Signal Processing: *Image Communication* 26 (2011) 162–174

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

# Adaptive Block-size Transform based Just-Noticeable Difference model for images/videos

Lin Ma*, King Ngi Ngan, Fan Zhang, Songnan Li

*Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, Hong Kong*

ARTICLE INFO

*Article history:*
Received 3 May 2010
Accepted 20 February 2011
Available online 2 March 2011

*Keywords:*
Just-Noticeable Difference (JND)
Adaptive Block-size Transform (ABT)
Human Visual System (HVS)
Spatial Content Similarity (SCS)
Motion Characteristic Similarity (MCS)

ABSTRACT

In this paper, we propose a novel Adaptive Block-size Transform (ABT) based Just-Noticeable Difference (JND) model for images/videos. Extension from $8 \times 8$ Discrete Cosine Transform (DCT) based JND model to $16 \times 16$ DCT based JND is firstly performed by considering both the spatial and temporal Human Visual System (HVS) properties. For still images or INTRA video frames, a new spatial selection strategy based on the Spatial Content Similarity (SCS) between a macroblock and its sub-blocks is proposed to determine the transform size to be employed to generate the JND map. For the INTER video frames, a temporal selection strategy based on the Motion Characteristic Similarity (MCS) between a macroblock and its sub-blocks is presented to decide the transform size for the JND. Compared with other JND models, our proposed scheme can tolerate more distortions while preserving better perceptual quality. In order to demonstrate the efficiency of the ABT-based JND in modeling the HVS properties, a simple visual quality metric is designed by considering the ABT-based JND masking properties. Evaluating on the image and video subjective databases, the proposed metric delivers a performance comparable to the state-of-the-art metrics. It confirms that the ABT-based JND consists well with the HVS. The proposed quality metric also is applied on ABT-based H.264/Advanced Video Coding (AVC) for the perceptual video coding. The experimental results demonstrate that the proposed method can deliver video sequences with higher visual quality at the same bit-rates.



## 1. Introduction

Just-Noticeable Difference (JND) accounts for the smallest detectable difference between a starting and a secondary level of a particular sensory stimulus in psychophysics [1], which is also known as the difference limen or differential threshold. JND model has given a promising way to model the properties of the Human Visual System (HVS) accurately and efficiently in many image/video processing research fields, such as perceptual image/video compression [2–4,11,12], image/video perceptual quality evaluation [5–7,18], watermarking [8], and so on.

Generally automatic JND model for images can be determined in the spatial domain or the transform domain, such as Discrete Cosine Transform (DCT) and Discrete Wavelet Transform (DWT), or the combination of the two schemes [17]. JND models generated in the spatial domain [9,10], named as the pixel-based JND, mainly focus on the background luminance adaptation and the spatial contrast masking. Yang et al. [11,12] deduce the overlapping effect of luminance adaptation and spatial contrast masking to refine the JND model in [9]. However pixel-based JND models do not consider the human vision sensitivities of different frequency components. Therefore it cannot describe the HVS properties accurately. JND models generated in the transform domain, namely the subband-based JND, usually

* Corresponding author. Tel.: +852 26098255; fax: +852 26035558.
*E-mail addresses:* lma@ee.cuhk.edu.hk (L. Ma), knngan@ee.cuhk.edu.hk (K.N. Ngan), fzhang@ee.cuhk.edu.hk (F. Zhang), snli@ee.cuhk.edu.hk (S. Li).

0923-5965/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2011.02.002

incorporate all the major effecting factors, such as Contrast Sensitivity Function (CSF), luminance adaptation, and contrast masking. In [2], the JND model is developed from the spatial CSF. Then the DCTune JND model [3] is developed by considering the contrast masking. Hontsch and Karam [4] modify the DCTune model by replacing a single pixel with a foveal region, and Zhang et al. [13] refine the JND model by formulating the luminance adaptation adjustment and contrast masking. More recently, Wei and Ngan [15] incorporate new formulae of luminance adaptation, contrast masking, and Gamma correction to estimate the JND threshold in the DCT domain. Zhang et al. [17] propose to estimate the JND profile by summing the effects in DCT and spatial domain together.

In order to extend the JND profile from spatial to temporal, temporal characteristics of the HVS are considered. The previous works mostly focus on the perceptual differences between an original video sequence and its processed version [7,18]. Actually, the temporal HVS properties are highly correlated with the video signals, and can be approximated by a computational model. In [9,11,12], an empirical function based on the luminance difference between adjacent frames is proposed to model the temporal masking property. Kelly [22] proposes to measure the spatio-temporal CSF model at a constant retinal velocity, which is tuned to a particular spatial frequency. Daly [26] refines the model by taking the retina movement compensation into consideration. Jia et al. [23] estimate the JND for video sequences by considering both the spatio-temporal CSF and eye movements. Wei and Ngan [14,15] take the directionality of the motion into consideration to generate the temporal modulation factor.

However all the existing DCT-based JND models are calculated based on the $8\times 8$ DCT, which do not consider the perceptual properties of the HVS over transforms of different block sizes. Recently Adaptive Block-size Transform (ABT) has attracted researchers' attention for its coding efficiency in image and video compression [19,20,27]. It will not only improve the coding efficiency but also provide subjective benefits, especially for High Definition (HD) movie sequences from the viewpoint of subtle texture preservation [34,35]. Specifically, transforms of larger blocks can better exploit the correlation within the block, while the smaller block size is more suitable for adapting to the local structures of the image [16]. Therefore by incorporating ABT into the JND, an adaptive JND model is obtained, which can more precisely model the spatio-temporal HVS properties. Furthermore, since ABT has been adopted in current video coding standards, the ABT-based JND model for images/videos should be considered for applications such as video compression, image/video quality assessment, watermarking, and so on.

In this paper, extension from $8\times 8$ DCT-based JND to $16\times 16$ DCT-based JND is performed by conducting a psychophysical experiment to parameterize the CSF for the $16\times 16$ DCT. For still images or the INTRA video frames, a new spatial selection strategy based on the Spatial Content Similarity (SCS) is utilized to yield the JND map. For the INTER video frames, a temporal selection strategy based on the Motion Characteristic Similarity (MCS) is employed to determine the transform size for generating the JND map. Furthermore, its applications on image/video quality assessment and perceptual video coding are demonstrated to evaluate its efficiency in modeling the HVS properties.

The rest of the paper is organized as follows. Section 2 briefly introduces the extension procedure from the $8\times 8$ JND to $16\times 16$ JND. The proposed spatial and temporal selection strategies are presented in Section 3. The experimental performances are demonstrated and compared with the existing relevant models in Section 4. Finally, Section 5 concludes the paper.

## 2. JND model based on transforms of different block sizes

JND model in the DCT domain is determined by a basic visibility threshold $T_{basic}$, the spatial and temporal modulation factors. It can be expressed as

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

$$T_{spatio}(m,n,i,j)=T_{basic}(i,j)\times\alpha_{lum}(m,n)\times\alpha_{cm}(m,n,i,j), \tag{2}$$

where $k$ denotes the frame index of the video sequence, $(m,n)$ is the position of DCT block in the current frame, $(i,j)$ indicates the DCT coefficient position, and $\alpha_{lum}$ and $\alpha_{cm}$, denoting the luminance adaptation and contrast masking, respectively, constitute the spatial modulation factor. The video JND model $T$ is obtained by modulating spatial JND model $T_{spatio}$ with the temporal modulation factor $\alpha_{tempo}$.

### 2.1. Extension from $8\times 8$ JND to $16\times 16$ JND

Based on the band-pass property of the HVS in the spatial frequency domain, the HVS sensitivity characteristics are modeled in [21,28] as

$$H(w)=(a+bw)\cdot\exp(-cw), \tag{3}$$

where $w$ is the specified spatial frequency. JND is defined as the reciprocal of the HVS sensitivity characteristics given by (3). Hence the basic JND threshold can be modeled as [15]

$$T_{basic}(i,j)=\frac{s}{\phi_i\phi_j}\frac{\exp(cw_{ij})/(a+bw_{ij})}{\gamma+(1-\gamma)\cos^2\varphi_{ij}}, \tag{4}$$

where $s=0.25$ denotes the summation effect factor, $\gamma$ is set as 0.6, $\phi_i$ and $\phi_j$ are the DCT normalization factors, and $\varphi_{ij}=\arcsin(2w_{i0}w_{0j}/w_{ij}^2)$ indicates the directional angle of the corresponding DCT subband. $w_{ij}$ is the spatial frequency of the $(i,j)$ subband. As claimed and verified in [27], $4\times 4$ DCT does not contribute much to the efficiency of HD video coding. Since the proposed JND model aims at improving the performance of the perceptual HD video coding, only the $8\times 8$ and $16\times 16$ DCTs are considered to constitute the ABT-based JND model.

In order to extend the $8\times 8$ JND to $16\times 16$, the DCT block dimension $N$ is set to 16, and a psychophysical experiment is carried out to parameterize the three parameters $a$, $b$, and $c$ in (4). For a $512\times 512$ image, with all pixel intensities are set as 128, noises are injected into several selected $16\times 16$ DCT subbands to decide whether

it is visible. The following two aspects need to be considered for the DCT subbands selection:

a) The selected DCT subbands should cover the low, middle, and high frequency components. We select at least one DCT subband located on each row and each column. Consequently, the selected spatial frequencies are uniformly distributed within the HVS sensitivity frequency range.
b) At least one selected DCT subband should be located on each diagonal. Therefore, the spatial frequencies with all directions are covered, with which the HVS directional sensitivities are taken into account.

Furthermore, we consider the oblique effect [31], where human eyes are more sensitive to the horizontal and vertical frequency components than the diagonal ones. The sensitivities of horizontal and vertical components appear to be nearly symmetrical. Consequently, only the DCT subbands of the upper-right portion (as shown in Fig. 1) are chosen by considering the two aforementioned aspects. For the selected DCT subbands, several amplitude levels of the noises are pre-defined. The initial amplitude of the noise for each selected DCT subband is obtained by referring to the spatial CSF presented in [21,28]. Then the noise amplitude is tuned into several levels that make the noise range from invisible to obviously visible based on the preliminary measure of the authors. During the tuning process, according to the CSF, larger magnitude alternations of the noises are performed in the subbands with lower sensitivities. The oblique effect [31] also results in lower HVS sensitivities for the subbands with larger directional angles. Therefore, the noise amplitude alternations in the subbands with larger directional angles should be larger.

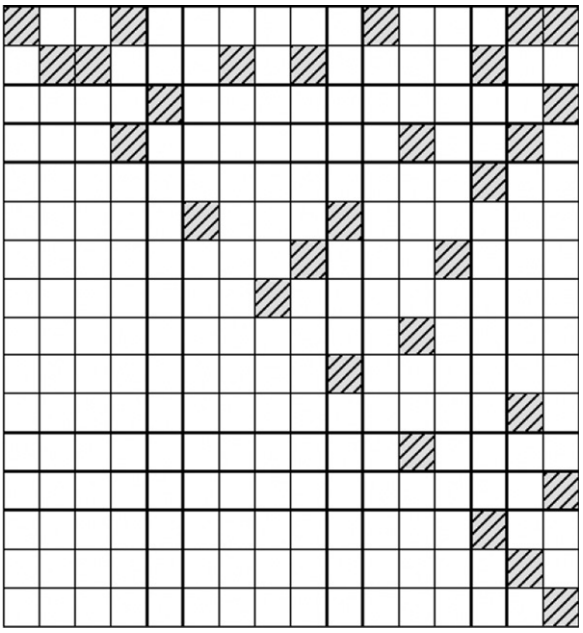

**Fig. 1.** Selected $16\times16$ DCT subbands for the psychophysical experiment (the shaded cells denote the selected DCT subbands).

Then the noise, with its amplitude as one of the pre-defined levels, is inserted into the selected DCT subbands of the image. The original image and the processed one (with noise insertion) are juxtaposed on the screen. Ten viewers vote on whether the noise is visible. If half of them choose "yes", the noise amplitude is recognized as above the JND threshold. A smaller amplitude noise will be inserted. Otherwise, a larger one will be chosen for injection. Finally, the obtained thresholds of the selected DCT subbands are employed to minimize the least squared error as given in (5) to parameterize $(a,b,c)$

$$(a,b,c)=\arg\min\sum_{w_{ij}}[T_{w_{ij}}-T_{basic}(i,j)]^2, \tag{5}$$

where $T_{w_{ij}}$ is the JND threshold obtained from the psychophysical experiment. The above procedure yields the parameters, $a=0.183$, $b=0.165$, and $c=0.16$ for the $16\times16$ JND model.

JND is influenced by the intensity scale of the digital image. It is reported that higher visibility threshold occurs in either dark or bright regions compared with the medium brightness regions. The luminance adaptation factor $\alpha_{lum}$ forms a U-shape curve [4,13,17,29,32]. Therefore, an empirical formula [15] is employed to depict the $\alpha_{lum}$

$$\alpha_{lum}=\begin{cases}(60-I_{ave})/150+1 & I_{ave}\le 60\\ 1 & 60<I_{ave}<170\\ (I_{ave}-170)/425+1 & I_{ave}\ge 170\end{cases}, \tag{6}$$

where $I_{ave}$ denotes the average intensity of the DCT block.

For the contrast masking factor, a block-based method [13–17] is utilized to accurately describe the different masking properties of different block categories. These methods categorize the blocks into different block types according to the DCT subband energy [13,17,23] or image spatial characteristics [14,15]. As in [15], we categorize the image block into three types, namely PLANE, EDGE, and TEXTURE, based on the proportion of the edge pixels in the $16\times16$ macroblock. The macroblock categorization is defined according to

$$Categ_{16}=\begin{cases}PLANE & \sum_{EP}<16\\ EDGE & 16\le\sum_{EP}\le 52\\ TEXTURE & \sum_{EP}>52\end{cases}, \tag{7}$$

where $\sum_{EP}$ denotes the number of edge pixels in a given macroblock. Considering the block category and the intra-band masking effect [13,15,17], the contrast masking factor $\alpha_{cm}$ for $16\times16$ JND is obtained. Detailed information about the contrast masking scheme can be found in [33].

For the temporal modulation factor $\alpha_{tempo}$, Robson [24] has shown that the form of the sensitivity fall-off at high spatial frequencies is independent of the temporal frequency and vice versa, while a sensitivity fall-off at low spatial frequencies occurs only when the temporal frequency is also low and vice versa. In [14,15], it demonstrates that the logarithms of the temporal contrast sensitivity values follow approximately the same slope (nearly $-0.03$) for different spatial frequencies. By further considering the band-pass characteristic at the lower spatial frequencies [22], the temporal modulation factor

is derived as

$$\alpha_{tempo} = \begin{cases} 1 & w_s < 5\,\text{cpd and } w_t < 10\,\text{Hz} \\ 10^{-0.03(w_t-10)} & w_s < 5\,\text{cpd and } w_t \geq 10\,\text{Hz} \\ 10^{-0.03w_t} & w_s \geq 5\,\text{cpd} \end{cases}, \quad (8)$$

where $w_s$ and $w_t$ denote the spatial and temporal frequency, respectively. $w_s$ is determined by the transform size and the viewing distance, while $w_t$ relies on both the spatial frequency $w_s$ and the motion information [25], which is approximated by the block-based motion estimation [14,15].

### 2.2. Why introduce ABT into JND?

The HVS sensitivities over transforms of different block sizes are illustrated in Fig. 2. Firstly, as explained before, the HVS sensitivities are constrained within a spatial frequency range, which is approximately from 0 to 25 cpd. Therefore, the HVS sensitivities can be modeled more accurately using a larger number of frequency bases. As shown in Fig. 2, the HVS sensitivities for the $8\times8$ DCT are very sparse compared with the ones for the $16\times16$ DCT. The HVS sensitivity properties cannot be accurately modeled by only employing the $8\times8$ DCT based sensitivity function. Secondly, the HVS directional sensitivities need to be considered. From Fig. 2, many points of the $16\times16$ sensitivities, which have nearly the same spatial frequency but different *Angle* information, demonstrate different HVS sensitivities. The higher the *Angle* information, the lower the HVS contrast sensitivities, which matches the HVS oblique effect [31]. However for the sensitivity values of $8\times8$, there are very few points with different *Angle* information. It cannot accurately represent the HVS directional properties. Considering the two aforementioned aspects, the sensitivities of $16\times16$ can more accurately model the HVS properties. It can help to find more accurate parameters $a$, $b$, and $c$ in (4) for depicting the HVS sensitivities.

From the viewpoint of energy compaction, a larger block size transform takes advantage of exploiting the correlation within a block. On the other hand, the smaller one is more adaptive to the local structural changes. Therefore, transforms of different block sizes adapting to the image content play a very important role in image/video processing tasks, especially in image/video compression. It has been claimed [35] that ABT can provide subjective benefits, especially for HD movie sequences from the viewpoint of subtle texture preservation, such as keeping film details and grain noises, which are crucial to the subjective quality [36]. We believe that ABT-based JND model will make the HVS properties modeling more accurate, and benefit the perceptual-related image/video applications.

As ABT has been adopted into the current video coding schemes such as H.264, it is therefore necessary to develop the ABT-based JND model. It can be easily incorporated into the current coding standards. In [11,12] perceptual video coding schemes employing the $8\times8$ DCT JND have been proposed. With the proposed ABT-based JND model, a more efficient perceptual coding scheme can be developed.

## 3. Selection strategy between transforms of different block sizes

In the last section, the formulations of the JND models for the $8\times8$ and $16\times16$ DCT transforms are described. Decision method for the proper transform size, i.e., $8\times8$ or $16\times16$, will be discussed in this section.

### 3.1. Spatial selection strategy for transforms of different block sizes

As the selection strategy is designed for each macroblock, the image is firstly divided into $16\times16$ macroblocks. For each macroblock, two JND models based on $8\times8$ and $16\times16$ DCT are obtained. For the still images or INTRA video frames, where there is no motion information, we propose the Spatial Content Similarity (SCS) to measure the image content homogeneity between a macroblock and its sub-blocks

$$SCS = \sum_{i=1}^{4} (Categ_{16} = \quad = Categ_8^i), \quad (9)$$

where $Categ_{16}$ and $Categ_8^i$ denote the categories of the macroblock and the $i$th $8\times8$ sub-block, respectively. SCS indicates the number of $8\times8$ sub-blocks with the same

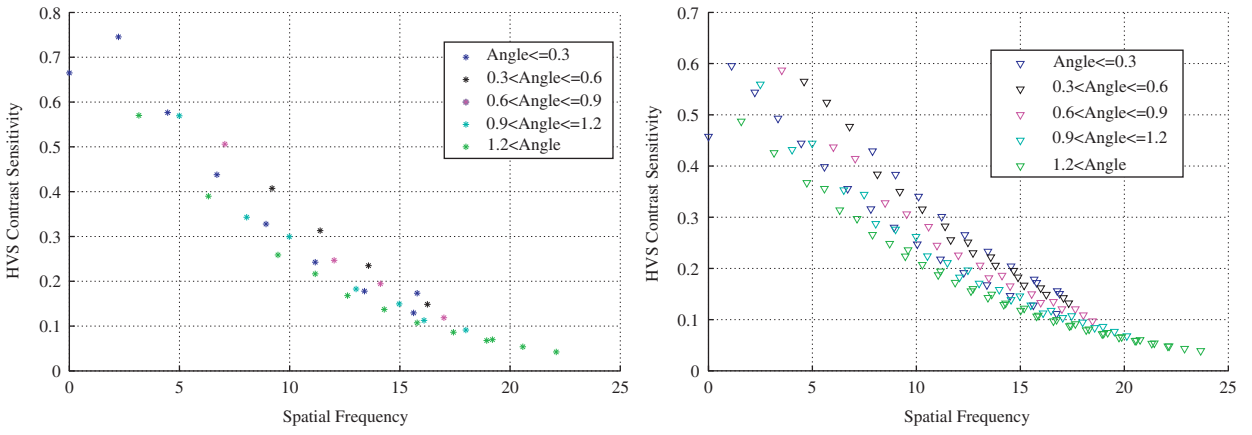


**Fig. 2.** Modeled HVS sensitivities over transforms of different block sizes by Eq. (4) (left: the HVS sensitivity over $8\times8$ DCT in [15]; right: the HVS sensitivity over $16\times16$ DCT).

categorization as the macroblock, which they belong to. If SCS is equal to 4, referring to the homogeneous content within the macroblock, the JND model based on $16 \times 16$ DCT will be utilized to yield the resulting JND model. On the contrary, if SCS is smaller than 4, the $8 \times 8$ JND model will be employed for adapting the local structures within the sub-blocks. The results of spatial selection strategy for LENA and PEPPERS are shown in Fig. 3. Most of the PLANE regions employ the $16 \times 16$ JND model, while the areas with local structure changes utilize $8 \times 8$ JND model. The results are consistent with the energy compaction capabilities of the $8 \times 8$ and $16 \times 16$ DCTs.

### 3.2. Temporal selection strategy for transforms of different block sizes

For INTER video frames, the JND model needs consider not only the spatial but also the temporal information. Therefore, we should include the temporal motion characteristics, which are depicted by motion vectors of different size blocks.

Based on the motion vectors of different size blocks, we propose a Motion Characteristics Similarity (MCS) to measure the motion consistency between a macroblock and its sub-blocks, which is expressed as

$$MCS = \sum_{i=1}^{4} ||Mv_8^i - Mv_{16}||_2^2/4, \tag{10}$$

where $Mv_8^i$ denotes the motion vector of the $i$th $8 \times 8$ sub-block, $Mv_{16}$ is the motion vector of the $16 \times 16$ macroblock, and $||\ \ ||_2^2$ denotes the Euclidean distance between the two motion vectors. Considering the spatial SCS and temporal MCS, we can make decision on which transform block size to use for the resulting JND.

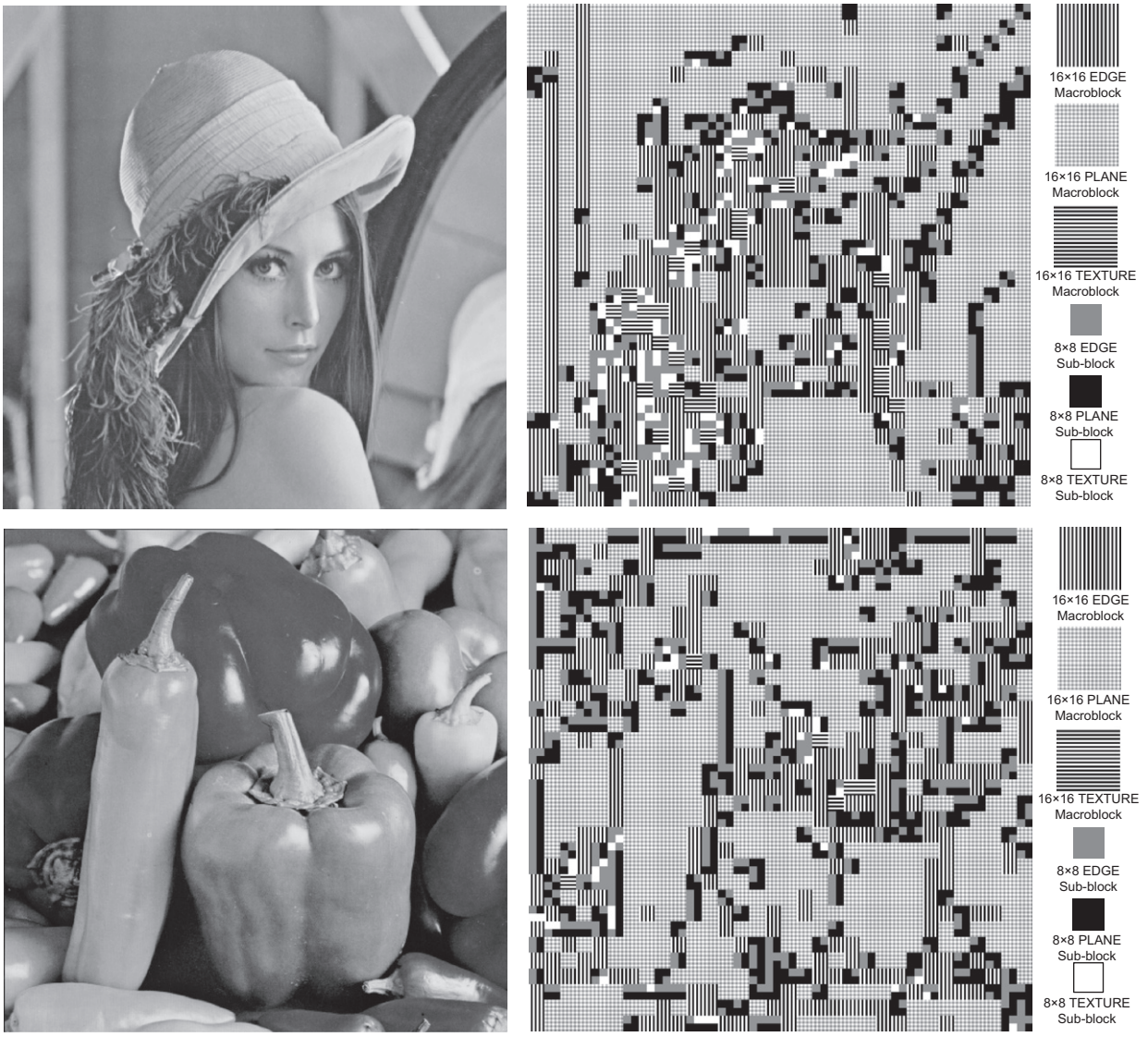


**Fig. 3.** Spatial selection results of LENA and PEPPERS (left: the original image; right: spatial selection results in terms of block category and transform block size).

If the calculated MCS is smaller than a threshold, it is deemed that the motion characteristics of the macroblock and its corresponding sub-blocks are nearly the same. In this paper, we empirically set the threshold as 1.25 pixels. When SCS is equal to 4 and MCS smaller than the threshold, the macroblock is considered to be a single unit. Therefore, $16\times16$ DCT based JND is utilized to generate the JND model. On the other hand, if the MCS is larger than the threshold, indicating that motion vectors of the macroblock and its sub-blocks are quite different, the macroblock should be separated into 4 sub-blocks because of the smaller SCS and larger MCS. The $8\times8$ DCT based JND for each sub-block is then employed to obtain the resulting JND model.

In order to further test the consistency between the spatial and temporal selection strategies, the Hit Ratio (HR) curve is used to demonstrate the hit rate for each INTER video frame. Firstly, we record the macroblock JND types determined by the aforementioned spatial and temporal selection strategies, respectively. The hit rate $h$ of each INTER frame measures the percentage of the macroblocks (as determined by the spatial and temporal selection strategies) are identical. In this case, the transform of the same block size is selected for a macroblock to generate the resulting JND model. The HR curves for each INTER frame of several typical CIF ($352\times288$) sequences are illustrated in Fig. 4. The hit rates $h$ are high, corresponding to the fact that the proposed temporal selection strategy accords well with the spatial selection strategy. The proposed selection strategy is efficient for depicting both spatial image content information and temporal video motion characteristics. Furthermore, the hit rates of FOOTBALL and FOREMAN are a bit lower than the other sequences, with the average hit rate as 77%. The reason is that both sequences contain high motion characteristics. Therefore, the consistency between spatial and temporal characteristics tends to be low. On the other hand, as the motion appears slightly in the other sequences, the hit rate becomes much higher, with the average value as 93%.

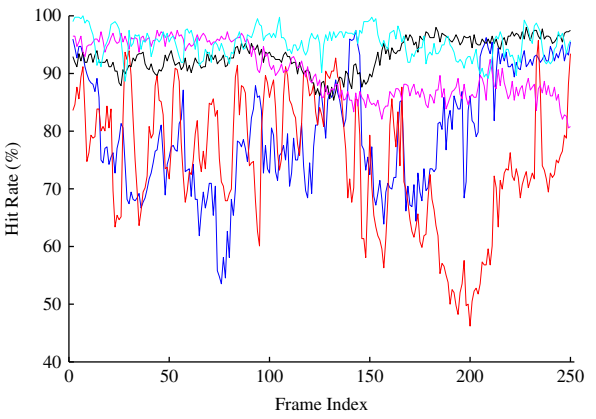


**Fig. 4.** HR curves of the macroblocks for each INTER frame of the test video sequences.

## 4. Performance

### 4.1. JND model performance evaluation

In order to demonstrate the efficiency of the proposed ABT-based JND model, the noise is injected into each DCT coefficient of each image or video frame to evaluate the HVS error tolerance ability

$$\tilde{I}_{typ}(k,m,n,i,j)=I_{typ}(k,m,n,i,j)+R\cdot T_{typ}(k,m,n,i,j), \tag{11}$$

where $\tilde{I}_{typ}$ is the noise-contaminated DCT coefficient, which is located on the $(i,j)$th position of the $(m,n)$th block in the $k$th frame. For still images, $k$ is set as 0. $R$ takes the value of $+1$ or $-1$ randomly to avoid introducing a fixed pattern of changes. $T_{typ}$ is the JND threshold obtained by the proposed ABT-based scheme, and $typ$ denotes the final transform block size to generate the resulting JND model.

#### 4.1.1. Evaluation on images

We test the proposed JND model on several typical $512\times512$ images and $768\times512$ Kodim images [54]. We compare the proposed method with Yang et al.'s method [11], which evaluated the JND in the image domain, and Wei and Ngan's method [15], which calculates the JND in the DCT domain. Comparisons in terms of PSNR are listed in Table 1, which shows that our proposed JND method yields smaller PSNR values compared with other JND models. Here if the image visual quality stays the same as the original one, it implies that our JND model can tolerate more distortions.

In order to provide a more convincing evaluation of the proposed JND model, subjective tests are conducted to assess the perceptual qualities of the noise-contaminated images. In the subjective test, two images were juxtaposed on the screen. One is the original image as the reference and the other is the noise-inserted version. In this experiment, the viewing monitor is a Viewsonic Professional series P225fb CRT display. The viewing distance is set as 4 times the image height. Ten observers (half of them are experts in image processing and the other half are not) are asked to offer their opinions on the subjective quality of the images, by following the quality comparison scale shown in Table 2. Their average subjective values are calculated to indicate the image visual quality, which is illustrated in Table 3. The mean and variance values of the subjective scores are also calculated. According to the

**Table 1**
PSNR comparison between different JND models.

| Image | Yang (dB) | Wei (dB) | Proposed JND (dB) |
|---|---|---|---|
| BABOON | 32.53 | 28.38 | 27.46 |
| BARBARA | 31.35 | 29.49 | 29.02 |
| BRIDGE | 30.96 | 29.01 | 28.53 |
| LENA | 32.72 | 29.97 | 29.51 |
| PEPPERS | 30.78 | 29.99 | 29.66 |
| Kodim06 | 32.21 | 29.02 | 28.61 |
| Kodim08 | 31.21 | 29.11 | 28.73 |
| Kodim13 | 30.59 | 28.75 | 28.42 |
| Kodim14 | 30.00 | 29.41 | 29.14 |
| Kodim21 | 32.15 | 29.43 | 29.06 |

quality comparison scale in Table 2, the smaller the subjective scores, the better quality of the noise contaminated images. The proposed method has the smallest mean value (only 0.37), demonstrating the best performance. From the subjective results, Yang et al.'s method can generate higher quality images, such as BABOON, Kodim13, and Kodim14. These images exhibit more texture information. For the images with much plane or edge information, such as PEPPERS and Kodim21, the visual quality will degrade significantly. Our method generates smaller variance compared with the other methods, indicating that the proposed scheme performs more consistently over images of different types. The noise-inserted images generated by our method can be found in [37].

#### 4.1.2. Evaluation on videos

The proposed JND model is evaluated on several typical CIF ($352 \times 288$) video sequences, with a frame rate of 30 fps. In our experiments, 250 frames of each sequence are tested, with the first frame as INTRA and the rest as INTER frames. We also compare the proposed method with Yang et al.'s [11], and Wei and Ngan's [15] JND models. Since we have evaluated the efficiency of ABT-based JND model for images, here only the average PSNR of the INTER frames is calculated. Comparisons in terms of PNSR are listed in Table 4. It is observed that the proposed JND model yields smaller PSNR values compared with other JNDs. It shows that the ABT-based JND model can tolerate more distortions.

The subjective test is conducted to further assess the perceptual quality of the noise-contaminated videos. Double Stimulus Continuous Quality Scale (DSCQS) method, as specified in ITU-R BT.500 [30], is employed to evaluate the perceptual quality. Two sequences are presented to viewers, one of which is original and the other is processed. Ten viewers (half of them are experts in image/video processing and the other half are not) are asked to offer their opinions. Mean Opinion Score (MOS) is scaled for testers to vote: Bad (0–20), Poor (20–40), Fair (40–60), Good (60–80), and Excellent (80–100). The difference between subjective scores of the original and noise-injected video sequence is calculated as the Differential Mean Opinion Score (DMOS). Hence, the smaller the DMOS, the higher is the quality of the noise-contaminated video. The testing conditions are the same as the image evaluation process. Detailed subjective test results are depicted in Table 5. The mean DMOS value of the proposed scheme is 6.89, which is smaller than Yang et al.'s and Wei et al.'s methods. It reflects that our proposed method can generate similar quality videos with the original ones. Also it can be found that variance of the DMOS value is the smallest. Compared with the other methods, our approach delivers more consistent results for both the fast-moving video sequences, e.g., FOOTBALL and STEFAN, and the slightly moving video sequences, e.g., SILIENCE and PARIS.

**Table 2**
Rating criterion for subjective quality evaluation.

| Subjective score | Descriptions |
|---|---|
| −3 | The right on is much worse than the left one |
| −2 | The right one is worse than the left one |
| −1 | The right one is slightly worse than the left one |
| 0 | The right one has the same quality as the left one |
| 1 | The right one is slightly better than the left one |
| 2 | The right one is better than the left one |
| 3 | The right one is much better than the left one |

**Table 3**
Subjective evaluation results (L: noise-contaminated image by different JND models; R: original image).

| Image | Yang | Wei | Proposed JND |
|---|---|---|---|
| BABOON | 0.5 | 0.2 | 0.2 |
| BARBARA | 1.2 | 0.4 | 0.5 |
| BRIDGE | 0.7 | 0.3 | 0.3 |
| LENA | 0.8 | 0.3 | 0.4 |
| PEPPERS | 1.0 | 0.4 | 0.4 |
| Kodim06 | 1 | 0.6 | 0.4 |
| Kodim08 | 1.2 | 0.5 | 0.6 |
| Kodim13 | 0.4 | 0.4 | 0.3 |
| Kodim14 | 0.5 | 0.3 | 0.2 |
| Kodim21 | 1.5 | 0.6 | 0.4 |
| Average | 0.88 | 0.40 | 0.37 |
| Variance | 0.362 | 0.133 | 0.125 |

### 4.2. Visual quality metric based on the proposed JND model

Traditional error measures for images/videos, such as Mean Square Error (MSE) and Peak Signal-to-Noise Ratio

**Table 4**
PSNR comparison between different JND models.

| Video | Yang (dB) | Wei (dB) | Proposed JND (dB) |
|---|---|---|---|
| TEMPETE | 31.68 | 27.42 | 27.04 |
| FOOTBALL | 34.43 | 28.39 | 28.17 |
| FOREMAN | 35.29 | 28.29 | 28.02 |
| MOBILE | 33.10 | 27.48 | 26.93 |
| SILENCE | 34.43 | 28.26 | 27.93 |
| TABLE | 36.37 | 27.81 | 27.33 |
| STEFAN | 35.20 | 27.83 | 27.38 |
| PARIS | 33.56 | 27.60 | 27.07 |
| FLOWER | 35.57 | 27.18 | 26.80 |
| WATERFALL | 33.88 | 27.83 | 27.52 |

**Table 5**
Subjective evaluation results (DMOS for noise-contaminated video sequences).

| Video | Yang | Wei | Proposed JND |
|---|---|---|---|
| TEMPETE | 7.3 | 6.6 | 6.4 |
| FOOTBALL | 7.6 | 6.2 | 5.6 |
| FOREMAN | 13.2 | 9.2 | 8.3 |
| MOBILE | 9.7 | 7.0 | 7.1 |
| SILENCE | 13.9 | 9.7 | 8.5 |
| TABLE | 6.9 | 6.2 | 5.2 |
| STEFAN | 7.2 | 6.0 | 5.4 |
| PARIS | 14.2 | 9.4 | 9.2 |
| FLOWER | 13.2 | 8.2 | 7.4 |
| WATERFALL | 6.5 | 5.6 | 5.8 |
| Average | 9.97 | 7.41 | 6.89 |
| Variance | 3.269 | 1.565 | 1.429 |

(PSNR), do not correlate well with the HVS for evaluating the image/video perceptual quality [38–43]. In this section, we design a very simple visual quality metric based on the proposed ABT-based JND model, which is defined as

$$\begin{aligned} Diff_{typ}(k,m,n,i,j) &= \begin{cases} 0, & \text{if } \left|I_{typ}(k,m,n,i,j)-I^{D}_{typ}(k,m,n,i,j)\right| \le T_{typ}(k,m,n,i,j) \\ \left|I_{typ}(k,m,n,i,j)-I^{D}_{typ}(k,m,n,i,j)\right|-T_{typ}(k,m,n,i,j), & \text{otherwise} \end{cases} \\ P_{dist}(k,m,n,i,j) &= \tau_{typ}\frac{Diff_{typ}(k,m,n,i,j)}{T_{typ}(k,m,n,i,j)} \\ V_Q &= 10\log_{10}\left(\underset{(k,m,n,i,j)}{mean}\left(P^2_{dist}(k,m,n,i,j)\right)\right), \end{aligned} \tag{12}$$

where $T_{typ}$ is the ABT-based JND, $typ$ denotes the transform block size for generating the JND, $I_{typ}$ is the DCT coefficients of the reference image/frame, $I^{D}_{typ}$ denotes the DCT coefficients of the distorted image/frame, and $Diff_{typ}$ denotes the DCT coefficient differences between the reference image/frame and the distorted one by considering the HVS error tolerance ability. Since the JND denotes the threshold for detecting the perceptual difference (as demonstrated in Section 4.1), the distortions below the JND thresholds cannot be perceived by the human eyes. They need not be accounted in measuring the visual quality, where the visual difference is set as 0. In (12) above, only the distortions larger than the JND thresholds are calculated for measuring the visual quality. The adjustable parameter $\tau_{typ}$ is introduced according to the different energy compaction properties, which are determined by the coding gains of different block transforms. The coding gain [51] for the block transform is defined as

$$G_{TC} = 10\log_{10}\left[\frac{\frac{1}{N}\sum_{i=0}^{N-1}\sigma_i^2}{\left(\prod_{i=0}^{N-1}\sigma_i^2\right)^{\frac{1}{N}}}\right], \tag{13}$$

where $N$ is the number of the transform subbands, and $\sigma_i^2$ is the variance of each subband $i$, for $0 \le i \le N-1$. Then $\tau_{typ}$ is defined according to

$$\tau_{typ} = \begin{cases} G^8_{TC}/G^{16}_{TC}, & typ \text{ is } 16\times 16 \\ 1, & typ \text{ is } 8\times 8 \end{cases}, \tag{14}$$

where $G^8_{TC}$ and $G^{16}_{TC}$ denote the coding gains of $8\times 8$ and $16\times 16$ DCT, respectively. After testing on the reference images of the LIVE database [44], the coding gain ratio $G^8_{TC}/G^{16}_{TC}$ appears to be nearly the same. Therefore, we simply set it as 0.95. $P_{dist}$ is the distortion masked by the proposed ABT-based JND model. The visual quality metric $V_Q$ is obtained by aggregating the $P_{dist}$ of all the transform blocks in one frame. If we evaluate the visual quality metric of an image, only the spatial JND model is employed and $k$ is set as 0. If the video quality is assessed, the proposed metric employs the spatio-temporal JND model. In our approach, the visual quality of each frame is measured individually. Hence the visual quality of the whole video sequence is given by the mean quality value of all the frames.

We have tested the performance of the proposed metric, as well as the state-of-the-art image quality metrics, such as SSIM [39], VIF [40], and VSNR [42] over the image subjective quality databases LIVE [44], A57 [46], and IRCCyN/IVC [45]. Table 6 lists some major characteristics of the image databases. They contain the most prevailing distortions, such as JPEG, JPEG 2000, blurring, additive Gaussian noise, and so on. Each distorted image in these subjective quality databases is assigned a subjective score, e.g., DMOS for LIVE image/video database, MOS for the IRCCyN/IVC database, and perceived distortion for the A57 database. These subjective scores are obtained from subjective viewing tests where many observers participated and provided their opinions on the visual quality of each distorted image. These subjective scores are regarded as the ground truths for evaluating the performances of different visual quality metrics. We follow the performance evaluation procedure adopted in Video Quality Experts Group (VQEG) HDTV test [49] and that in [48]. After non-linear mapping, three standard criteria named as Correlation Coefficient (CC), Spearman-Rank Order Correlation Coefficient (SROCC), and Root Mean Square prediction Error (RMSE) are employed to

**Table 6**
Major characteristics of the subjective image/video databases.

| | Database | No. of reference images/videos | No. of distortion types | No. of distorted images/videos | Typical image/video size |
|---|---|---|---|---|---|
| Image | LIVE | 29 color images | 5 (JPEG, JPEG2000, blurring, fast-fading, and additive Gaussian noise) | 779 color images | $768\times 512$/ $512\times 768$ |
| | IRCCyN/IVC | 10 color images | 4 (JPEG, JPEG2000, LAR coding, and blurring) | 185 color images | $512\times 512$ |
| | A57 | 3 gray images | 6 (JPEG, JPEG2000, additive Gaussian noise, blurring, etc.) | 54 gray images | $512\times 512$ |
| Video | LIVE | 10 YUV 420 sequences | 4 (Wireless distortion, IP distortion, H.264 compression, and MPEG-2 compression) | 150 YUV 420 sequences | $768\times 432$ |

**Table 7**
Performances of different image quality metrics.

| Database | | PSNR | SSIM | VSNR | VIF | $16 \times 16$ JND | $8 \times 8$ JND | Proposed |
|---|---|---|---|---|---|---|---|---|
| LIVE image | CC | 0.8716 | 0.904 | 0.637 | 0.956 | 0.907 | 0.921 | 0.933 |
| | SROCC | 0.8765 | 0.910 | 0.648 | 0.958 | 0.911 | 0.925 | 0.934 |
| | RMSE | 13.392 | 11.68 | 21.13 | 7.99 | 11.544 | 10.663 | 9.881 |
| IRCCyN/IVC | CC | 0.704 | 0.776 | 0.800 | 0.903 | 0.910 | 0.893 | 0.913 |
| | SROCC | 0.679 | 0.778 | 0.798 | 0.896 | 0.903 | 0.885 | 0.909 |
| | RMSE | 0.866 | 0.769 | 0.731 | 0.524 | 0.503 | 0.548 | 0.498 |
| A57 | CC | 0.644 | 0.415 | 0.942 | 0.618 | 0.877 | 0.910 | 0.913 |
| | SROCC | 0.570 | 0.407 | 0.936 | 0.622 | 0.870 | 0.891 | 0.901 |
| | RMSE | 0.192 | 0.224 | 0.083 | 0.193 | 0.118 | 0.103 | 0.101 |

evaluate the metric performances. According to their definitions [49], the larger the CC and SROCC, the better is the visual quality metric. In contrast, the smaller the RMSE, the better is the visual quality metric. The performances of different image quality metrics are illustrated in Table 7. And the scatter-plots of different quality metrics are illustrated in Fig. 5 and [37]. It can be observed that our proposed method scatter closely around the fitted curve, which indicates a good performance.

Furthermore, we test the proposed visual quality metric on the LIVE video subjective quality database [47], whose major characteristics are listed in Table 6. The video subjective quality index is obtained by averaging the frame $V_Q$ scores, the same as PSNR, SSIM, and VIF. And we also compare with the most popular video quality metrics VQM [41] and MOVIE [43]. As usual, after non-linear mapping, CC, SROCC, and RMSE are employed for evaluating the performances, as shown in Table 8. It is observed that the proposed method outperforms other video quality metrics, while slightly inferior to MOVIE. The scatter-plots are provided in Fig. 6 and [37]. The results of our proposed method scatter closely around the fitted curve, indicating a good performance.

We have implemented a visual quality metric solely based on $16 \times 16$ or $8 \times 8$ JND. It means that both the INTRA and INTER JND models are generated by only considering the $16 \times 16$ or $8 \times 8$ JND. There is no selection strategy. The experimental results of these visual quality metrics are illustrated in Tables 7 and 8. The two metrics based on $16 \times 16$ and $8 \times 8$ JNDs can efficiently evaluate the image/video quality better than PSNR. However, both of them are inferior to the proposed ABT-based metric. The $16 \times 16$ JND can more accurately model the HVS property. However, for distorted images/videos in practical applications, the $8 \times 8$ DCT is quite frequently utilized, such as JPEG-coded images and MPEG2-coded videos. By considering the $8 \times 8$ JND, the distortion can be more precisely depicted, which can improve the quality metric performance. Therefore, the $8 \times 8$ and $16 \times 16$ JNDs are considered for designing the visual quality metric. Thus we introduce the ABT-based JND into the visual quality metric.

From the test results, our proposed visual quality metric performs comparably with the stat-of-the-art quality metrics. It clearly demonstrates that the proposed ABT-based JND model can incorporate the HVS properties

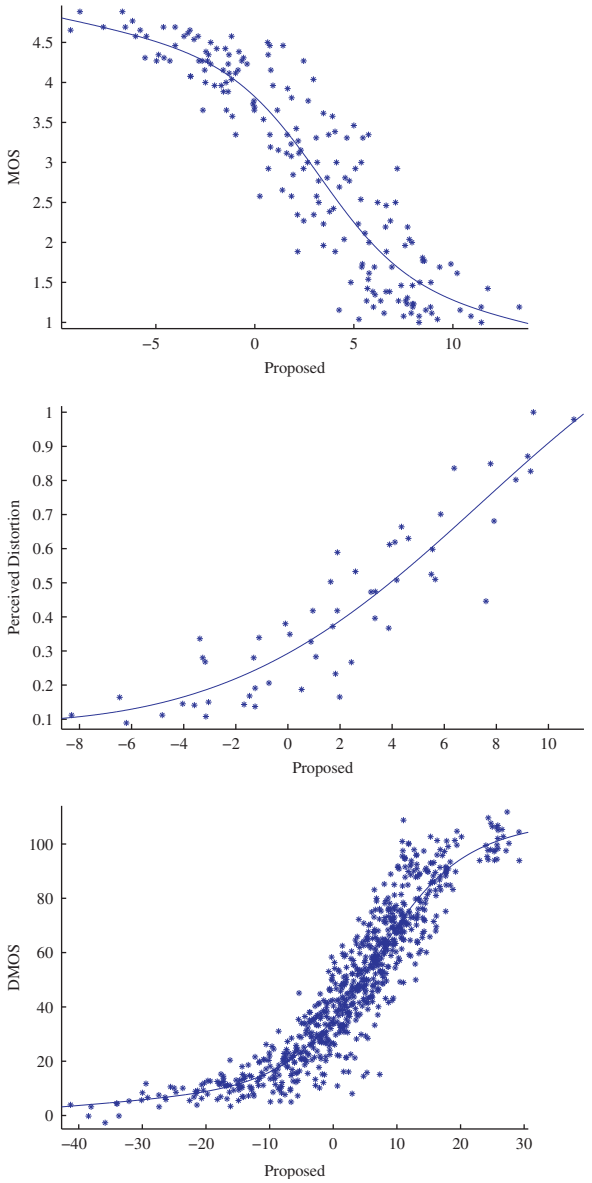


**Fig. 5.** Scatter plots of the DMOS values versus model prediction on the image subjective databases (top: IRCCyN/IVC image database; middle: A57 image database; bottom: LIVE image database).

**Table 8**
Performances of different video quality metrics.

| Database | | PSNR | SSIM | VIF | VQM | MOVIE[a] | 16 × 16 JND | 8 × 8 JND | Proposed |
|---|---|---|---|---|---|---|---|---|---|
| Live video | CC | 0.5398 | 0.4999 | 0.5735 | 0.7160 | 0.8116 | 0.611 | 0.602 | 0.780 |
| | SROCC | 0.5234 | 0.5247 | 0.5564 | 0.7029 | 0.7890 | 0.585 | 0.579 | 0.761 |
| | RMSE | 9.241 | 9.507 | 8.992 | 7.664 | – | 8.692 | 8.764 | 6.935 |

[a] CC and SROCC value of MOVIE are obtained directly from [50], which does not provide the RMSE value.

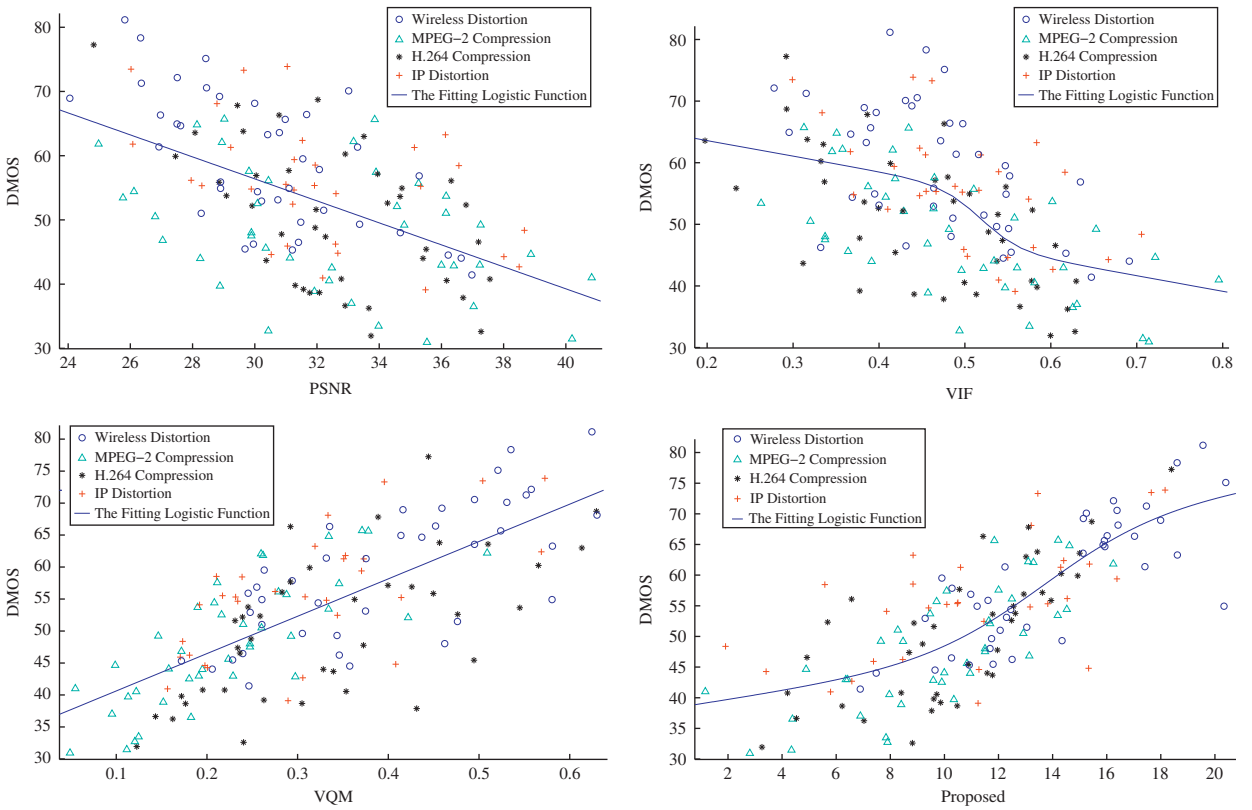


**Fig. 6.** Scatter plots of the DMOS values versus model prediction on the LIVE video subjective database (top left: PSNR; top right: VIF; bottom left: VQM; bottom right: the proposed metric).

into the context of visual quality assessment. It is found that SSIM and VIF perform very well on image quality evaluation. But they fail in assessing the video subjective quality. The reason is that SSIM and VIF succeed to depict the spatial distortions, but fail to capture the temporal distortions. That is the reason why VQM outperforms SSIM and VIF, for it has considered the temporal effect. However, the temporal effect in VQM is simply modeled by the fame differences. It cannot efficiently depict the temporal distortions, resulting in a slightly better performance. MOVIE is developed by considering the complex temporal and spatial distortion modeling, leading to the best performance. However, it is very complex and time-consuming, hence cannot be easily applied in practical applications. As the proposed visual quality metric has modeled both the spatial and temporal HVS properties, it performs comparably with VIF and MOVIE. It maintains a very simple formulation in DCT domain. Therefore, the proposed visual quality metric can be easily applied to image/video applications, especially the perceptual video coding.

### 4.3. Perceptual video coding based on the ABT-based JND

In this section, the ABT-based JND is incorporated into the video coding scheme for pursuing higher visual quality with the same bit-rates according to

$$Re_{typ}(k,m,n,i,j) = DCT_{typ}\{I(k,m,n,i,j) - I_{pre}(k_{ref},m,n,i,j)\}$$
$$Re'_{typ}(k,m,n,i,j) = \begin{cases} 0, & \text{if } |Re_{typ}(k,m,n,i,j)| \le JND_{typ}(k,m,n,i,j) \\ sign(Re_{typ}(k,m,n,i,j))(|Re_{typ}(k,m,n,i,j)| - JND_{typ}(k,m,n,i,j)), & \text{otherwise} \end{cases}, \quad (15)$$

where $I$ is the macroblock to be encoded, $I_{pre}$ is the predicted macroblock by inter motion estimation or intra prediction, $typ$ denotes the transform size ($8\times8$ or $16\times16$ DCT), $Re_{typ}$ is the DCT coefficients of the prediction error, $JND_{typ}$ is the calculated JND thresholds for different transform sizes. According to the definition of JND and the quality metric in (12), the HVS cannot detect the distortions, which are smaller than the JND threshold. Therefore, the distortions below the JND threshold need not be accounted. The perceptual redundancies in the video signals are removed according to (15), which will not cause any visual degradation. Then the resulting DCT coefficients $Re'_{typ}$ without perceptual redundancies are encoded.

For the traditional video coding strategy, MSE is utilized to calculate the distortions in Rate-Distortion Optimization (RDO), which is justified to be inconsistent with the HVS perception [38]. In this paper, $P_{dist}$ is employed for depicting the HVS responses of the distortions, which is defined as

$$P_{dist}(k,m,n,i,j)=\tau_{typ}\frac{Re'_{typ}(k,m,n,i,j)}{T_{typ}(k,m,n,i,j)}. \tag{16}$$

The sum squared error of $P_{dist}$ will be utilized as the distortion measurement for the Modified RDO (M-RDO) process. As demonstrated in Section 4.2, $P_{dist}$ correlates better with the HVS than MSE, which is believed to benefit the perceptual video coding. During the encoding process, a suitable $\lambda$ needs to be determined for the M-RDO process

$$\text{Cost}=D_p+\lambda R, \tag{17}$$

where $D_p$ is the sum squared error of $P_{dist}$, and $R$ denotes the bit-rate. In our experiments, four 720P sequences, Crew, Harbor, Sailormen, and Spincalendar are encoded with the H.264 platform provided by [27]. The test conditions are listed in Table 9 (only 100 frames), with QP ranging from 28 to 40. Then $D_p$ is used to evaluate the coded sequences. According to the derivation in [52,53], the optimal $\lambda$ is set as

$$\lambda=-\frac{dD_p}{dR}. \tag{18}$$

In our experiments, the tangent slopes at each identical QP point of the four testing sequences appear to be similar. Therefore, the average value of the tangent slopes is employed as $\lambda$ in the M-RDO process.

In the encoding process, the M-RDO process is employed to determine the best transform type. We believe that the proposed selection strategy has strong ties with the M-RDO process. For one macroblock, if the spatial content is homogenous within its sub-blocks, and the motion vector differences between the macroblock and its sub-blocks are small, the macroblock is regarded as a unit. The $16\times16$ DCT is chosen by the proposed selection strategy. During the encoding process, the macroblock can be well predicted by the $16\times16$ macroblock motion estimation. The prediction error will be very small. The $16\times16$ DCT thus can efficiently compact the energy, which will be chosen by the M-RDO process. Otherwise, the $8\times8$ DCT will be determined by both the selection strategy and M-RDO.

In order to demonstrate the relationship between the selection strategy and the M-RDO process, the Hit Ratio (HR) curve is employed to demonstrate the hit rates. The transform type ($8\times8$ or $16\times16$ DCT) is first determined by the proposed selection strategy for each macroblock. Then the video sequences are encoded by the proposed perceptual coding scheme. The QP is fixed as 20 and the test conditions are listed in Table 9. During the encoding process, the transform type ($8\times8$ or $16\times16$ DCT) for each macroblock as determined by the M-RDO process is also recorded. The hit rate $h$ of each video frame measures the percentage of the macroblocks whose transform types determined by the M-RDO process and the proposed selection strategy are identical. It indicates that the selection strategy and M-RDO choose the same size transform. The HR curves of several typical CIF ($352\times288$) sequences are illustrated in Fig. 7. The hit rates are high, with the average hit rate higher than 80%. It means that the proposed selection strategy correlates well with the M-RDO process. During the video encoding, the M-RDO process will take the role to determine the transform size to be used. For other applications, such as visual quality assessment, watermarking, and so on, where the M-RDO process is not applicable, the proposed selection strategy will determine the transform size to be utilized.

The test 720P sequences, Crew, Harbor, Sailormen, and Spincalendar, are coded with fixed QP parameters. The H.264/AVC software platform used and compared is the JM 11 with ABT implementation [27]. The test conditions are listed in Table 9. With different QP parameters, nearly the same bit-rates are ensured by the traditional ABT codec and the proposed ABT-based JND codec, as shown in Table 10. It can be observed that there is a slight PSNR loss. As explained before, the PSNR correlates poorly with

**Table 9**
Test conditions.

| | |
|---|---|
| Platform | JM 11 (H.264) [27] |
| Sequence structure | IBBPBBP |
| Intra period | 10 frames |
| Transform Size | $8\times8$, and $16\times16$ |
| Entropy coding | CABAC |
| Deblocking filter | On |
| R-D optimization | On |
| Rate control | Off |
| Reference frame | 2 |
| Search Range | $\pm32$ |
| Frame rate | 30 frames/s |
| Total frame number | 199 |

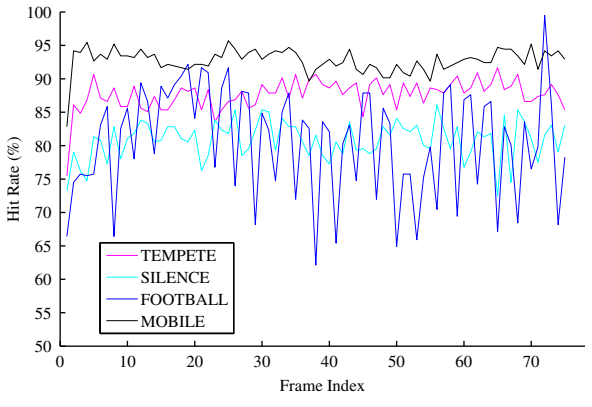


**Fig. 7.** HR curves for each CIF sequence.

**Table 10**
Performance comparison between the tradition ABT codec and the proposed ABT-based JND codec.

| Video | ABT Codec [27] | | | | The proposed codec | | | | ΔPSNR | $\Delta V_Q$ | ΔDMOS |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Bit-rates (Kbit/s) | PSNR (dB) | $V_Q$ | DMOS | Bit-rates (Kbit/s) | PSNR (dB) | $V_Q$ | DMOS | | | |
| Crew | 807.79 | 36.68 | 2.88 | 25.0 | 806.28 | 36.42 | 2.76 | 22.3 | −0.26 | −0.12 | −2.7 |
| Harbor | 1068.34 | 30.05 | 13.32 | 37.3 | 1056.37 | 29.83 | 13.22 | 32.5 | −0.22 | −0.10 | −4.8 |
| Sailormen | 572.40 | 30.92 | 10.09 | 33.8 | 576.51 | 30.86 | 9.78 | 30.5 | −0.06 | −0.31 | −3.3 |
| Spincalendar | 683.91 | 31.23 | 8.40 | 30.5 | 688.70 | 31.05 | 8.23 | 25.3 | −0.18 | −0.17 | −5.2 |

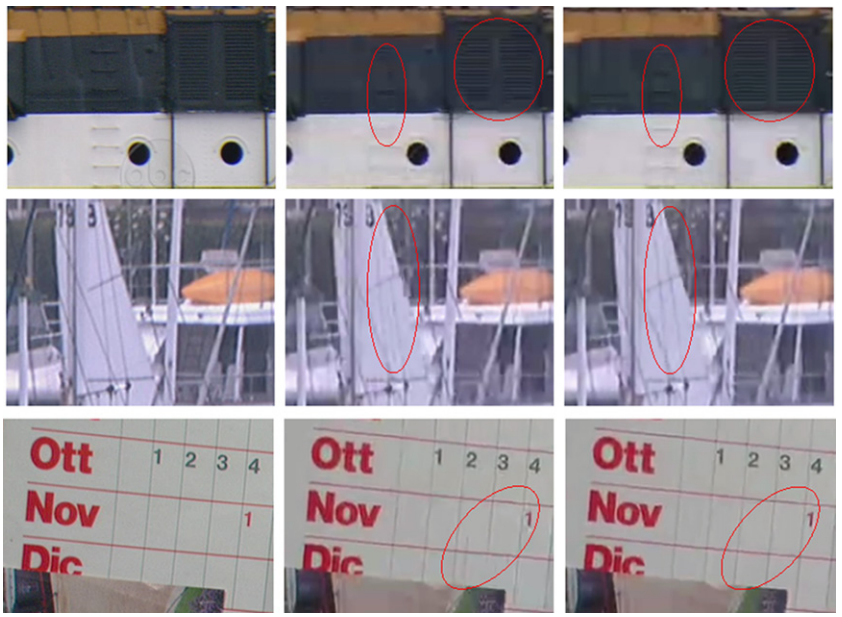


**Fig. 8.** Visual quality comparison of regions of the reconstructed frames generated by different video codec. Left: original frame; middle: reconstructed frame from ABT codec [27]; right: reconstructed frame for the proposed ABT-based JND codec; top: 113th frame of Sailormen; center: 109th frame of Harbor; bottom: 40th frame of Spincalendar.

the HVS perception, which makes it an improper criterion for visual quality assessment. The proposed visual quality index $V_Q$ has demonstrated better performances in matching subjective ratings. We calculate the $V_Q$ indexes of the distorted sequences. According to its definition in (12), the smaller the index, the better is the visual quality. It can be observed from Table 10 that the sequences generated by our proposed method possess smaller $V_Q$ indexes, compared to the sequences processed by [27].

In order to demonstrate the perceptual gain of our proposed video codec, the same DSCQS subjective test in Section 4.2 is conducted to evaluate the visual qualities of the coded video sequences. And the DMOS value for each coded sequence is listed in Table 10. As explained before, the smaller the DMOS value, the better the visual quality. Therefore, it can be observed that the proposed method can improve the visual quality of the coded video sequences with the constraint of the same bit-rates. Fig. 8 shows some pictures of videos coded and decoded with the JM 11 with ABT implementation [27] on one hand and with the proposed method on the other hand. Generally, the proposed method generates frames with higher visual quality, especially the detailed information, such as the lines and edges of Harbor and Spincalendar sequences.

## 5. Conclusions

In this paper, a novel ABT-based JND profile for images/videos is proposed by exploiting the HVS properties over transforms of different block sizes. Novel spatial and temporal selection strategies are designed to determine which block-size transform is employed, for still images/video INTRA frames and INTER video frames, respectively. The experimental results have demonstrated that the ABT-based JND profile can effectively model the HVS properties. Based on the proposed JND model, a simple visual quality metric is formulated. By evaluating on different image/video subjective quality databases, our visual quality metric performs comparably with the state-of-the-art metrics. The ABT-based JND profile is further applied to video coding, resulting in higher visual quality videos with the same bit-rates. It further confirms the efficiency of our proposed ABT-based JND in modeling the HVS characteristics.

## Acknowledgment

This work was partially supported by a grant from the Chinese University of Hong Kong under the Focused

Investment Scheme (Project 1903003). Thanks to the area editor and all the anonymous reviewers for their constructive comments and useful suggestions that led to the improvements in the quality, presentation and organization of the paper. The authors are grateful to Dr. Zhenyu Wei and Prof. Weisi Lin for providing their JND codes for comparisons, and thank Dr. Jie Dong for her valuable discussions and suggestions.

## References

[1] Weber's Law of Just Noticeable Differences, 〈http://www.usd.edu/psyc301/WebersLaw.htm〉.
[2] A.J. Ahumada, H.A. Peterson, Luminance-model-based DCT quantization for color image compression, Proceedings of the SPIE, Human Vision, Visual Processing, and Digital Display III 1666 (1992) 365–374.
[3] A.B. Watson, DCTune: a technique for visual optimization of DCT quantization matrices for individual images, Society for Information Display (SID) Digest 24 (1993) 946–949.
[4] I. Hontsch, L.J. Karam, Adaptive image coding with perceptual distortion control, IEEE Transactions on Image Processing 11 (2002) 213–222.
[5] W. Lin, L. Dong, P. Xue, Visual distortion gauge based on discrimination of noticeable contrast changes, IEEE Transactions on Circuits and Systems for Video Technology. 15 (2005) 900–909.
[6] Z. Lu, W. Lin, X. Yang, E. Ong, S. Yao, Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation, IEEE Transactions on Image Processing 14 (2005) 1928–1942.
[7] A.B. Watson, J. Hu, J.F. McGowan, Digital video quality metric based on human vision, Journal of Electronic Imaging 10 (2001) 20–29.
[8] R.B. Wolfgang, C.I. Podilchuk, E.J. Delp, Perceptual watermarks for digital images and video, Proceedings of IEEE 87 (1999) 1108–1126.
[9] C. Chou, C. Chen, A perceptual optimized 3-D subband codec for video communication over wireless channels, IEEE Transactions on Circuits and Systems for Video Technology 6 (1996) 143–156.
[10] Y. Chin, T. Berger, A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancements, IEEE Transactions on Circuits and Systems for Video Technology 9 (1999) 438–450.
[11] X. Yang, W. Lin, Z. Lu, E. Ong, S. Yao, Motion-compensated residue pre-processing in video coding based on just-noticeable-distortion profile, IEEE Transactions on Circuits and Systems for Video Technology 15 (2005) 742–750.
[12] X. Yang, W. Ling, Z. Lu, E. Ong, S. Yao, Just noticeable distortion model and its applications in video coding, Signal Processing: Image Communication 20 (2005) 662–680.
[13] X. Zhang, W. Lin, P. Xue, Improved estimation for just-noticeable visual distortion, Signal Processing 85 (2005) 795–808.
[14] Z. Wei, K.N. Ngan, A temporal just-noticeable distortion profile for video in DCT domain, in: Proceedings of the International Conference on Image Processing, 2008, pp. 1336–1339.
[15] Z. Wei, K.N. Ngan, Spatial–temporal just noticeable distortion profile for grey scale image/video in DCT domain, IEEE Transactions on Circuits and Systems for Video Technology 19 (2009) 337–346.
[16] Y. Huh, K. Panusopone, K.R. Rao, Variable block size coding of images with hybrid quantization, IEEE Transactions on Circuits and Systems for Video Technology 6 (1996) 679–685.
[17] X. Zhang, W. Lin, P. Xue, Just-noticeable difference estimation with pixels in images, Journal of Visual Communication and Image Representation 19 (2008) 30–41.
[18] S.J.P. Westen, R.L. Lagendijk, J. Biemond, A quality measure for compressed image sequences based on an eye movement compensated spatio-temporal model, in: Proceedings of the International Conference on Image Processing, 1997, pp. 279–282.
[19] J. Dong, J. Lou, C. Zhang, L. Yu, A. New, Approach to compatible adaptive block-size transforms, Proceedings of VCIP (2005).
[20] H. Qi, W. Gao, S. Ma, D. Zhao, Adaptive block-size transform based on extended integer $8\times 8/4\times 4$ transforms for H.264/AVC, in: Proceedings of the International Conference on Image Processing, 2006, pp. 1341–1344.
[21] K.N. Ngan, K.S. Leong, H. Singh, Adaptive cosine transform coding of image in perceptual domain, IEEE Transactions on Acoustics, Speech, and Signal Processing 37 (1989) 1743–1750.
[22] D.H. Kelly, Motion and vision II. Stabilized spatio-temporal threshold surface, Journal of the Optical Society America 69 (1979) 1340–1349.
[23] Y. Jia, W. Lin, A.A. Kassim, Estimating just-noticeable distortion for video, IEEE Transactions on Circuits and Systems for Video Technology 16 (2006) 820–829.
[24] G. Robson, Spatial and temporal contrast sensitivity functions of the visual system, Journal of the Optical Society America 56 (1966) 1141–1142.
[25] Y. Wang, J. Ostermann, Y. Zhang, Video Processing and Communications, Prentice Hall, 2002.
[26] S. Daly, Engineering observations from sptaiovelocity and spatiotemporal visual models, Proceedings of the SPEI 3299 (1998) 180–191.
[27] J. Dong, K.N. Ngan, C. Fong, W.K. Cham, 2D order-16 integer transforms for HD video coding, IEEE Transaction on Circuit System and Video Technology 19 (2009) 1463–1474.
[28] N. Nill, A visual model weighter cosine transform for image compression and quality assessment, IEEE Transactions on Communications 33 (1985) 551–557.
[29] N. jayant, J. johnsto, R. Sagranek, Signal compression based on models of human perception, Proceedings of the IEEE (1993) 1385–1422.
[30] Methodology for the Subjective Assessment of the Quality of Television Pictures, ITU-R BT.500.11, 2002.
[31] B. Li, M.R. Peterson, R.D. Freeman, Oblique effect: a neural basis in the visual cortex, Journal of Neurophysiology (2003) 204–217.
[32] R.J. Safranek, J.D. Johnston, A perceptually tuned subband image coder with image dependent quantization and post-quantization data compression, Proceedings of the IEEE ICASSP (1989) 1945–1948.
[33] L. Ma, K.N. Ngan, Adaptive block-size transform based just-noticeable difference profile for images, Proceedings of the PCM (2009).
[34] C. Zhang, L Yu, J. Lou, W. Cham, J. Dong, The technique of prescaled integer transform: concept, design and applications, IEEE Transaction on Circuit System and Video Technology 18 (2008) 84–97.
[35] S. Gordon, ABT for Film Grain Reproduction in High Definition Sequences, Doc. JVT-H029, Geneva, Switzerland, May 2003.
[36] T. Wedi, Y. Kashiwagi, T. Takahashi, H.264/AVC for next generation optical disc: a proposal on FREXt profile, Doc. JVT-K025. Munich, Germany, March 2004.
[37] [Available] 〈http://www.ee.cuhk.edu.hk/~lma/welcome_files/SPIC_2011_Experiments.html〉.
[38] B. Girod, What's wrong with mean-squared error, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993.
[39] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.
[40] H.R. Sheikh.and, A.C. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2006) 430–444.
[41] M.H. Pinson, S. Wolf, A new standardized method for objectively measuring video quality, IEEE Transaction on Broadcasting 50 (2004) 312–322.
[42] D.M. Chandler, S.S. Hemami, VSNR: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Transaction on Image Processing 16 (2007) 2284–2298.
[43] K. Seshadrinathan, A.C. Bovik, Motion tuned spatio-temporal quality assessment of natural videos, IEEE Transaction on Image Processing 19 (2010) 335–350.
[44] [Available] LIVE image quality assessment database: 〈http://live.ece.utexas.edu/research/quality/〉.
[45] [Available] IRRCyN/IVC database: 〈http://www2.irccyn.ec-nantes.fr/ivcdb/〉.
[46] [Available] A57 database: 〈http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html〉.
[47] [Available] LIVE Video Quality Assessment Database: 〈http://live.ece.utexas.edu/research/quality/〉.
[48] H.R. Sheikh, M.F. Sabir, A.C. Bovik, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Transaction on Image Processing 15 (2006) 3441–3452.
[49] VQEG. (2000) Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment. [online] Available: 〈http://www.vqeg.org〉.
[50] K. Seshadrinathan, R. Soundararajan, A.C. Bovik, L.K. Cormack, Study of subjective and objective quality assessment of video, IEEE Transaction on Image Processing 19 (2010) 1427–1441.
[51] A.C. Bovik, The Essential Guide to Video Processing, second edition, Elsevier, 2009.
[52] T. Wiegand, B. Girod, Lagrange Multiplier Selection in Hybrid Video Coder Control, ICIP, 2001.
[53] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, G.J. Sullivan, Rate-constrained coder control and comparison of video coding standards, IEEE Transaction on Circuit System and Video Technology 13 (2003) 688–703.
[54] Kodak Lossless True Color Image Suite. [online] Available: 〈http://r0k.us/graphics/kodak/〉.