective quality ("+": the proposed JND is better, "−": the JND in [8] is better)

Subject index	Mean	Standard deviation
1	−0.173	0.888
2	−0.056	0.348
3	−0.130	1.548
4	−0.280	1.160
5	+0.221	1.370
6	−0.019	0.590
7	−0.319	1.005
8	+0.312	2.361
Average	−0.055	1.159

$(t-1)$th frame [7]:

$$ild_\theta(x,y,t) = 0.5(I_\theta(x,y,t) - I_\theta(x,y,t-1) + \overline{I}_\theta(x,y,t) - \overline{I}_\theta(x,y,t-1)), \quad (13)$$

where $\overline{I}_\theta(x,y,t)$ is the average intensity and $f(ild_\theta(x,y,t))$ is the function defined as in Fig. 6.

3. Perceptual video coding scheme

The image-domain JND profile provides useful information in visual perceptual quality evaluation because it gives a local threshold of visibility, and how visual quality is judged affects many important decisions in video codecs. In this section, possibilities are explored to incorporate the JND profile into the motion prediction loop for hybrid video coding towards better coding efficiency and smaller perceptual distortion at a given bit rate. The general framework for the proposed PVC is given in Fig. 7. Apart from the JND profile generation, the difference between the PVC and a typical hybrid coder is the JND-adaptive motion estimation and the JND-adaptive residue filter, which will be presented in the following subsections.

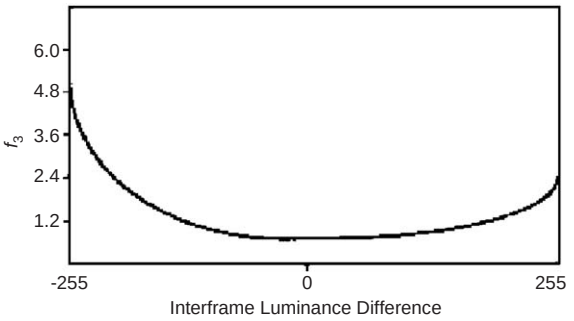

Fig. 6. Temporal masking effect.

Perceptual coding quality can be measured by the peak signal-to-perceptual-noise ratio (PSPNR) [8] that only takes into account the distortion that exceeds the JND profile:

$$\mathrm{PSPNR}_\theta(t) = 10\log_{10}\frac{255\times 255}{\frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}(|I_\theta(x,y,t)-\hat{I}_\theta(x,y,t)| - \mathrm{JND}_\theta(x,y,t))^2\delta_\theta(x,y,t)} \quad (14)$$

with

$$\delta_\theta(x,y,t) = \begin{cases} 1, & \text{if } |I_\theta(x,y,t)-\hat{I}_\theta(x,y,t)| \geqslant \mathrm{JND}_\theta(x,y,t), \\ 0, & \text{otherwise}, \end{cases} \quad (15)$$

where $\hat{I}_\theta(x,y,t)$ denotes the reconstructed intensity component for a pixel located at (x,y) for color channel θ of tth coded frame. If $\mathrm{JND}_\theta(x,y,t) \equiv 0$,

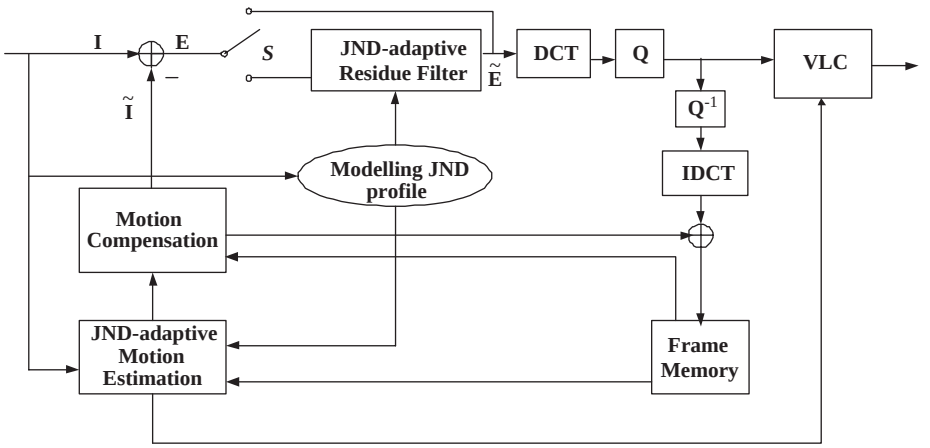

Fig. 7. The framework of the proposed perceptual-adaptive video encoder based on the new JND profile.

(14) becomes the conventional signal distortion criterion, $\mathrm{PSNR}_\theta(t)$ (peak signal-to-noise ratio).

3.1. JND-adaptive motion estimation

Motion estimation is the most computationally expensive module in hybrid video coding, and conventionally motion estimation for a macroblock or a block is to find a motion vector associated with a zero, minimum or sufficiently-small SAD for luminance. However, the effect of an SAD on perceptual quality depends upon not only its magnitude but also the local JND. In this work, we propose the sum of absolute perceptual difference (SAPD) as the perceptual motion estimation matching criterion. The SAPD for the (k,l)th macroblock/block in the current (tth) frame with a tentative motion vector (p,q) is defined as

$$\mathrm{SAPD}(k,l,p,q,t)=\sum_{s=1}^{B^2}(|i_Y^{k,l}(s,t)-\hat{i}_Y^{k,l,p,q}(s,t-1)| - \mathrm{JND}_Y^{k,l}(s,t))\delta_Y^{k,l}(s,t), \quad (16)$$

where $i_Y^{k,l}(s,t)$ and $\hat{i}_Y^{k,l,p,q}(s,t-1)$ are the sth luminance pixels in the current macroblock/block and the candidate macroblock/block to be matched in the previous reconstructed frame, respectively; $s=1,2,\ldots,B^2$, corresponding to the raster scanning order for pixels in a macroblock/block whose dimension is B (B is 16 for macroblock and 8 for block and, $k=0,1,\ldots,M/B-1$, $l=0,1,\ldots,N/B-1$); $-R\leqslant p,q\leqslant R$, and R is the maximum possible displacement of motion vectors; $\mathrm{JND}_Y^{k,l}(s,t)$ represents the luminance JND at the sth pixel of the current macroblock/block; and

$$\delta_Y^{k,l}(s,t)=\begin{cases}1, & \text{if } |i_Y^{k,l}(s,t)-\hat{i}_Y^{k,l,p,q}(s,t-1)| \\ & \quad \geqslant \mathrm{JND}_Y^{k,l}(s,t),\\ 0, & \text{otherwise,}\end{cases} \quad (17)$$

SAPD becomes SAD if $\mathrm{JND}_Y^{k,l}(j,t)\equiv 0$.

Motion vector is determined by matching the current macroblock/block to a reconstructed macroblock/block in the previous frame with the smallest SAPD:

$$\mathrm{MV}(k,l,t)=\arg\min_{p,q}\mathrm{SAPD}(k,l,p,q,t). \quad (18)$$

The process in (18) yields better PSPNR as discussed below, and can accelerate the motion estimation process.

The complexity of motion estimation depends on: (1) the number of candidate motion vectors in the search window (i.e. search points); (2) the computation to perform each of the candidate motion vectors. Thus, fast motion estimation techniques strive to reduce either the number of search points (fast search) (e.g., [28]) or the computational cost of matching each candidate motion vector (fast match) (e.g., [24]). The concept of SAPD can be easily incorporated into the full

search motion estimation, any fast search algorithms and fast match schemes. For fast search algorithms, SAPD increases the occurrence of zeros since it does not take any objective distortion below the JND profile into account, and avoids the unnecessary further search once the macroblock/block difference is not perceptually significant. Once a zero or sufficiently small SAPD is encountered, the motion estimation for the current macroblock/block can be terminated.

In [24], a fast match algorithm was proposed to compute the partial SAD (PSAD) recursively for each candidate motion vector so that a matching process can be pruned when the PSAD exceeds that of the best candidate encountered so far. The PSAD-based matching scheme is more efficient when combined with a center-biased search scheme, such as spiral scanning scheme, because of the center-biased motion vector distribution characteristics in the real-world video [5].

SAPD can replace SAD in the aforementioned recursive process via the partial SAPD (PSAD) expressed as

$$\begin{aligned}&\mathrm{PSAPD}_s(k,l,p,q,t)\\&\quad=\mathrm{PSAPD}_{s-1}(k,l,p,q,t)+(|i_Y^{k,l}(s,t)\\&\qquad-\hat{i}_Y^{k,l,p,q}(s,t-1)|-\mathrm{JND}_Y^{k,l}(s,t))\delta_Y^{k,l}(s,t),\end{aligned}\quad(19)$$

where $\mathrm{PSAPD}_0(k,l,p,q,t)=0$; $\mathrm{PSAPD}_s(k,l,p,q,t)<\mathrm{PSAPD}_{B^2}(k,l,p,q,t)=\mathrm{SAPD}(k,l,p,q,t)$, for all $s<B^2$, $\mathrm{PSAPD}_s(k,l,p,q,t)$ reduces to $\mathrm{PSAD}_s(k,l,p,q,t)$ if $\mathrm{JND}_Y^{k,l}(s,t)\equiv 0$. In each recursion, PSAPD_s is compared with the SAPD of the "best found-so-far" motion vector (SAPD_{BSF}). Whenever PSAPD_s is greater than SAPD_{BSF}, the recursion in (19) terminates and the process moves on to evaluate the next candidate motion vector.

Formulae (19), in comparison with its PSAD counterpart, needs one extra comparison and one extra substraction, but one addition less in the case with $\delta_Y^{k,l}(j,t)=0$. It is expected that the number of recursions with (19) is much less than that based on PSAD for the real-world video. The experimental results in the next section will demonstrate that the motion estimation based on PSAPD metric outperforms that based on PSAD in terms of the overall computational complexity.

3.2. JND-adaptive residue filter

The residue image after motion compensation is

$$E_\theta(x,y,t)=I_\theta(x,y,t)-\tilde{I}_\theta(x,y,t),\quad(20)$$

where $\tilde{I}_\theta(x,y,t)$ denotes the intensity component for a motion-compensated pixel.

Since any $E_\theta(x,y,t)$ below the associated JND is invisible, it is unnecessary to code such a residue from the perceptual distortion point of view. Let q_τ denote the average quantization step of the previously inter-encoded frame. The residues that need to be DCT-transformed can be obtained via a JND-adaptive residue filtering manipulation:

$$E'_\theta(x,y,t)=\begin{cases}0 & \text{if } |E_\theta(x,y,t)|<\mathrm{JND}_\theta(x,y,t)\\ & \quad\text{and } q_\tau>T_q,\\ E_\theta(x,y,t) & \text{otherwise,}\end{cases}\quad(21)$$

where T_q is a threshold to ensure that the filtering is activated only when the frame is with very low motion (in this work, T_q is empirically determined as 10). For simplicity but without loss of generality, we assume that no scene change occurs (in practice, a scene change detector can be used to detect scene change so that the first frame of a new scene is adaptively set to be an intra-frame); therefore, a bigger q_τ indicates higher motion because it is caused by a lower level in the stipulated "bit reservoir" (or higher buffer fullness in applications involving transmission). The switch is illustrated in Fig. 7.

When the JND-adaptive residue filter is switched on, those residues below JND are discarded. If all residues in a block are less than their corresponding JNDs, such a block becomes a zero block after JND-adaptive residue filtering and can therefore be skipped at the DCT phase. The saved bits can be used for better coding of those residues above JND. If only a portion of residues in the block are less than their corresponding JNDs, the JND-adaptive residue filtering process reduces the variance of the DCT coefficients in the block. From the rate-distortion viewpoint, signal variance is a good measurement of compressibility [3,13]. For a given bit-rate, smaller variance of

signal results in less objective distortion of the reconstructed signal. Therefore, the JND-adaptive residue filter reduces not only perceptual distortion but also objective distortion if the better representation of the residues above JND sufficiently compensates the loss of the residues below JND in overall effect for the given bits (as it will be further analyzed in Section 3.3).

When the allowed bits are sufficient (q_τ is small), there is no need to allocate more bits for the more significant residues. In consequence, the JND-adaptive residue filter is switched off so that all residues are coded since both the residues below and those above JND can be coded sufficiently by the available bits.

The additional check with q_τ in (21) is a provision for coding frames with very slow motion to make full use of the available bandwidth. For slow-motion frames, many motion-compensated residues are below JNDs; this leads to possible buffer underflow and bandwidth under-usage if the JND-adaptive filtering is used without the check with q_τ. With the proposed JND-adaptive filtering scheme, when the average motion in the previously inter-coded frame is small (i.e., q_τ is small), those residues below JNDs are still coded to make full use of the available bandwidth, for minimum objective distortion.

3.3. Rate-distortion analysis for JND-adaptive residue filtering

Signal distortion of conventional video coder is uniquely introduced by the quantizer (uniform or non-uniform quantizer) for DCT coefficients. With the proposed PVC scheme, additional consideration is the JND-adaptive residue filtering scheme. Let $X_{u,v}$ and $X'_{u,v}$ represent collections of DCT coefficients of $E_\theta(x,y,t)$ and $E'_\theta(x,y,t)$, respectively, with u and v being the frequency indices $(0 \leqslant u < U, 0 \leqslant v < V)E^J_\theta(x,y,t)$ is the difference signal caused by the JND-adaptive residue filtering scheme:

$$E^J_\theta(x,y,t) = E_\theta(x,y,t) - E'_\theta(x,y,t) = \begin{cases} 0, & \text{if } |E_\theta(x,y,t)| \geqslant \mathrm{JND}_\theta(x,y,t) \text{ or } q_\tau \leqslant T_q, \\ E_\theta(x,y,t), & \text{otherwise.} \end{cases} \quad (22)$$

For the convenience of presentation, the following notations are defined:

$\mathbf{X} = \{X_{u,v}, 0 \leqslant u < U, 0 \leqslant v < V\}$,
$\mathbf{X}' = \{X'_{u,v}, 0 \leqslant u < U, 0 \leqslant v < V\}$,
$\mathbf{E}^J$—the collection of $E^J_\theta(x,y,t)$'s corresponding to $\mathbf{X}$ and $\mathbf{X}'$,
$\overline{D}(\mathbf{X})$—average signal distortion measure with MSE for conventional video coder,
$\overline{D}(\mathbf{X}')$—average signal distortion measured with MSE for the video coder with the JND-adaptive residue filter,
$\overline{D}(\mathbf{E}^J)$—average signal distortion measured with MSE due to $\mathbf{E}^J$:

$$\overline{D}(\mathbf{E}^J) = \frac{1}{MN}\sum_{x=1}^{M}\sum_{y=1}^{N}(E^J_\theta(x,y,t))^2. \quad (23)$$

When a conventional video coder and a coder with the proposed PVC scheme is compared, the following question arises:

Does the proposed PVC scheme satisfy Inequality (24)?

$$\overline{D}(\mathbf{X}) - \overline{D}(\mathbf{X}') > \overline{D}(\mathbf{E}^J). \quad (24)$$

In other words, can the better representation of the residues above JND sufficiently compensate the loss of those below JND in the overall signal distortion for a given bit-rate?

Since the proposed PVC does not modify the intra-coding process, the signal distortion of I-frames is largely unchanged, so only the signal distortion of inter-coded frames needs to be considered. Under the assumption that $X_{u,v}$ and $X'_{u,v}$ follow Laplacian distribution, it has been proven (see 5) that for medium and high bit coding

with uniform quantization over a block,

$$\overline{D}(\mathbf{X}) = 1.2e^{-1.386\overline{b}}F(\mathbf{X}) \quad \text{and}$$
$$\overline{D}(\mathbf{X}') = 1.2e^{-1.386\overline{b}}F(\mathbf{X}'), \tag{25}$$

where $\overline{b}$ is the average number of bits assigned for coding the signal under consideration; and

$$F(\mathbf{X}) = \left(\prod_{\substack{0\leqslant u<U\\0\leqslant v<V}} \sigma^2_{X_{u,v}}\right)^{1/UV}, \tag{26}$$

where σ^2 represents signal variance.

Inequality (24) is held when $\sigma^2_{X'}$ (with JND-adaptive filtering) is sufficiently smaller than σ^2_X (without JND-adaptive filtering).

The experiments with standard test video sequences show the approximate Laplacian distribution for $X_{u,v}$ and $X'_{u,v}$ (the probability function of Laplacian distribution is given in (37) and (38) in 5) and more aggregation for $X'_{u,v}$ in the near-zero region than for $X_{u,v}$. As an example, Fig. 8 illustrates the distribution of $X_{u,v}$ and $X'_{u,v}$ with $(u,v) = (0,0)$ and (0,1), respectively, over the first 120 frames of "Harp" sequence. Table 4 is the comprehensive list of $\sigma^2_{X_{u,v}}$ and $\sigma^2_{X'_{u,v}}$ for "Harp" sequence, and by the coefficient-by-coefficient comparison, it can be seen that $\sigma^2_{X'_{u,v}}$'s are smaller than $\sigma^2_{X_{u,v}}$'s. It is therefore more efficient to encode $X'_{u,v}$ than $X_{u,v}$. Table 5 shows that Inequaty (24) is held by listing the three variables defined in Inequaty (24) for the eight video sequences.

4. Overall performance of the PVC scheme

The proposed PVC scheme is illustrated in Fig. 7, and has been implemented by incorporating the aforementioned JND-adaptive motion estimation and residue filtering scheme into the MPEG-2 Test Model 5 (TM5) [16] coder. Eight test sequences are used for experiments, as shown in Fig. 9. These sequences span a spectrum of various motion, zooming, color and texture. The first three sequences are with frame rate of 25 fps and resolution of 720×576 pixels, and the rest are with frame rate of 30 fps and resolution of 720×480 pixels. The target bit rate is set as 5 Mbits/s and the GOP length is 12 frames. Only forward frame prediction mode is used and full search with spiral scanning is used for motion estimation with $R = 15$. The uniform quantization scheme is adopted. A Pentium IV 2.2-GHz processor is utilized in the experiments.

Fig. 10 illustrates the frame-by-frame comparison of PSNR and PSPNR of the motion-compensated frames, with the JND-adaptive motion estimation (using PSAPD as matching criterion) and conventional motion estimation (using PSAD as matching criterion), respectively, on the "Harp" sequence. The average PSNR and PSPNR for "Harp" and other test sequences are summarized in Table 6. As it can be seen from Fig. 10 and Table 6, the JND-adaptive motion estimation outperforms conventional motion estimation in PSPNR for most of the frames and with

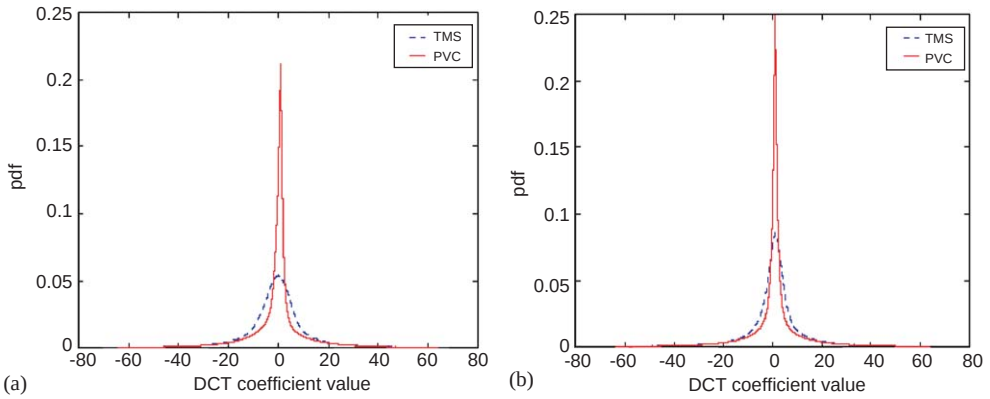

Fig. 8. Distribution of $X_{u,v}$ and $X'_{u,v}$ for "Harp" sequence, with (u,v): (a) (0,0), (b) (0,1).

Table 4
Variances of DCT coefficients of "Harp" sequence for different u, v: (a) $\sigma^2_{X_{u,v}}$, (b) $\sigma^2_{X'_{u,v}}$

	0	1	2	3	4	5	6	7
(a)								
0	9825	367.12	401.9	462.73	270.71	264.72	163.37	114.9
1	219.17	257.24	294.5	335.96	236.42	235.67	177.71	155.37
2	211.71	217.94	240.04	263.39	221.34	228.77	205.1	200.26
3	198.22	182.51	182.16	186.12	187.99	200.85	206.81	224.52
4	164.56	140.59	128.5	123.06	132.89	143.28	164.35	196.66
5	125.88	100.93	88.25	80.19	85.29	92.88	109.58	133.27
6	90.12	73.00	63.52	57.4	59.06	62.59	71.76	86.16
7	42.66	34.85	30.25	27.46	27.093	27.44	30.41	35.51
(b)								
0	260.07	343.71	377.6	437.33	254.94	249.75	152.92	106.13
1	196.32	237.79	277.14	319.27	223.5	222.94	168.11	145.57
2	188.63	199.56	223.44	248.24	210.45	217.2	193.97	187.35
3	177.17	165.54	167.18	174.29	177.74	190.37	195.31	210.1
4	146.98	127.05	116.93	113.72	124.94	135.23	154.6	183.08
5	110.3	90.13	79.36	73.50	79.58	86.89	101.99	122.96
6	77.74	63.62	56.374	51.82	54.72	57.59	65.967	77.99
7	38.78	32.28	28.99	26.96	27.28	27.40	30.01	34.05

Table 5
Analysis of signal distortion measured with MSE

Video sequence	$\overline{D}(\mathbf{X})$	$\overline{D}(\mathbf{X}')$	$\overline{D}(\mathbf{X}) - \overline{\mathbf{D}}(\mathbf{X}')$	$\overline{D}(\mathbf{E}')$	Inequality (24)
Harp	130.71	119.78	10.94	4.55	True
Barcelona	89.41	86.40	2.99	2.84	True
Mobile & calender	166.38	159.29	7.08	4.94	True
Autumn leaves	14.59	13.71	0.87	0.53	True
Football	42.88	39.07	3.80	1.84	True
Sailboat	31.71	30.15	1.56	1.09	True
Susie	7.803	6.509	1.29	0.46	True
Tempete	65.58	58.20	7.37	6.051	True
Average over 8 Seq.	68.63	64.14	4.49	2.79	—

an average of 0.25 dB improvement for the sequences under test, while maintaining almost the same PSNR. Computational complexity of motion estimation can be measured by the average number of search points, average number of recursions in (19) and its counterpart for PSAD, and the average CPU time per frame. Fig. 11 shows that these three computational complexity measures for the PSAPD matching metric are significantly lower than those for the PSAD metric, on the frame basis for the "Harp" sequence. Table 7 confirms that this is also true for other video test sequences (with average improvement of 5.5% in number of search points, 27.3% in recursion for matching and 18.6% in CPU time per frame).

Fig. 12 illustrates the overall PSNR and PSPNR comparison of the reconstructed video for the "Harp" sequence compressed by the MPEG-2 TM5 coder and the proposed PVC scheme (with

Fig. 9. Thumbnail of the test sequences: (a) harp, (b) barcelona, (c) mobile & calender, (d) autumn leaves, (e) football, (f) sailboat, (g) Susie, (h) tempete.

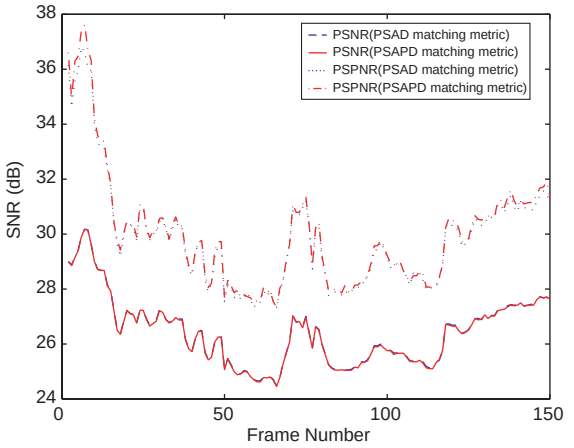

Fig. 10. PSNR and PSPNR comparison of the motion-compensated frames after respectively applying PSAPD and PSAD matching criteria in full motion estimation for the "Harp" sequence.

Table 6
Signal distortion (PSNR) and perceptual distortion (PSPNR) of the motion-compensated pictures obtained via PSAD and PSAPD matching metric, respectively

Video sequence	PSAD matching metric		PSAPD matching metric	
	PSNR (dB)	PSPNR (dB)	PSNR (dB)	PSPNR (dB)
Harp	26.53	29.67	26.51	29.81
Barcelona	27.44	30.62	27.43	30.75
Mobile & calender	25.44	28.84	25.44	29.03
Autumn leaves	34.94	41.02	34.89	41.36
Football	25.81	27.55	25.81	27.60
Sailboat	33.42	41.65	33.39	43.05
Susie	38.06	45.15	38.06	45.78
Tempete	27.14	30.29	27.13	30.42
Average over 8 Seq.	29.85	34.35	29.83	34.60

the JND-adaptive residue filtering). It can be seen that both PSNR and PSPNR are improved in the proposed PVC scheme. The results for all video sequences under test are listed in Table 8: an average PSNR gain of 0.17 dB and an average PSPNR gain of 0.6 dB are achieved by the proposed PVC scheme.

5. Conclusions

In this paper, the possibilities have been investigated for the image-domain JND profile to be applied to the motion prediction loop of the standardized hybrid video coding (H.261/263, MPEG-1/2/4 and the emerging H.264 video coding).

A new image-domain JND estimator is firstly proposed for color video. The general nonlinear additivity model for masking (NAMM) has been

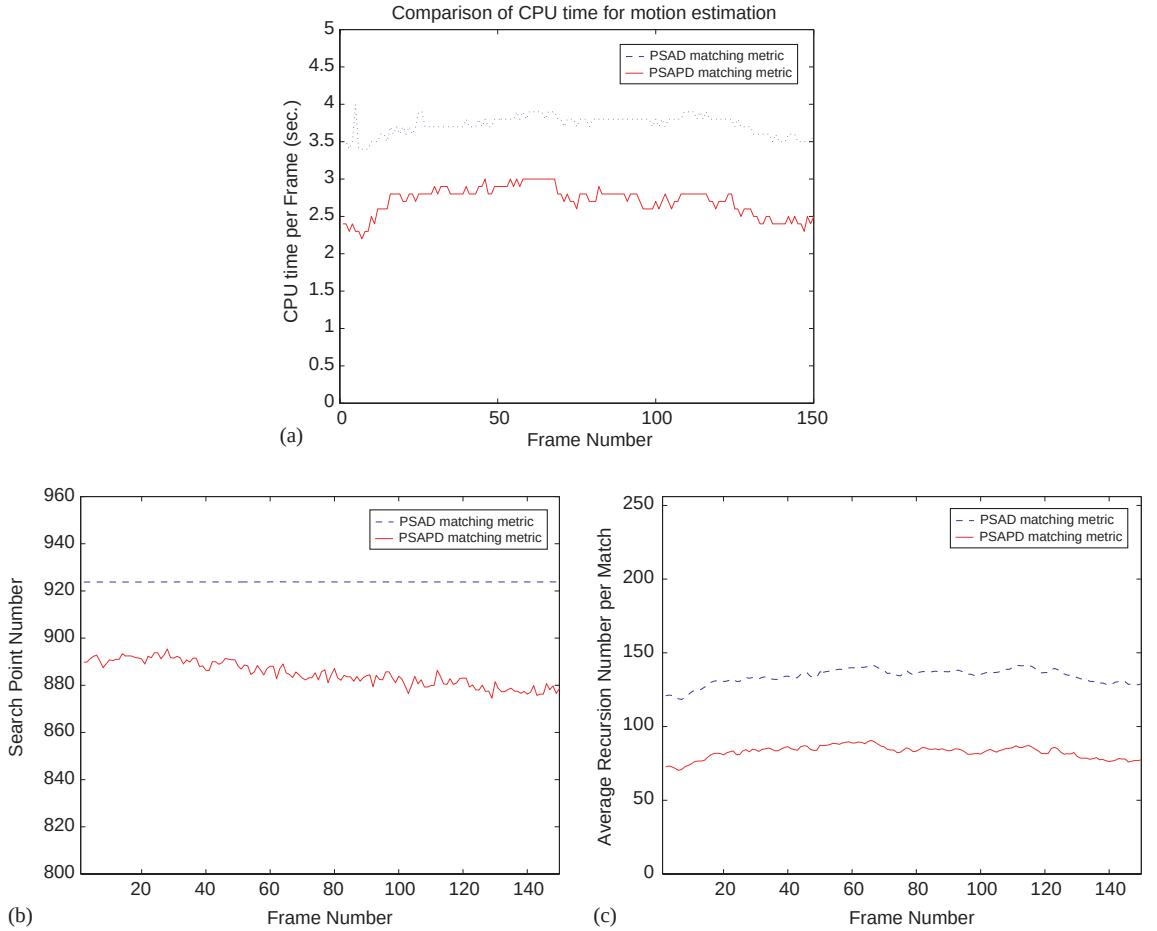

Fig. 11. Computational complexity comparison of motion estimation using conventional PSAD and the proposed PSAPD matching criteria on the "Harp" sequence: (a) CPU time per frame; (b) average number of search points in motion estimation; (c) average recursion number in matching ((19) and its counterpart for PSAD).

devised to integrate different spatial masking factors and applied to all color components in the image/video. The introduction of the nonlinear accumulation of various masking effects allows the exploitation of larger JND values without jeopardizing the visual quality. Factors considered at the current stage of implementation includes luminance masking, texture masking (for edge and non-edge regions), and temporal masking. Extensive subjective tests confirm that the proposed scheme provides a more accurate JND profile towards the actual JND bound in the HVS. This is because at the same level of perceptual quality, it allows an average, more than 2 dB of extra perceptually-lossless data redundancy than the most related existing method.

The major contributions of the new JND estimator are: the general formulae for nonlinear accumulation of different masking effects, allowing a higher visibility threshold (i.e., bigger perceptually-lossless data redundancy) than that derived in [8,7] when more maskers are present; chrominance components being considered for additional perceptually-lossless data redundancy; edge regions being distinguished from non-edge ones for better accuracy. The JND estimators in [8,6] are special cases of the proposed NAMM.

Table 7
Computational complexity of motion estimation based on PSAD and PSAPD matching metric, respectively

Video sequence	PSAD matching metric			PSAPD matching metric		
	Search points per MV	Recursions per match	CPU time (s)	Search points per MV	Recursions per match	CPU time (s)
Harp	921.17	134.90	3.72	882.49	83.87	2.71
Barcelona	921.12	75.00	1.99	916.91	50.57	1.81
Mobile & calender	921.15	72.97	1.91	862.81	61.70	1.80
Autumn leaves	916.07	68.01	1.27	889.89	43.04	1.02
Football	916.17	148.14	2.77	897.21	104.48	2.59
Sailboat	916.08	68.81	1.34	800.29	63.72	1.00
Susie	916.10	102.61	2.19	862.26	59.30	1.46
Tempete	916.10	89.47	2.19	846.60	74.49	1.72
Average	917.99	94.99	2.12	869.81	70.15	1.76
Average improvement with the PSAPD metric (%)				5.5	27.3	18.6

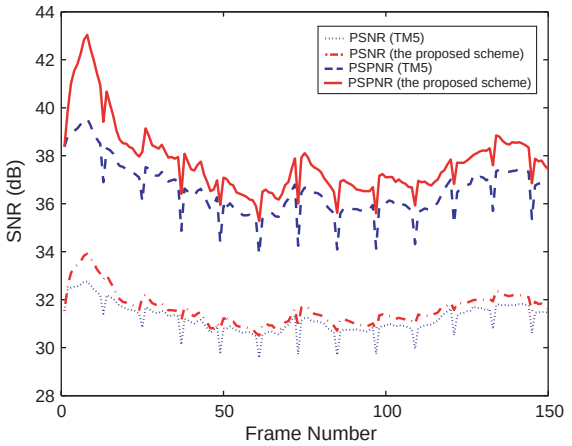

Fig. 12. PSNR and PSPNR comparison of "Harp" sequence compressed by: (a) the MPEG-2 TM5 coder; (b) the proposed PVC.

The new perceptually-adaptive video coding (PVC) scheme has been developed by incorporating the derived image-domain JND profile into the motion prediction loop and the filtering process on motion compensated residues before DCT coding. Both the JND-adaptive motion estimation and residue filtering scheme improve the perceptual coding quality (measured by peak signal-to-perceptual-noise ratio—PSPNR) and video encoding efficiency (in terms of search points in motion estimation, number of recursions and execution time).

Perceptual visual coding is usually expected to lower the objective quality measure because the objective measure is no longer the metric in the process. However, the proposed JND-adaptive residue filtering has the tendency to improve the objective quality measure (such as PSNR) since it forces the insignificant residues to give way to the significant ones in bit allocation. Theoretical explanation has been presented for this tendency.

The proposed JND-based perceptual coding scheme achieves operational savings and also allows the scarce bits to be allocated for higher visual quality with coded signal. The scheme works with various motion estimation strategies and any prevalent video coding standards. As an example for demonstration, the proposed PVC scheme has been implemented in the MPEG-2 TM5 coder, and achieved average improvement of over 18% in motion estimation efficiency, 0.6 dB in the peak signal-to perceptual-noise ratio (PSPNR) and 0.17 dB in the objective coding quality measure (PSNR).

Table 8
The overall signal distortion (PSNR) and perceptual distortion (PSPNR) of the motion-compensated frames encoded by the original MPEG-2 TM5 coder and the proposed PVC scheme, respectively

Video sequence	TM5		PVC	
	PSNR (dB)	PSPNR (dB)	PSNR (dB)	PSPNR (dB)
Harp	31.12	36.55	31.48	37.38
Barcelona	28.95	33.06	29.30	33.67
Mobile & calender	27.97	32.16	28.13	32.60
Autumn leaves	36.53	44.78	36.64	45.23
Football	32.24	36.97	32.25	37.21
Sailboat	34.68	44.12	34.85	44.97
Susie	40.60	52.25	40.82	52.88
Tempete	31.12	37.42	31.20	38.15
Average over 8 Seq.	32.91	39.66	33.08	40.26
Average improvement for PVC (dB)			0.17	0.60

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable advice towards the improvement of this paper.

Appendix. Derivation for $\overline{D}(\mathbf{X})$ and $\overline{D}(\mathbf{X}')$

In this appendix, the relationship is to be derived between $\overline{D}(\mathbf{X})$ and σ_X^2 (same process applying to $\overline{D}(\mathbf{X}')$ and $\sigma_{X'}^2$ as well).

It is known that, except for very low bit rate, the number of bits (b) versus MSE-based signal distortion (D) for a zero-mean independent and identically distributed (i.i.d.) source $\mathbf{X}$ can be approximated by the following formula when the quantized data are entropy-coded [3,13,12,29]:

$$b(\Delta) = \frac{1}{\alpha}\log_e\left(\varepsilon\beta\,\frac{\sigma_{\mathbf{X}}^2}{\Delta^2}\right) \tag{27}$$

and

$$D(\Delta) = \frac{\Delta^2}{\beta}. \tag{28}$$

Combining the two equations above,

$$D(b) = \varepsilon e^{-\alpha b}\sigma_{\mathbf{X}}^2, \tag{29}$$

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

$$\varepsilon \simeq \begin{cases} 1 & \text{for uniform distribution,} \\ 1.4 & \text{for Gaussian distribution,} \\ 1.2 & \text{for Laplacian distribution.} \end{cases} \tag{30}$$

In most practical image/video coders, signal to be quantized is a $U \times V$ block of DCT coefficients, i.e., $\mathbf{X} = \{X_{u,v}, 0 \leqslant u < U, 0 \leqslant v < V\}$, with smaller (u, v)'s denoting lower DCT frequencies. In general, the quantization step size is

$$\Delta_{u,v} = q_s W_{u,v}, \tag{31}$$

where q_s is a quantization scaling factor for the entire block and $[W_{u,v}]$ is a matrix whose elements are used as multiplicative factors to produce the actual step sizes for quantization.

A uniform quantizer assumes that all the step sizes in a block are identical [20,21], that is, $W_{u,v} \equiv 1$ for all (u, v)'s.

A non-uniform quantizer takes the frequency-dependent visual sensitivity of human perception [29] into account: bits are assigned to a frequency component according to its perceptual threshold. Thus, $W_{u,v}$ (and therefore $\Delta_{u,v}$) have different values [17–19].

For a collection of $X_{u,v}$'s with the same q_s, the corresponding number of bits can be found by substituting (31) into (27):

$$b_{u,v}(q_s) = \frac{1}{\alpha}\log_e\left(\frac{\beta\varepsilon_{u,v}\sigma_{X_{u,v}}^2}{q_s^2 W_{u,v}^2}\right). \tag{32}$$

Then the average number of bits is

$$\begin{aligned}\overline{b}(q_s) &= \frac{1}{UV}\sum_{u=0}^{U-1}\sum_{v=0}^{V-1} b_{u,v}(q_s) \\ &= \frac{1}{\alpha}\log_e\left[\frac{\beta}{q_s^2}\left[\prod_{\substack{0\leqslant u<U \\ 0\leqslant v<V}}\left(\frac{\varepsilon_{u,v}\sigma_{X_{u,v}}^2}{W_{u,v}^2}\right)\right]^{1/UV}\right].\end{aligned} \tag{33}$$

From (29) and (31), the average distortion corresponding to $\overline{b}$:

$$\overline{D}(\overline{b}) = \frac{1}{UV}\sum_{u=0}^{U-1}\sum_{v=0}^{V-1} D_{u,v}(q_s) = \frac{q_s^2}{\beta UV}\sum_{u=0}^{U-1}\sum_{v=0}^{V-1} W_{u,v}^2 \tag{34}$$

substituting (33) into (34):

$$\overline{D}(\overline{b}) = \frac{1}{UV}\mathrm{e}^{-\alpha\overline{b}}\left(\prod_{\substack{0\leqslant u<U\\0\leqslant v<V}}\left(\frac{\varepsilon_{u,v}\sigma_{X_{u,v}}^2}{W_{u,v}^2}\right)\right)^{1/UV} \times \sum_{u=0}^{U-1}\sum_{v=0}^{V-1} W_{u,v}^2 = \frac{1}{UV}\mathrm{e}^{-\alpha\overline{b}}F(\mathbf{X}) \tag{35}$$

with

$$F(\mathbf{X}) = \left(\prod_{\substack{0\leqslant u<U\\0\leqslant v<V}}\left(\frac{\varepsilon_{u,v}\sigma_{X_{u,v}}^2}{W_{u,v}^2}\right)\right)^{1/UV}\sum_{u=0}^{U-1}\sum_{v=0}^{V-1} W_{u,v}^2. \tag{36}$$

$F(\mathbf{X})$ can be regarded as a general measure of source complexity, indicating how difficult for the given source $\mathbf{X}$ to be encoded using a particular quantization matrix $[W_{u,v}]$.

It can be seen from (35) and (36) that larger $\sigma_{X_{u,v}}^2$'s lead to larger $\overline{D}(\overline{b})$. So (24) can be held if $\sigma_{X'_{u,v}}^2$'s is significantly smaller than $\sigma_{X_{u,v}}^2$'s.

In the case of uniform quantization, Formula (35) and (36) can be simplified since $W_{u,v}$ is a constant. For instance, in the uniform quantizer of the MPEG-2 TM5 coder [16], $W_{u,v} \equiv 16$. The DCT coefficients of residues after motion compensation is with Laplacian (double-sided exponential) distribution [13,32] (see examples in Fig. 8), with the probability of X:

$$p(X) = \frac{\lambda}{2}\mathrm{e}^{-\lambda|X|} \tag{37}$$

with

$$\lambda = \sqrt{2}/\sigma_{\mathbf{X}}. \tag{38}$$

For Laplacian distribution, $\varepsilon_{u,v} \equiv 1.2$. With values of $W_{u,v}$, $\varepsilon_{u,v}$ and α substituted, (35) is simplified as

$$\overline{D}(\overline{b}) = 1.2\,\mathrm{e}^{-1.386\overline{b}}\left(\prod_{\substack{0\leqslant u<U\\0\leqslant v<V}}\sigma_{X_{u,v}}^2\right)^{1/UV} = 1.2\,\mathrm{e}^{-1.386\overline{b}}F(\mathbf{X}) \tag{39}$$

and

$$F(\mathbf{X}) = \left(\prod_{\substack{0\leqslant u<U\\0\leqslant v<V}}\sigma_{X_{u,v}}^2\right)^{1/UV}. \tag{40}$$

These are the ground for Formula (25) and (26).

References

[1] A.J. Ahumada, H.A. Peterson, Luminance-model-based DCT quantization for color image compression, in: SPIE International Conference on Human Vision, Visual Processing and Digital DisplayIII, 1992.

[2] A.P. Bradley, A wavelet visible difference predictor, IEEE Trans. Image Processing 8 (5) (May 1999) 717–730.

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

[4] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (6) (1986) 679–698.

[5] C.-K. Cheung, L.-M. Po, Normalized partial distortion search algorithm for block motion estimation, IEEE Trans. Circuits Syst. Video Technol. 10 (3) (April 2000) 417–422.

[6] Y.J. Chiu, T. Berger, A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancement, IEEE Trans. Circuits Syst. Video Technol. 9 (3) (April 1999) 438–450.

[7] C.-H. Chou, C.-W. Chen, A perceptually optimized 3-D subband image codec for video communication over wireless channels, IEEE Trans. Circuits Syst. Video Technol. 6 (2) (1996) 143–156.

[8] C.-H. Chou, Y.-C. Li, A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile, IEEE Trans. Circuits Syst. Video Technol. 5 (6) (1995) 467–476.

[9] M.P. Eckert, A.P. Bradley, Perceptual quality metrics applied to still image compression, Signal Processing 70 (1998) 177–200.

[10] J.H. Elder, R.M. Goldberg, Local scale control for edge detection and blur estimation, IEEE Trans. Pattern Anal. Mach. Intell. 20 (7) (July 1998) 699–716.
[11] B. Girod, What's wrong with mean-squared error? in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, 1993.
[12] H. Gish, J.N. Pierce, Asymptotically efficient quantizing, IEEE Trans. Inform. Theory 14 (September 1968) 676–683.
[13] H.-M. Hang, J.-J. Chen, Source model for transform video coder and its application—part I: fundamental theory, IEEE Trans. Circuits Syst. Video Technol. 7 (2) (April 1997) 287–298.
[14] I. Hntsch, L.J. Karam, Locally adaptive perceptual image coding, IEEE Trans. Image Processing 9 (9) (September 2000) 1472–1483.
[15] I. Hntsch, L.J. Karam, Adaptive image coding with perceptual distortion control, IEEE Trans. Image Processing 11 (3) (March 2002) 213–222.
[16] ISO/IEC JTC1/SC29 WG11, MPEG-2 test model 5. MPEG93/457, April 1993.
[17] ISO/IEC 11172-2, Coding of moving pictures and associated audio for digital storage media at up to about 1.5 mbits/s (MPEG-1), part 2: Video, 1993.
[18] ISO/IEC 13818-2, Generic coding of moving pictures and associated audio (MPEG-2), Part 2: Video, 1995.
[19] ISO/IEC JTC 1/SC 29/WG 11/14496, Information technology-coding of audio-visual objects (MPEG-4). part 2: Visual. MPEG2001/N4350, 2001.
[20] ITU-T Recommendation H.261, Video codec for audio-visual services at p x 64 kbits/s, 1993.
[21] ITU-T Recommendation H.263 Version 2, Video coding for low bitrate communication, 1998.
[22] N.S. Jayant, J.D. Johnston, R.J. Safranek, Signal compression based on models of human perception, Proc. IEEE 81 (1993) 1385–1422.
[23] Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, Editor's proposed draft text modifications for joint video specification (ITU-T Rec. H.264 — ISO/IEC 14496-10 AVC), draft 7. 2002.
[24] K. Lengwehasatit, A. Ortega, Probabilistic partial-distance fast matching algorithms for motion estimation, IEEE Trans. Circuits Syst. Video Technol. 11 (2) (February 2001) 139–152.
[25] P. Longere, X. Zhang, P.B. Delahunt, D.H. Brainaro, Perceptual assessment of demosaicing algorithm performance, Proc. IEEE 90 (1) (January 2002) 123–132.
[26] J. Malo, J. Gutierrez, I. Epifanio, F.J. Ferri, J.M. Artigas, Perceptual feedback in multigrid motion estimation using an improved DCT quantization, IEEE Trans. Image Processing 10 (10) (October 2001) 1411–1427.
[27] D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information, W.H. Freeman and Co., 1982.
[28] MPEG-4 Video Group, Text of iso/iec 14496-7 cd registration and consideration (optimized code for MPEG-4 visual standards), ISO/IEC/JTEC1/SC29/WG11, N3325, March 2000.
[29] A.N. Netravali, B.G. Haskell, Digital pictures: representation and compression, Plenum, New York, NY, 1988.
[30] H.C. Nothdurft, Salience from feature contrast: additivity across dimensions, Vision Research 40 (2000) 1183–1201.
[31] R.J. Safrenek, J.D. Johnson, A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression, in: International Conference on Acoustics, Speech, and Signal Processing, 1989.
[32] S.R. Smoot, L.A. Rowe, Study of DCT coefficient distributions, in: SPIE: Human Vision and Electronic Imaging, vol. 2657, 1996.
[33] H.Y. Tong, A.N. Venetsanopoulos, A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking, in: International Conference on Image Processing, 1998.
[34] T.D. Tran, R. Safranek, A locally adaptive perceptual masking threshold for image coding, in: International Conference on Acoustics, Speech, and Signal Processing, 1996.
[35] A.B. Watson, DCT quantization matrices visually optimized for individual images, in: SPIE International Conference on Human Vision, Visual Processing and Digital DisplayIV, vol. 87, 1913, 1993.
[36] A.B. Watson, G.Y. Yang, J.A. Solomon, J.A. Villasenor, Visibility of wavelet quantization noise, IEEE Trans. Image Processing 6 (8) (August 1997) 1164–1175.