Chapter 2
Computational Models for Just-Noticeable Differences

The growing demand for transmission and storage of images has spurred much effort in improving image compression techniques. To achieve this goal, one promising approach is to integrate properties of the HVS into image compression techniques [JJS93]. The central idea of such approach is to embed coding distortion beneath the spatial visibility threshold of the HVS. This threshold is commonly referred as the JND threshold [JJS93] as it specifies the minimum sensory difference that is detectable by the HVS. In the context of image compression, a perceptually perfect image is obtained at the lowest possible bit-rate [JJS93] if the coding error of each pixel in a compressed image is exactly at level of JND. Over the years, several computational models for JND have been developed and employed in image compression. These computational models for JND models are computed using subbands [SJ89, Wat93, TS96, HK00, HK02, ZLX05, ZLX08, WN09] and pixels [CL95, CC96, CB99, YLL03, YLL05, LLP10] of an image.

The first few models of the HVS [Sch56, MS74, Fau79] were developed using a single channel approach. Such models regard the HVS as a single spatial filter, which is defined by the CSF. One of the first few HVS based image quality metrics for luminance images was developed by Mannos and Sakrison [MS74]. By inferring some properties of the human vision from psychophysical experiments, Mannos and Sakrison derived a closed-form expression describing the contrast sensitivity of the HVS as a function of spatial frequency.

It is later argued that the HVS is a multi-channel system with each channel tuned to different ranges of spatial frequencies and orientations [Dau80], and many multi-channel models were subsequently proposed. Multi-channel HVS models are employed in metrics such as visual differences predictor (VDP) proposed by Daly [Dal92, Dal93], and the visual discrimination model (VDM) proposed by Lubin [Lub93, Lub95]. These image quality metrics are intended for general applicability, but are computationally expensive to implement.

A priori knowledge of the image processing algorithm (such as image compression) permits the use of specialized vision models. Although specialized vision models are not as versatile as the generalized models, specialized models can perform very well in a given application scope. Such vision models are usually simpler and computationally efficient. One example of an image coder based on a

E.-L. Tan and W.-S. Gan, *Perceptual Image Coding with Discrete Cosine Transform*, SpringerBriefs in Signal Processing,
DOI 10.1007/978-981-287-543-3_2

specialized vision model is the DCTune [Wat93], which permits higher image compression by exploiting the limitations of the HVS.

The general block diagram of a computational model for JND is shown in Fig. 2.1. Before the computational model for JND is applied, pre-processing such as color space transformation and frequency decomposition might be performed on the input image. In general, most JND models incorporate four properties of the HVS, namely, spatial contrast sensitivity, luminance adaptation, contrast masking, and temporal masking [Bov05]. The last three properties of the HVS are considered as elevation parameters of the base threshold which are determined by the spatial contrast sensitivity.

Since no masking is present in the measurement of contrast sensitivity, the effect of the background luminance on contrast sensitivity is typically accounted as luminance adaptation, or luminance masking [Wat93]. Contrast masking refers to the change of visibility of one image component due to the presence of another. The strongest contrast masking occurs when both components are of the same or at similar spatial frequency, orientation, and location. Temporal masking refers to the reduced contrast sensitivity due to the temporal variation of light intensity falling into the eye, and is commonly adopted in video compression. Since this monograph focuses on image processing, only the first three properties of the HVS shall be introduced in the following sections of this chapter.

The input image is decomposed into several components (also known as channels or subbands) in multi-channel HVS models. Numerous decomposition methods are used in PICs and image quality metrics, which include Fourier decomposition [CR68, MS74], Gabor decomposition [Dau88, LB90], DCT [Wat93, HK02, YLL03, YLL05, ZLX05, ZLX08], wavelet transform [TH94a, WHM97, LK00], and polar separable wavelet transform [Wat87, TH94b]. To combine the error of each spatial frequency, orientation band, and location into a single number or a distortion map [Wat79, RG81], many image quality metrics and PICs implement error pooling after CSF, luminance adaptation, and contrast masking.

This chapter begins with a review of the concepts on psychophysics of the human vision that are applied to image quality metrics and computational models for JND. In particular, this chapter emphasizes on image quality metrics and

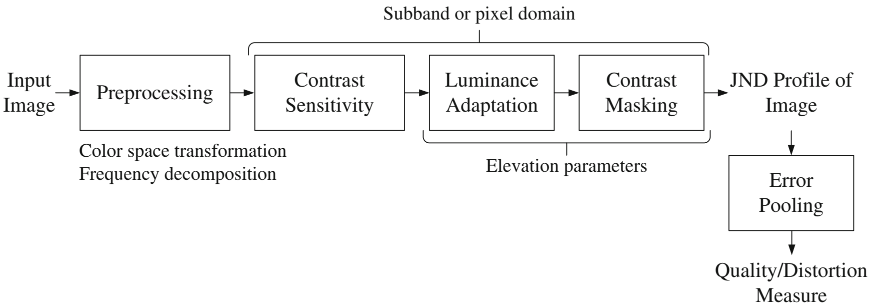

Fig. 2.1 Block diagram of a computational model for JND

computational models for JND that are designed specifically for image compression. As DCT is widely used in image and video compression standards (e.g. JPEG, MPEG-1/2/4, H.261/3), we focus our discussion in this chapter on image quality metrics computed using DCT subbands. It is also useful to consider pixel-based image quality metrics since it is possible to convert the contrast sensitivity from the pixel domain to the DCT subband domain and vice versa.

The cortex filters, which provide a good approximation of the multi-channel response of HVS, and the frequency decomposition using cortex filters are introduced in Sect. 2.1. This is followed by mapping of the cortex filters to DCT-II subbands (or coefficients). In Sect. 2.2, the widely adopted spatial CSF proposed by Ahumada and Peterson [AP92] is discussed. The detection threshold of every DCT subband is inversely proportional to contrast sensitivity, and can be derived from the spatial CSF. Apart from the contrast sensitivity, the detection threshold is varied by the local mean luminance and local spatial content, which are referred as luminance adaptation and contrast masking, respectively. Section 2.3 illustrates the effects of luminance adaptation using Weber's law. Subsequently, several techniques estimating luminance adaptation from the subbands and pixels of an image are reviewed. Next, intra- and inter-band contrast masking are discussed in Sect. 2.4. Intra-band contrast masking is typically adopted in many PICs due to its simple formulation. Discussions on estimating inter-band masking using cortex filters and block classification are also included. The final step of many image quality metrics, known as error pooling, is presented in Sect. 2.5.

2.1 Frequency Decomposition

The multi-channel response of HVS approximates a dyadic system [Dau80] that is well-matched by a multi-resolution filterbank or a wavelet decomposition. Examples of multi-resolution filterbank are cortex transform [Wat87] and cortex filter [Dal92, Dal93]. The cortex transform was first conceived by Watson [Wat87], which was inspired by neurophysiology [HW62, DAT82] and psychophysical studies in masking [BC69, SJ72]. The cortex transform is then adapted by Daly as the cortex filters in VDP. The decomposition of the frequency plane adopted by Watson and Daly is shown in Fig. 2.2. The main difference between Watson's and Daly's implementations of the cortex filtering is that Daly used six orientation bands [PDT77, DYH82], instead of four (in the case of Watson's cortex transform), to better approximate the orientation selectivity of the HVS. Several HVS models [WR84, Wat87, Dal92, Dal93] use six spatial channels, and it was found that six spatial channels show good agreement with psychophysical data [WLM90].

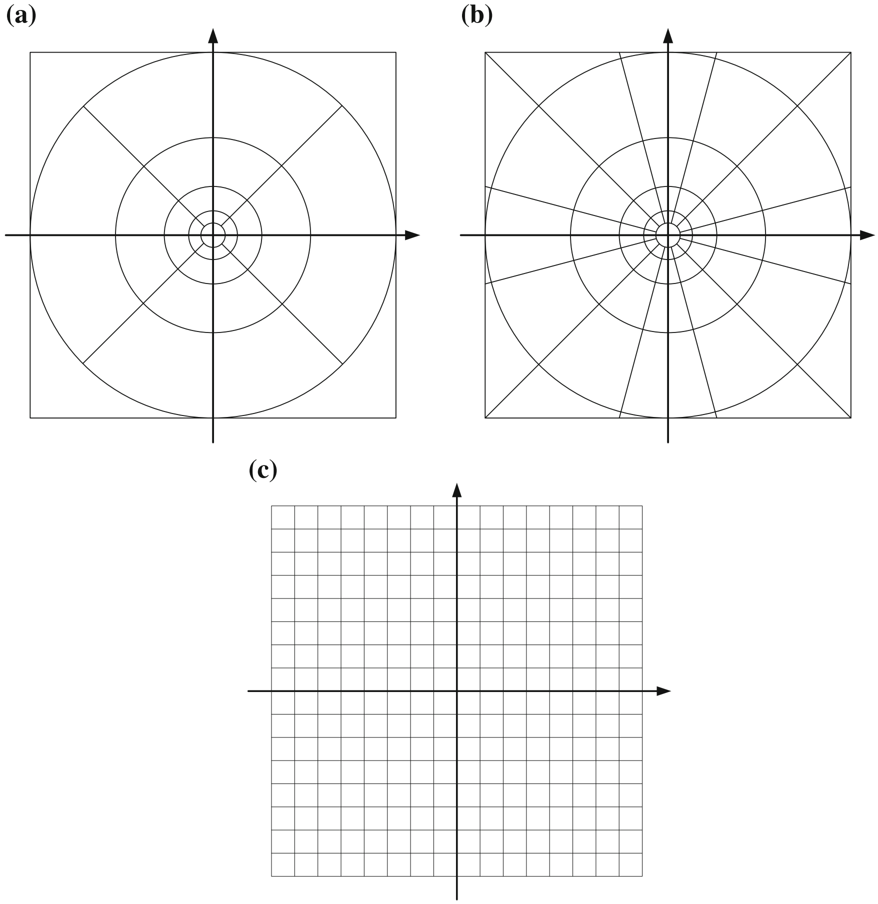

Fig. 2.2 Decomposition of the frequency plane corresponding to **a** Watson's cortex transform [Wat87]. **b** Daly's cortex filters [Dal92, Dal93], and **c** DCT-II. Range of each axis is from $-f_s/2$ to $f_s/2$, where f_s is the sampling frequency

2.1.1 Cortex Filters

The cortex filters model the spatial (or radial) frequency selectivity and the orientational selectivity of the HVS. These filters are formed by cascading two filters, which model the radial frequency bands and orientation bands of the HVS. The radial frequency filters are formed by the difference of two dimensional (2-D) lowpass mesa filters. The mesa filter possesses a flat passband, a Hanning window transition band, and a flat stopband as shown in Fig. 2.3.

The mesa filter [Dal92] is completely characterized by its half-amplitude frequency $d_{1/2}$ and transition width tw. Let s denote the spatial frequency in cycles per degree (cpd). The mesa filter mesa(s) is expressed as

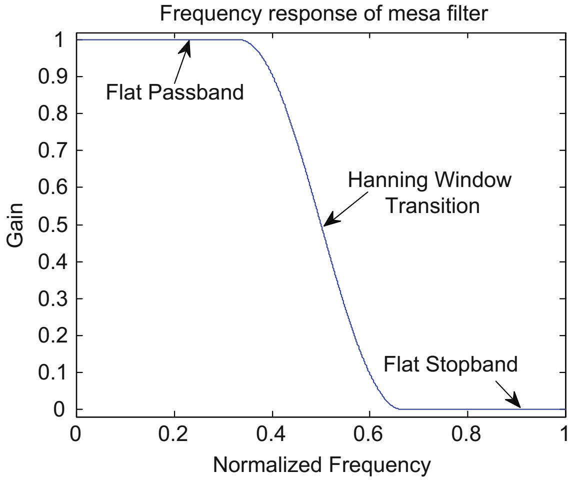

Fig. 2.3 Frequency response of a mesa filter [Dal92, Dal93]

$$\mathrm{mesa}(s) = \begin{cases} 1, & \text{for } s < s_{1/2} - \frac{tw}{2}, \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi\left(s - s_{1/2} - tw/2\right)}{tw}\right)\right), & \text{for } s_{1/2} - \frac{tw}{2} \le s \le d_{1/2} + \frac{tw}{2}, \\ 0, & \text{for } s > s_{1/2} + \frac{tw}{2}, \end{cases} \tag{2.1}$$

where $tw = 2s_{1/2}/3$. The radial frequency selectivity of the HVS is modelled by the difference of two mesa filters with different half amplitude frequencies. The difference of the mesa (DOM) filter $\mathrm{dom}(d, s)$ is given by

$$\mathrm{dom}(d, s) = \mathrm{mesa}(s)|_{s_{1/2}=2^{-(d-1)}} - \mathrm{mesa}(s)|_{s_{1/2}=2^{-d}}, \tag{2.2}$$

where $d = 0, 1, \ldots, D - 1$, and D is the number of DOM filters. The choice of tw yields a set of cortex bands with approximately constant behaviour on a log frequency axis with a bandwidth of one octave [SJ72, MTT78, DAT82]. The orientation sensitivity of the HVS can be modelled by a set of fan filters [Dal92], which is expressed as

$$\mathrm{fan}(f, \theta) = \begin{cases} \frac{1}{2}\left(1 - \cos\left(\frac{\pi|\theta - \theta_{cr}(f)|}{\theta_{tw}}\right)\right), & |\theta - \theta_c(f)| \le \theta_{tw}, \\ 0, & \text{otherwise}, \end{cases} \tag{2.3}$$

where θ_{tw} is the angular transition width in degree; $\theta_{cr}(f)$ is the orientation of the center angular frequency of the fth fan filter in degree, $f = 0, 1, \ldots, F - 1$, and F is the number of fan filters. $\theta_{cr}(f)$ is given by

$$\theta_{cr}(f) = (f - 1)\theta_{tw} - 90^\circ, \tag{2.4}$$

where $\theta_{tw} = 180°/F$. The cortex filter at the **b**th band cortex$(\mathbf{b}, s, \theta)$ is formed by the product of the *d*th DOM and *f*th fan filters, which is given as

$$\mathrm{cortex}(\mathbf{b}, s, \theta) = \begin{cases} \mathrm{dom}(d, s)\mathrm{fan}(f, \theta), & \text{for } d = 1, \ldots, D-1;\ f = 0, 1, \ldots F-1, \\ \mathrm{base}(s), & \text{for } d = D, \end{cases} \tag{2.5}$$

where $\mathbf{b} = (d, f)$, and base(*s*) is the cortex filter having the lowest spatial frequency without orientational selectivity. In [TS96], the base(*s*) filter is implemented using a truncated Gaussian function, which is given as

$$\mathrm{base}(s) = \begin{cases} e^{-\frac{s^2}{2\sigma^2}}, & \text{for } s < s_{1/2} + \frac{tw}{2}, \\ 0, & \text{for } s \geq s_{1/2} + \frac{tw}{2}, \end{cases} \tag{2.6}$$

where $\sigma = (2^{-D} + tw/2)/3$. Six spatial channels ($D$ = 6) and six orientation bands (F = 6) are used in Daly's implementation of the cortex filters.

Since the cortex filters model the spatial frequency selectivity and orientation selectivity of the HVS, it would be useful to consider the mapping of DCT-II coefficients to the cortex bands. The general idea behind mapping of DCT-II coefficients to the cortex bands is to group the DCT-II coefficients that belong to the same cortex bands [TS96].

An example of this mapping is illustrated in Fig. 2.4. In order to map the partially covered DCT-II coefficients that fall within a cortex band, Tran and Safranek divide each DCT-II coefficients into $M \times M$ smaller blocks (referred as sub-bins). Subsequently, these sub-bins are grouped into corresponding **b**th band of the cortex filters. Let **k** denote the **k**th DCT-II coefficient of an $N \times N$ DCT-II block, where $\mathbf{k} = (k_1, k_2)$ and $k_1, k_2 = 0, 1, \ldots, N-1$. The overlapping area between the **k**th DCT coefficient and the corresponding band cortex band is computed as

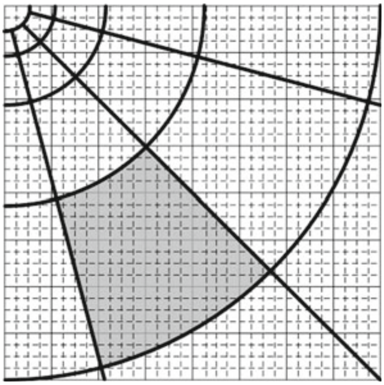

Fig. 2.4 Mapping of DCT coefficients (*thin line*) to cortex bands (*thick line*) [TS96]. The *shaded area* denotes the DCT coefficients which fall within the same cortex band. *Dashed lines* denote the sub-bins of each DCT coefficient

$$\mathrm{overlap}(\mathbf{k},\mathbf{b}) = \sum_{m_1=k_1M}^{(k_1+1)M}\sum_{m_2=k_2M}^{(k_2+1)M}\mathrm{cortex}(\mathbf{b},m_1,m_2), \tag{2.7}$$

which leads to T_{CF} $N \times N$ matrices, where T_{CF} is the number of cortex filters. Each T_{CF} matrix contains the information of the overlapping area of the N^2 DCT-II coefficients.

2.2 Spatial Contrast Sensitivity Function

In this section, we shall review the widely adopted spatial CSF [Wat93, HK02, YLL03, YLL05, ZLX05, ZLX08], which was proposed by Ahumada and Peterson [AP92]. Their formulation of the CSF is very useful as it takes into account of display luminance levels, veiling luminance levels, and spatial frequencies.

We consider the base detection threshold $T_{\mathrm{D}}(\mathbf{k},\mathbf{n})$ of the $\mathbf{k}$th DCT-II subband located at $\mathbf{n}$ of an image, where $\mathbf{n}=(n_1,n_2)$; $n_1=0,1,\ldots,H/N-1$; $n_2=0,1,\ldots,W/N-1$; H denotes the height of an image and W denotes the width of an image. Let $f(\mathbf{k})$ and $\theta(\mathbf{k})$ denote the spatial frequency of a grating and the angle between two gratings, respectively. Based on van Nes and Bouman's measurements [NB67], Ahumada and Peterson approximated the detection threshold using a parabola in log spatial frequency, and they expressed the detection threshold $T_D(\mathbf{k},\mathbf{n})$ as

$$\log_{10}(T_D(\mathbf{k},\mathbf{n})) = \log_{10}\left(\frac{T_{\min}(\mathbf{n})}{0.7+0.3\cos^2\theta(\mathbf{k})}\right) + K(\mathbf{n})(\log_{10}f(\mathbf{k}) - \log_{10}f_{\min}(\mathbf{n}))^2,$$
$$k_1 = 0 \text{ or } k_2 = 0, \tag{2.8}$$

where $\theta(\mathbf{k}) = \sin^{-1}\left(2f(k_1,0)f(0,k_2)/f^2(\mathbf{k})\right)$, $f(\mathbf{k}) = \sqrt{(k_1/w_x)^2+(k_2/w_y)^2}$, w_x and w_y are the horizontal width and vertical height of a pixel, respectively. $T_{\min}(\mathbf{n})$, $K(\mathbf{n})$, and $f_{\min}(\mathbf{n})$ are the functions of the total luminance $L(\mathbf{n})$, where $L(\mathbf{n})$ is the sum of veiling luminance and the luminance of the image located at $\mathbf{n}$. Based on Ahumada and Peterson's formulation, $T_{\min}(\mathbf{n})$, $K(\mathbf{n})$, and $f_{\min}(\mathbf{n})$ are computed as

$$T_{\min}(\mathbf{n}) = \begin{cases} 0.0263L(\mathbf{n})^{0.649}, & L(\mathbf{n}) \leq 13.45\ \mathrm{cd/m^2}, \\ 0.0106L(\mathbf{n}), & \text{otherwise}, \end{cases} \tag{2.9}$$

$$f_{\min}(\mathbf{n}) = \begin{cases} 2.401L(\mathbf{n})^{0.182}, & L(\mathbf{n}) \leq 300\ \mathrm{cd/m^2}, \\ 6.78, & \text{otherwise}, \end{cases} \tag{2.10}$$

and

$$K(\mathbf{n}) = \begin{cases} 2.0891L(\mathbf{n})^{0.0706}, & L(\mathbf{n}) \leq 300 \text{ cd/m}^2, \\ 3.125, & \text{otherwise.} \end{cases} \tag{2.11}$$

Since van Nes and Bouman found negligible difference between the CSFs for luminance ranging from 290 to 1880 cd/m^2, (2.10) and (2.11) are clipped at 300 cd/m^2. It should be noted that Kelly [Kel85] stated that this parabola model of the CSF may not be valid for low spatial frequencies, and Peterson et al. [PMP91] suggested a conservative estimate for $T_D(0,0,\mathbf{n})$, which is the smaller value of $T_D(1,0,\mathbf{n})$ and $T_D(0,1,\mathbf{n})$.

Watson [Wat93] used the DC DCT-II coefficient to estimate the local luminance of an image. Höntsch and Karam [HK02] estimated the local luminance from the foveal region, which typically covers two degrees of the visual angle, as

$$L(\mathbf{n}) = L_{\min} + \frac{L_{\max} - L_{\min}}{M}\left(\sum_{(0,0,m_1,m_2)\in F(0,0,\mathbf{n})} \frac{C(0,0,m_1,m_2)}{N_F N} + \bar{m}\right), \tag{2.12}$$

where $F(0,0,\mathbf{n})$ denotes the foveal region centers at $\mathbf{n}$ in DC subband; $C(0,0,m_1,m_2)$ denotes the DC DCT-II coefficient at (m_1,m_2); N_F denotes the number of DCT-II coefficients at $\mathbf{n}$ in DC subband that fall inside the foveal region; and $\bar{m}$ is the mean of the image; M is the number of gray levels in the image; $L_{\max}$ and $L_{\min}$ are the maximum and minimum luminance levels of the display, respectively. N_F is computed as

$$N_F = \left(\left\lfloor \frac{2VR_x}{N} \tan\left(\frac{\theta_f}{2}\right)\right\rfloor\right)\left(\left\lfloor \frac{2VR_y}{N} \tan\left(\frac{\theta_f}{2}\right)\right\rfloor\right), \tag{2.13}$$

where the operator $\lfloor . \rfloor$ returns the nearest smallest integer; V is the viewing distance in inches; R_x and R_y are the height and width of the display resolution in pixel per inch, respectively; and θ_f is the visual angle (approximately 2°) covered by the foveal region.

Assuming an image is displayed on a gamma corrected screen, we can linearly map signal intensity values into luminance levels. Thereby, the base detection threshold $T_{\mathrm{b}}(\mathbf{k},\mathbf{n})$ for the $\mathbf{k}$th DCT-II subband located at $\mathbf{n}$ is computed as

$$T_{\mathrm{b}}(\mathbf{k},\mathbf{n}) = \frac{MT_{\mathrm{D}}(\mathbf{k},\mathbf{n})}{\alpha_{k_1}\alpha_{k_2}(L_{\max} - L_{\min})}. \tag{2.14}$$

To ensure the quantization error remains invisible to the HVS, the quantization of each DCT-II coefficient should not be greater than $2T_{\mathrm{b}}(\mathbf{k},\mathbf{n})$.

The JND threshold for DCT-II subband is formulated as a product of the detection threshold $T_{\mathrm{b}}(\mathbf{k},\mathbf{n})$ and its elevation parameters given by luminance adaptation and contrast masking. Let $e_{\mathrm{la}}(\mathbf{n})$ and $e_{\mathrm{cm}}(\mathbf{k},\mathbf{n})$ denote the luminance adaption and contrast masking, respectively. Hence, the JND threshold $T(\mathbf{k}, \mathbf{n})$ for the $\mathbf{k}$th DCT-II subband located at $\mathbf{n}$ is given as

$$T(\mathbf{k}, \mathbf{n}) = T_{\mathrm{b}}(\mathbf{k}, \mathbf{n}) e_{\mathrm{la}}(\mathbf{n}) e_{\mathrm{cm}}(\mathbf{k}, \mathbf{n}). \tag{2.15}$$

Since the luminance of a digital image spans a small luminance range of the spatial CSF experiment conducted by van Nes and Bouman [NB67], a single spatial CSF (based on the mean luminance of the image) can be used for the whole image [ZLX05]. Therefore, the detection threshold $T_{\mathrm{D}}(\mathbf{k}, \mathbf{n})$ can be simplified to $T_{\mathrm{D}}(\mathbf{k})$ by replacing the total luminance $L(\mathbf{n})$ with the mean luminance L of the display [Wat93, ZLX05, ZLX08].

2.3 Luminance Adaptation

Weber's law is widely used to model luminance adaptation, and the Weber fraction $K = \Delta I / I_{bg}$ is found to be nearly constant for a wide range of intensities [Hec24], where I_{bg} is the background intensity and ΔI is the just-noticeable incremental intensity over the background. However, Weber's law does not hold for a wide range of background intensities and spatial frequencies. For an 8-bit grayscale image, it is found that the Weber's fraction stays fairly constant for gray levels from 50 to 235; and higher contrast sensitivity [SW96] is found for gray levels lower and higher than 50 and 235, respectively. These observations are similar to those reported in [SJ89, CL95]. From the empirical model of the CSF in [Bar04], it is also observed that the contrast sensitivity remains relatively constant at low spatial frequencies for luminance levels between 10 and 1000 cd/m^2. However, the contrast sensitivity for these luminance levels vary significantly as the spatial frequency increases.

In the DCT domain, Watson [Wat93] estimated the luminance adaptation for $\mathbf{n}$th DCT-II block using

$$e_{\mathrm{la}}^{\mathrm{Wat}}(\mathbf{n}) = \left(\frac{C(0, 0, \mathbf{n})}{\bar{C}_L} \right)^{0.649}, \tag{2.16}$$

where $\bar{C}_L$ refers to the DC DCT-II coefficient corresponding to the mean luminance ($\bar{C}_L$ = 1024 for a 8-bit image). On the other hand, Zhang et al. [ZLX05, ZLX08] considered different luminance adaptation at low and high luminance, and they estimated the luminance adaptation as

$$e_{\mathrm{la}}^{\mathrm{ZLX}}(\mathbf{n}) = \begin{cases} 2\left(1 - \frac{C(0,0,\mathbf{n})}{128N}\right)^3 + 1, & \text{for } C(0, 0, \mathbf{n}) \leq 128N, \\ 0.8\left(\frac{C(0,0,\mathbf{n})}{128N} - 1\right)^2 + 1, & \text{otherwise.} \end{cases} \tag{2.17}$$

Using a similar approach, Wei and Ngan [WN09] computed luminance adaptation using

$$e_{\mathrm{la}}^{\mathrm{WN}}(\mathbf{n}) = \begin{cases} \left(\frac{60N - C(0,0,\mathbf{n})}{150N}\right) + 1, & \text{for } C(0,0,\mathbf{n}) \leq 60N, \\ 1, & \text{for } 60N < C(0,0,\mathbf{n}) < 170N, \\ \left(\frac{C(0,0,\mathbf{n}) - 170N}{425N}\right) + 1, & \text{for } C(0,0,\mathbf{n}) \geq 170N. \end{cases} \tag{2.18}$$

In the pixel domain, Chou and Li [CL95, YLL05] empirically determined the luminance adaptation of a pixel at $\mathbf{x}$, where $\mathbf{x} = (x_1, x_2)$, $x_1 = 0, 1, \ldots, H-1$, and $x_2 = 0, 1, \ldots, W-1$, using the following:

$$e_{\mathrm{la}}^{\mathrm{CL}}(\mathbf{x}) = \begin{cases} 17\left(1 - \sqrt{\frac{L_s(\mathbf{x})}{127}}\right) + 3, & \text{for } L_s(\mathbf{x}) \leq 127, \\ \frac{3}{128}(L_s(\mathbf{x}) - 127) + 3, & \text{for } L_s(\mathbf{x}) > 127, \end{cases} \tag{2.19}$$

where $L_s(\mathbf{x})$ is the local luminance at $\mathbf{x}$, and (2.19) was obtained for a distance of six times of the image height. Chou and Li determined the local luminance $L_s(\mathbf{x})$ as

$$L_s(\mathbf{x}) = \frac{1}{32} \sum_{p_1=0}^{4} \sum_{p_2=0}^{4} i(x_1 - 2 + p_1, x_2 - 2 + p_2) B(p_1, p_2), \tag{2.20}$$

where $i(\mathbf{x})$ denotes the pixel of an image at $\mathbf{x}$ and the operator B is depicted in Fig. 2.5.

Fig. 2.5 Operator to determine average local luminance (B) [CL95]

1	1	1	1	1
1	2	2	2	1
1	2	0	2	1
1	2	2	2	1
1	1	1	1	1

B

2.4 Contrast Masking

Contrast masking refers to the reduction of visibility of one image signal due to the presence of another signal. The masking characteristic of the HVS is known to be strongest when both signals are of the same spatial frequency, orientation, and location [LF80]. Contrast masking can be classified as inter- and intra-band masking. Sometimes, the term "texture masking" (inter-band masking) is used to refer to a "broadband" masker, where the masking effect is contributed by multiple frequency and orientation channels. On the other hand, intra-band masking refers to the masking due to a masker within the same frequency and orientation channel. Based on the estimation of contrast masking reported in [SJ89], Höntsch and Karam [HK00] proposed a more elaborate adjustment for contrast masking, which incorporates both intra- and inter-band masking. Let $e_{\text{inter}}(\mathbf{k},\mathbf{n})$ and $e_{\text{intra}}(\mathbf{k},\mathbf{n})$ denote the amount of intra- and inter-band masking at $\mathbf{n}$ of $\mathbf{k}$th subband of an image, respectively. The elevation parameter $e_{\text{cm}}(\mathbf{k},\mathbf{n})$ is computed as

$$e_{\text{cm}}(\mathbf{k},\mathbf{n}) = e_{\text{inter}}(\mathbf{k},\mathbf{n})e_{\text{intra}}(\mathbf{k},\mathbf{n}). \tag{2.21}$$

In [SJ89], Safranek and Johnston proposed a subband image coder that employs a 4 × 4 band generalized QMF (GQMF) to decompose an image into 16 subbands. Let $tex^{\text{SJ}}(\mathbf{b},\mathbf{n}')$ denote the texture energy of the $\mathbf{b}$th subband at location $\mathbf{n}'$, and $\text{wCSF}(\mathbf{b})$ is the $\mathbf{b}$th weighting factor empirically derived from a CSF [Cor90], where $\mathbf{n}' = (n'_1, n'_2)$, $n'_1 = 0, 1, \ldots, H/2-1$, $n'_2 = 0, 1, \ldots, W/2-1$, $\mathbf{b} = (b_1, b_2)$, and $0 \le b_1, b_2 \le 3$. Safranek and Johnston defined contrast masking (only inter-band masking is considered) as follows:

$$e^{\text{SJ}}_{\text{inter}}(\mathbf{b},\mathbf{n}') = \max\left\{1, \left(\sum_{\mathbf{b}} \text{wCSF}(\mathbf{b})\, tex^{\text{SJ}}(\mathbf{b},\mathbf{n}')\right)^{0.15}\right\}. \tag{2.22}$$

The texture energy of the $\mathbf{b}$th subband at location $\mathbf{n}'$ is computed as

$$tex^{\text{SJ}}(\mathbf{b},\mathbf{n}') = \begin{cases} \text{var}(\mathbf{n}'), & \text{for } \mathbf{b} = (0,0), \\ \text{energy}(\mathbf{b},\mathbf{n}'), & \text{otherwise}, \end{cases} \tag{2.23}$$

where $\text{energy}(\mathbf{b},\mathbf{n}')$ computes the energy of the $\mathbf{b}$th subband at $\mathbf{n}'$ and $\text{var}(\mathbf{n}')$ computes the variance at $\mathbf{n}'$ of subband zero over the area given by (n'_1, n'_2), $(n'_1 + 1, n'_2)$, $(n'_1, n'_2 + 1)$, and $(n'_1 + 1, n'_2 + 1)$.

Based on the masking model in [LF80], Watson [Wat93] adjusted the base detection threshold to account for contrast masking (only intra-band masking is considered) using the following:

$$e_{\text{intra}}^{\text{Wat}}(\mathbf{k},\mathbf{n}) = \begin{cases} 1, & \mathbf{k} = (0,0), \\ \max\left\{1, \left(\frac{|C(\mathbf{k},\mathbf{n})|}{T_{\text{b}}(\mathbf{k},\mathbf{n})e_1^{\text{Wat}}(\mathbf{n})}\right)^{0.7}\right\}, & \mathbf{k} \neq (0,0). \end{cases} \tag{2.24}$$

It is assumed that there is no contrast masking in the DC DCT-II coefficient. However, the DC DCT-II coefficient indirectly affects contrast masking via $e_1^{\text{Wat}}(\mathbf{n})$ in the denominator of (2.24) for all DCT-II coefficients except for the DC DCT-II coefficient.

In [HK00], Höntsch and Karam estimated intra-band masking using

$$e_{\text{intra}}^{\text{HK}}(\mathbf{k},\mathbf{n}) = \begin{cases} 1, & \mathbf{k} = (0,0), \\ \max\left\{1, \left(\frac{|C(\mathbf{k},\mathbf{n})|}{T_{\text{b}}(\mathbf{k},\mathbf{n})}\right)^{0.36}\right\}, & \mathbf{k} \neq (0,0), \end{cases} \tag{2.25}$$

and inter-band masking is computed using (2.22) with an exponent of 0.035. Taking into account of the foveal region for intra-band masking, Höntsch and Karam [HK02] proposed the adjustment for intra-band masking as

$$e_{\text{intra}}^{\text{HK2}}(\mathbf{k},\mathbf{n}) = \begin{cases} 1, & \mathbf{k} = (0,0), \\ \max\left\{1, \left(\frac{|\bar{C}_F(\mathbf{k},\mathbf{n})|}{T_{\text{b}}(\mathbf{k},\mathbf{n})}\right)^{0.6}\right\}, & \mathbf{k} \neq (0,0), \end{cases} \tag{2.26}$$

where $\bar{C}_F(\mathbf{k},\mathbf{n})$ is the average magnitude of the DCT-II coefficients in the foveal region.

Yang et al. [YLL05] improved their estimate of contrast masking by differentiating the contribution of masking from edge and texture. Edges are structurally simpler than textures, and it is generally observed that edges tend to be easily recognized by the HVS. Furthermore, a typical observer would have prior knowledge of how an edge looks like [EB98]. Girod [Gir93] found that the HVS has acute sensitivity at or near the luminance edge. Based on these observations in [EB98, Gir93], Yang et al. defined the JND threshold at a texture region to be three times higher than those at an edge region.

To date, classification of plain, edge, and texture blocks are performed in [YLL05, ZLX05, ZLX08, WN09, LLP10] to effectively estimate contrast masking in an image. Zhang et al. [ZLX05, ZLX08] employed a block classification method in the DCT domain [TV98], which was first proposed in [PJJ94]. To perform block classification in the DCT domain, the DCT-II coefficients of an $N \times N$ sub-image are divided into four groups as shown in Fig. 2.6. Let $L_{\text{T}}(\mathbf{n})$, $M_{\text{T}}(\mathbf{n})$, and $H_{\text{T}}(\mathbf{n})$ denote the sum of DCT-II coefficients (absolute magnitude) in the low-frequency (LF), mid-frequency (MF), and high-frequency (HF) groups, respectively, of the $\mathbf{n}$th DCT-II block. Based on these sums, three measures are formulated to determine the texture energy of the $\mathbf{n}$th DCT-II block, and these measures are defined as

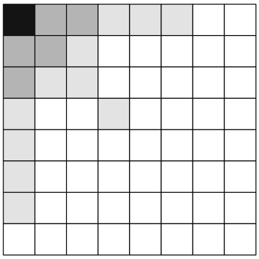

Fig. 2.6 DCT-II block classification for contrast masking. LF, MF, and HF are represented by the *dark gray*, *light gray* and *white boxes*, respectively [ZLX05, ZLX08]

$$\begin{aligned} tex_1^{\mathrm{ZLX}}(\mathbf{n}) &= M_{\mathrm{T}}(\mathbf{n}) + H_{\mathrm{T}}(\mathbf{n}), \\ tex_2^{\mathrm{ZLX}}(\mathbf{n}) &= \frac{(\bar{L}_{\mathrm{T}}(\mathbf{n}) + \bar{M}_{\mathrm{T}}(\mathbf{n}))}{\bar{H}_{\mathrm{T}}(\mathbf{n})}, \\ tex_3^{\mathrm{ZLX}}(\mathbf{n}) &= \frac{\bar{L}_{\mathrm{T}}(\mathbf{n})}{\bar{M}_{\mathrm{T}}(\mathbf{n})}, \end{aligned} \tag{2.27}$$

where $\bar{L}_{\mathrm{T}}(\mathbf{n})$, $\bar{M}_{\mathrm{T}}(\mathbf{n})$, and $\bar{H}_{\mathrm{T}}(\mathbf{n})$ are the means of $L_{\mathrm{T}}(\mathbf{n})$, $M_{\mathrm{T}}(\mathbf{n})$, and $H_{\mathrm{T}}(\mathbf{n})$, respectively.

Each DCT-II block is classified into PLAIN, EDGE, or TEXTURE class using $tex_1^{\mathrm{ZLX}}(\mathbf{n})$, $tex_2^{\mathrm{ZLX}}(\mathbf{n})$, and $tex_3^{\mathrm{ZLX}}(\mathbf{n})$ as shown in Table 2.1. DCT-II blocks that are generally smooth with few spatial activities are classified as PLAIN, DCT-II blocks containing a lot of complex spatial activities are classified as TEXTURE, and DCT-II blocks containing clear edges are classified as EDGE.

Based on the block classification result, inter-band contrast masking is computed by

Table 2.1 Conditions used in classification of DCT-II blocks [ZLX05, ZLX08]

Case	Conditions	Block classification
I	$tex_1^{\mathrm{ZLX}}(\mathbf{n}) \leq 125$	DCT-II block is classified as PLAIN
II	$125 < tex_1^{\mathrm{ZLX}}(\mathbf{n}) \leq 290$ and $\max\left(tex_2^{\mathrm{ZLX}}(\mathbf{n}), tex_3^{\mathrm{ZLX}}(\mathbf{n})\right) \geq 7$ $\min\left(tex_2^{\mathrm{ZLX}}(\mathbf{n}), tex_3^{\mathrm{ZLX}}(\mathbf{n})\right) \geq 5$ or $tex_2^{\mathrm{ZLX}}(\mathbf{n}) \geq 16$	DCT-II block is classified as EDGE, otherwise PLAIN
III	$290 < tex_1^{\mathrm{ZLX}}(\mathbf{n}) \leq 900$ and $\max\left(tex_2^{\mathrm{ZLX}}(\mathbf{n}), tex_3^{\mathrm{ZLX}}(\mathbf{n})\right) \geq 7$ $\min\left(tex_2^{\mathrm{ZLX}}(\mathbf{n}), tex_3^{\mathrm{ZLX}}(\mathbf{n})\right) \geq 5$ or $tex_2^{\mathrm{ZLX}}(\mathbf{n}) \geq 16$	DCT-II block is classified as EDGE, otherwise TEXTURE
IV	$tex_1^{\mathrm{ZLX}}(\mathbf{n}) > 900$ and $\max\left(tex_2^{\mathrm{ZLX}}(\mathbf{n}), tex_3^{\mathrm{ZLX}}(\mathbf{n})\right) \geq 0.7$ $\min\left(tex_2^{\mathrm{ZLX}}(\mathbf{n}), tex_3^{\mathrm{ZLX}}(\mathbf{n})\right) \geq 0.5$ or $tex_2^{\mathrm{ZLX}}(\mathbf{n}) \geq 16$	DCT-II block is classified as EDGE, otherwise TEXTURE

$$e_{\text{inter}}^{\text{ZLX}}(\mathbf{n}) = \begin{cases} 1 + \frac{tex_1^{\text{ZLX}}(\mathbf{n}) - 290}{1208}, & \text{for TEXTURE block,} \\ 1.25, & \text{for EDGE block and } L(\mathbf{n}) + M(\mathbf{n}) > 400, \\ 1.125, & \text{for EDGE block and } L(\mathbf{n}) + M(\mathbf{n}) \leq 400, \\ 1, & \text{for PLAIN block.} \end{cases} \tag{2.28}$$

Zhang et al. considered similar adjustment as (2.24) for intra-band contrast masking, and the amount of adjustment for contrast masking is computed as

$$e_{\text{intra}}^{\text{ZLX}}(\mathbf{k},\mathbf{n}) = \begin{cases} 1, & \text{for EDGE block} \\ & \text{at } \mathbf{k} \in LF \cup MF, \\ \max\left\{1, \left(\frac{|C(\mathbf{k},\mathbf{n})|}{T_b(\mathbf{k},\mathbf{n}) e_{\text{la}}^{\text{ZLX}}(\mathbf{k},\mathbf{n})}\right)^{0.36}\right\}, & \text{otherwise.} \end{cases} \tag{2.29}$$

To avoid over-estimation of JND threshold at the EDGE block, the LF and MF regions of the EDGE block are excluded from the estimation of intra-band contrast masking.

Differing from Zhang's method, Wei and Ngan [WN09] performed block classification in the pixel domain. Using an edge map of the image obtained with the Canny edge detector [Can86], Wei and Ngan computed the edge density $\bar{p}_{\text{edge}}(\mathbf{n})$ at $\mathbf{n}$ as the ratio of the number of edge pixels in each $N \times N$ sub-image to N^2. Based on the edge density, the $\mathbf{n}$th DCT-II block is classified as

$$\text{Block Type}(\mathbf{n}) = \begin{cases} \text{PLAIN} & \text{for } \bar{p}_{\text{edge}}(\mathbf{n}) \leq 0.1, \\ \text{EDGE} & \text{for } 0.1 < \bar{p}_{edge}(\mathbf{n}) \leq 0.2, \\ \text{TEXTURE} & \text{for } \bar{p}_{\text{edge}}(\mathbf{n}) > 0.2. \end{cases} \tag{2.30}$$

Using the block classification results from (2.30), the inter-band masking is computed as

$$e_{\text{inter}}^{\text{WN}}(\mathbf{k},\mathbf{n}) = \begin{cases} 1 & \text{for PLAIN and EDGE block,} \\ 2.25 & \text{for } \left(k_1^2 + k_2^2\right) \leq 16 \text{ in TEXTURE block,} \\ 1.25 & \text{for } \left(k_1^2 + k_2^2\right) > 16 \text{ in TEXTURE block.} \end{cases} \tag{2.31}$$

Finally, Wei and Ngan computed intra-band contrast masking as

$$e_{\text{intra}}^{\text{WN}}(\mathbf{k},\mathbf{n}) = \begin{cases} 1, & \text{for } \left(k_1^2 + k_2^2\right) \leq 16 \text{ in} \\ & \text{PLAIN and EDGE block,} \\ \min\left\{4, \max\left\{1, \left(\frac{|C(\mathbf{k},\mathbf{n})|}{T_b(\mathbf{k},\mathbf{n}) e_{\text{la}}^{\text{WN}}(\mathbf{k},\mathbf{n})}\right)^{0.36}\right\}\right\}, & \text{otherwise.} \end{cases} \tag{2.32}$$

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
1	8	0	-8	1
0	3	0	-3	0
0	1	0	-1	0

G_4

Fig. 2.7 Operators to determine weighted average of luminance changes (G_p)

For a viewing distance of six times of the image height, Chou and Li [CL95] estimated contrast masking in the pixel domain using the following expression:

$$e_2^{\mathrm{CL}}(\mathbf{x}) = 0.01L(\mathbf{x})(0.01G(\mathbf{x}) - 1) + 0.115G(\mathbf{x}) + 0.5, \tag{2.33}$$

where $G(\mathbf{x})$ is the maximal weighted average of the gradient around the pixel at $\mathbf{x}$. $G(\mathbf{x})$ is calculated by

$$G(\mathbf{x}) = \max_{j=1,2,3,4}\left\{\left|grad_j(\mathbf{x})\right|\right\}, \tag{2.34}$$

where

$$grad_j(\mathbf{x}) = \frac{1}{16}\sum_{p_1=0}^{4}\sum_{p_2=0}^{4} c(x_1 - 2 + p_1, x_2 - 2 + p_2)G_j(p_1, p_2), \tag{2.35}$$

and $G_j(\mathbf{x})$ are the four directional highpass filters shown in Fig. 2.7.

2.5 Error Pooling

The final step of many image quality metrics is to combine the errors normalized by $T(\mathbf{k}, \mathbf{n})$ computed for every spatial frequency (from DCT-II subbands) at all spatial location $\mathbf{n}$ into a single distortion measure [SJ89, Wat93]. Alternatively, these normalized errors can be combined into an error map using error pooling, which describes the amount of error of each pixel in the image.

An example of error pooling using the Minkowski metric can be expressed as

$$P(\mathbf{n}) = \left(\sum_{\mathbf{k}}\left|\frac{C(\mathbf{k}, \mathbf{n}) - \hat{C}(\mathbf{k}, \mathbf{n})}{T(\mathbf{k}, \mathbf{n})}\right|^{\beta_f}\right)^{1/\beta_f}, \tag{2.36}$$

where $\hat{C}(\mathbf{k}, \mathbf{n})$ is the quantized **k**th DCT-II coefficient of the **n**th DCT-II block and β_f is a constant for summation across frequency bands. For summation across frequency band, it is found that $\beta_f \approx 4$ [Wat82, GRN78, Leg78a, Leg78b, RG81, PAW93b, RAW97]. By summing all the errors in (2.36) over **n**, a single value describing the distortion of an image is then obtained. For spatial error pooling over **n**, several values of β_s have been adopted. Teo and Heeger [TH94b], Lubin [Lub93, Lub95], and Watson [Wat93] adopted β_s as 2, 2.4, and 4, respectively. Alternatively, error pooling can be performed over **n**, followed by over frequency bands [Wat93].

At near JND threshold, probability summation is well accepted as the basis for summation of signal energy (or distortion) across frequency and spatial domains [EB98]. For summation across frequency band, it is reported in [GRN78, Leg78a, Leg78b, RG81] that $\beta_f = 3.5$ [Wat82] is consistent with subjective evaluation. It has been found that summation across frequency bands with DCT-II basis functions is well modeled with $\beta_f = 2.4$ [PAW93b]. In target detection experiments [RAW97], it is found that $\beta_s = 4$ provides the closest match to psychophysical results for spatial summing.

To obtain a single distortion value describing the amount of distortion in a compressed image, spatial summing is performed after summation across frequency bands or vice versa. If summation across frequency bands is first performed, the perceptual distortion score P_1 of an image becomes

$$P_1 = \left(\sum_{\mathbf{n}} P(\mathbf{n})^{\beta_s} \right)^{1/\beta_s}. \tag{2.37}$$

Alternatively, localized pooling of an image can be performed. One such example is found in [HK02], where spatial summing is performed within the foveal region $F(\mathbf{k}, \mathbf{n})$. The distortion within the foveal region is given as

$$P_F(\mathbf{k}, \mathbf{n}) = \left(\sum_{(\mathbf{k}', \mathbf{n}') \in F(\mathbf{k}, \mathbf{n})} \left| \frac{C(\mathbf{k}', \mathbf{n}') - \hat{C}(\mathbf{k}', \mathbf{n}')}{T(\mathbf{k}', \mathbf{n}')} \right|^{\beta_F} \right)^{1/\beta_F}, \tag{2.38}$$

where $\beta_F = 4$. Using the foveal distortion $P_F(\mathbf{k}, \mathbf{n})$, the distortion for the **k**th DCT-II coefficient is computed as

$$P_F(\mathbf{k}) = \max_{\mathbf{n}} \{ P_F(\mathbf{k}, \mathbf{n}) \}, \tag{2.39}$$

and the single distortion measure of the image becomes

$$P_F = \max_{\mathbf{k}} \{ P_F(\mathbf{k}) \}. \tag{2.40}$$

Zhang et al. [ZLX05, ZLX08] suggested the following expression for spatial summing:

$$P(\mathbf{k}) = \begin{cases} \left(\sum_{\mathbf{n}} |d_{\mathrm{JND}}(\mathbf{k},\mathbf{n})|^{2.3}\right)^{1/2.3}, & \text{for } \mathbf{k} = (0,0),(1,0),(0,1), \\ \left(\sum_{\mathbf{n}} |d_{\mathrm{JND}}(\mathbf{k},\mathbf{n})|^{4}\right)^{1/4}, & \text{otherwise,} \end{cases} \tag{2.41}$$

where $d_{\mathrm{JND}}(\mathbf{k},\mathbf{n}) = \big(C(\mathbf{k},\mathbf{n}) - \hat{C}(\mathbf{k},\mathbf{n})\big)/T(\mathbf{k},\mathbf{n})$. In this case, the perceptual distortion score P_2 is computed as:

$$P_2 = \left(\sum_{\mathbf{k}} P(\mathbf{k})^{\beta_f}\right)^{1/\beta_f}. \tag{2.42}$$

2.6 Summary

In this chapter, we reviewed three properties of the HVS, namely, CSF, luminance adaptation, and intra- and inter-band contrast masking. These properties play important roles in the design of image quality metric and computational model for JND. It is known that DCT does not match the channel decomposition mechanism of the HVS. To mitigate the issues arise from the mismatch of frequency decomposition of the HVS and DCT, Tran and Safranek [TS96] introduced a mapping from DCT-II coefficients to the cortex bands. Section 2.1 introduced the cortex filters, and reviewed the mapping of DCT-II coefficients to the cortex bands. Section 2.2 presented a widely adopted CSF proposed by Ahumada and Peterson [AP92], which is used to compute the base detection threshold of DCT subband.

Elevation in the base detection threshold is attributed by the luminance adaptation and contrast masking. These elevation parameters were reviewed in Sects. 2.3 and 2.4, respectively. Luminance adaptation refers to the variation of the base detection threshold due to the local luminance. Two forms of contrast masking, namely, the intra- and inter-band contrast masking were described in Sect. 2.4. Most PICs account for intra-band contrast masking due to its simple formulation; however more accurate representation of the JND threshold should also include inter-band contrast masking. Two estimations of the inter-band contrast masking using block classification and cortex filtering were shown in Sect. 2.4.

In Sect. 2.5, we discussed how a PIC uses a single distortion measure or distortion map to determine the permissible compression of an entire image (using a single distortion measure) or different regions of the image (using a distortion map) at a predefined image quality. The next chapter shall review the integration of these computational models in DCT-based image coders.

http://www.springer.com/978-981-287-542-6

Perceptual Image Coding with Discrete Cosine Transform
Tan, E.-L.; Gan, W.-S.
2015, VIII, 69 p. 35 illus., 6 illus. in color., Softcover
ISBN: 978-981-287-542-6