Just Noticeable Distortion Model and Its Application in Image Processing

JIA YUTING

NATIONAL UNIVERSITY OF SINGAPORE

2005

Just Noticeable Distortion Model and Its Application in Image Processing

JIA YUTING

(B.SCI., PEKING UNIVERSITY, BEIJING, CHINA)

A THESIS SUBMITTED

FOR THE DEGREE OF MASTER OF ENGINEERING

DEPARTMENT OF ELECTRICAL AND COMPUTER

ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2005

Acknowledgements

With the completion of this master's thesis, the author would like to thank many people for their kind help and precious suggestions in the entire course of postgraduate study. Firstly, I would like to express the deepest gratitude to my supervisors, Associate Professor Ashraf Kassim and Dr. Lin Weisi, for their pertinent and helpful guidance. Because of their insightful vision, I entered into the very promising realm of perceptual image/video processing. Because of their patience and encouragement, I could get through the research difficulties successfully and make constant development during the project.

Many thanks should go to the seniors in the Embedded Video Lab as well as the Vision and Image Processing Lab in National University of Singapore. I would like to thank Lee Weisiong, Yan Pingkun, Li Ping and Wang Heelin for sparing their time to discuss with me. Their experience and support really unveiled some research doubts in my mind, which paved the way for the thesis. In addition, I am also grateful to the other peers and friends in these two labs for creating an aspiring and enjoyable atmosphere for studying.

I should not forget to thank my dearest parents in China and my uncle and aunt in Singapore. Their concerns and supports give me more strength to meet the challenges and seek development.

Last but not the least, I would like to express the sincere gratitude to my lovely housemates and friends. With all of you, I have spent a good time in Singapore. That is of particular importance to my master study.

Table of Contents

Summary

Advances in vision research are contributing to the development of image processing. Digital communication systems can be optimized by incorporating the perceptual properties of the human eye to ensure that the resulting images are more appealing to human viewers.

This thesis discusses the relevant properties of the human visual system (HVS) and presents a spatio-temporal just-noticeable distortion (JND) model in the discrete cosine transform (DCT) domain. The proposed JND model thus incorporates the relatively well developed spatial mechanism of the HVS (including luminance adaptation and contrast masking) as well as the temporal mechanisms with the aim of deriving a vision model which is consistent for both image and video applications. Subjective experiments show that the proposed model outperforms the related existing JND models, especially when high motion takes place.

The JND model facilitates perceptual image/video processing. Based on an improved pixel-based JND profile for the image, an image compression scheme for both perceptually lossless and perceptually optimized lossy compression have been then proposed and discussed. Experiments show that the proposed coding scheme leads to

higher compression in the perceptually lossless mode and better visual quality in perceptually optimized lossy mode compared with related coding methods.

List of Figures

List of Tables

CHAPTER

1

Introduction

1.1 Motivation

Modern design of the visual communication system aims at using the least resources to achieve the highest visual quality with respect to the coding constraint (e.g., bit-rate, complexity and delay). In most circumstances, the human visual system (HVS) makes final evaluations on the quality of images and videos that are processed, transmitted and displayed. Thus it is essentially futile to spend significant effort on encoding those signals that are beyond human perception. Researchers have already realized the importance of considering human visual properties and implemented some of them in the existing image/video coding standards. For example, the quantization tables of JPEG & MPEG can be adjusted to fit human visual sensitivity [1-5].

The characteristics of HVS influence the human perception in many aspects. Luminance adaptation property explains the fact that it is safer to insert noise into low-intensity or high-intensity regions than mid-intensity regions. The contrast

masking phenomenon gives good reasons why more distortion can be tolerated in texture areas of an image. The contrast sensitivity theory indicates that the human eye is actually sensitive to the contrast rather than the absolute intensity of the signal and the human perceptive capability highly depends on the frequency of the signal. This finding gives sound foundation for assigning a higher quantization step for high-frequency component in image/video compression. In video sequences, the temporal mechanism can not be ignored. The contrast sensitivity property has its extension in the temporal domain and the temporal component interweaves with the spatial component for different spatio-temporal frequencies. For example, in the region where high motion (high temporal frequency) takes place, details (signals of high spatial frequency) are not so crucial for perception; but in the low-motion region, detailed information is quite obvious and should be carefully managed. In addition, the human eye tends to track moving objects, and this mechanism helps alleviate the blurring effect of motion. Only by properly considering the combination effect of those factors above can we derive a comprehensive model to predict the perception of HVS.

An effective and convenient way to realize perception-based application is through deriving the *just-noticeable distortion* (JND) map for images or video sequences. JND, which accounts for the smallest distortion that the human eye perceives [6], serves as the benchmark perceptual threshold to guide an image/video processing task. In image compression schemes, JND can be used to optimize the quantizer [7-10] or to facilitate rate-distortion control [11]. Information of higher perceptual significance is given

more bits and preferentially encoded, so that the resultant image is more appealing. In video compression schemes, JND plays more diverse roles. As in image compression, JNDs for video can be used to improve quantizers and bit allocation [12,13]; moreover, motion estimation can be facilitated with the help of the JND profile [14]. For both image and video, objective quality evaluation based on the characteristics of the HVS can be achieved by using the JND [15-21].

JND estimation for images has been relatively well developed. However, there has not been much work on the study of JND for videos. The majority of the related work has been devoted to the evaluation of perceptual error between an original video sequence and its processed version [16,18,19,20,21,22,23], without explicit mathematical expressions for JND. In fact, JND is a property of video itself, even when no processing is performed on it. Therefore, it is meaningful to derive an explicit formula for the calculation of JND with any frame in a given video sequence, after incorporating the temporal characteristics of the HVS. Furthermore, a *stand-alone* JND estimator for the video signal would facilitate wider and/or more convenient applications in visual processing of different nature and constraints.

HVS-based technology is becoming a good tool in the information processing field, providing guidance for determining which information should be maintained and which can be safely omitted. As more and more psychophysical properties of HVS are unveiled, perceptual technology will keep on developing.

1.2 Objectives

This thesis mainly aims at explicit JND estimation based upon the perceptual characteristics of the human visual system. An estimator that can be adopted for both image and video in the DCT domain is proposed first. This JND model combines the effects of eye-movement compensated spatio-temporal contrast sensitivity function, luminance adaptation and contrast masking, thus providing a more accurate estimation of distortion thresholds than previous models. Secondly, a perceptual image compression scheme based on an enhanced pixel-based JND model is proposed. This coding method gives an example of how the JND model can be applied to image/video processing.

1.3 Contributions

The contributions of this thesis can be summarized as follows:

- Major properties of human perception with regard to the proposed model and scheme are explored and investigated, and well-known perceptual models related to the proposed JND model are discussed.
- A new spatio-temporal DCT-based CSF model, which takes into account the effect of eye movement on visual perception, is proposed. The spatio-temporal CSF model is combined with luminance adaptation and contrast masking to form a complete JND model. Subjective testing shows that our model outperforms existing models in JND value prediction, and therefore achieves better noise mask

in the image/ video.

- According to the different response of the human eye to the distortion in different areas (*smooth, edge, texture*) of an image, a block classification module is adopted for contrast masking. Incorporating the more accurately predicted contrast masking based on the local texture activity, an improved JND model for the image is achieved. This JND model is among the few perceptual models that estimate the visual threshold in the pixel domain.
- Based on the modified pixel-based JND estimator for the image, an image compression scheme for both perceptually lossless and perceptually optimized lossy compression is proposed. Experiments show that our scheme is effective and efficient for both modes compared with related coding schemes.

1.4 Organization

The thesis is outlined as follows:

Chapter 2 discusses the properties of the human visual system and its contribution to human perception. Temporal properties including temporal contrast sensitivity function, temporal masking and eye movement effect are presented in detail because of their importance to the proposed perceptual model.

Chapter 3 presents several models of the human visual system particularly those

spatio-temporal contrast sensitivity function (CSF) models and just-noticeable distortion (JND) models for images, because they are the basis for our proposed JND model. The human vision models designed for video applications have also been summarized in this chapter.

Chapter 4 shows the design of the proposed JND estimation model. Firstly, the eye movement compensated spatio-temporal CSF is elaborated because of its essential role in the calculation of JND calculation. Secondly, luminance adaptation and the improved contrast masking scheme are included to derive a comprehensive model for JND estimation.

Chapter 5 gives the experimental results and discussions for the model validation. The proposed model is compared with related existing JND estimators by specially designed experiments.

Chapter 6 introduces a modified version of a pixel-based JND model for the image. Based on the JND model, a perceptual image compression scheme is designed for both perceptually lossless and perceptually optimized lossy compression. Experiments are conducted to show that this human vision based coding scheme is superior to the traditional coding scheme (without perceptual consideration) for both modes.

Chapter 7 concludes the thesis with discussions and suggestions for the future research

endeavors.

CHAPTER

2

Perceptual Characteristics of Human Vision

2.1 Introduction

The working of the human visual system (HVS) can be divided into two stages: lower level processing and higher level processing. Lower level processing involves the functions of the optics, retina, lateral geniculate nucleus, and striate cortex, while higher level processing incarnates more complex mechanism such as attentive vision, Gestalt and figure/ground effects [24]. Since higher level processing elements are not understood well enough and their effects are not that predictable as those in lower level processing stage, current HVS models mostly focus on the simulation of lower level mechanisms. The effective application of these models justifies the approximation.

In general, the basic elements that influence the visual sensitivity include *contrast sensitivity function* (CSF), *luminance adaptation* and *contrast (texture) masking*. For video applications, temporal properties such as *temporal CSF* and *temporal masking* can be added. In this chapter, these spatial and temporal mechanisms of the early-stage

human perception as well as their roles in perception will be discussed.

2.2 Contrast Sensitivity Function

The *contrast sensitivity function* (also called the *modulation transfer function*) demonstrates the varying visual acuity of the human eye towards signals of different spatial and temporal frequencies. Instead of the absolute intensity of signal, the human eye responds to contrast. In psychophysical experiments, the threshold contrasts are measured for viewing traveling sine wave gratings (Figure 2.1) at various spatial frequencies and velocities (the standing sine waves can be regarded as traveling waves at 0 velocity and counterphase flicker stimuli can be decomposed into two opposing traveling waves [10]). The contrast sensitivity function (CSF) is defined as the inverse of this measured threshold contrast.

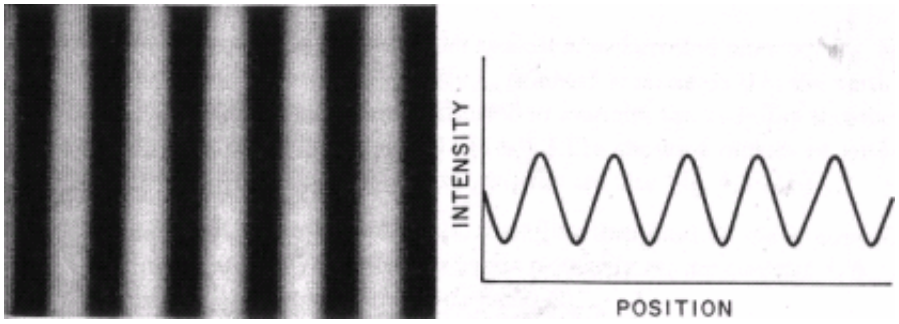


Figure 2.1 Illustration of traveling sine wave gratings [25]

Spatial contrast sensitivity function, as shown in Figure 2.2, describes the influence of the spatial frequency on visual sensitivity. The parabola curves show that the human eye has different acuity for different spatial frequency. Specifically, the acuity for high spatial frequencies is comparatively low. This fact has been utilized to design

perceptually optimized coding schemes where few bits are given to high spatial frequency components. In the measurement of the contrast sensitivity, it should be noticed that spatial frequencies are in units of cycles per degree of visual angle [24]. This implies that the contrast sensitivity function also varies with the viewing distance. For instance, the imperceptible details of an image may become visible when the viewer moves closer to it. Therefore, a minimum viewing distance needs to be clarified when a visual model is derived. Strictly speaking, the HVS is not perfectly isotropic and orientation has some adjustive effects on CSF [24]. However, for a visual model, isotropic assumption can be a rational approximation.

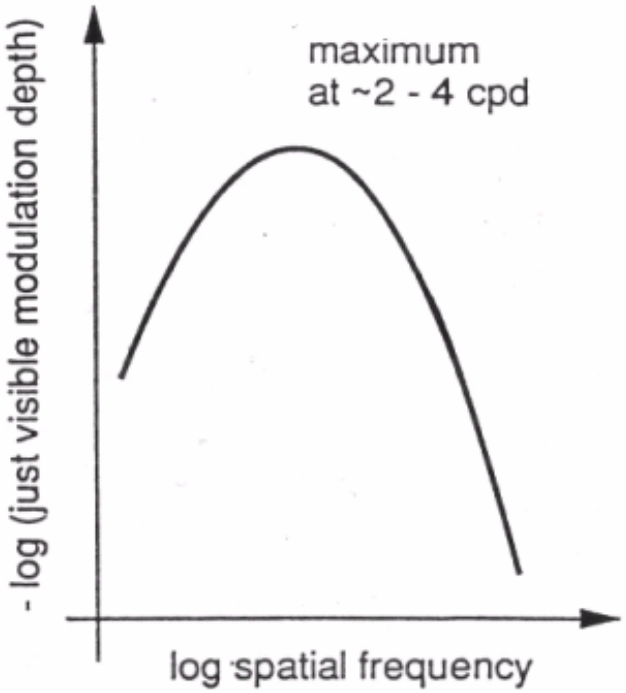


Figure 2.2 Typical spatial contrast sensitivity function [26]

Another notable factor that affects the CSF is the background luminance. We define it as *luminance adaptation* and will discuss it in details in Section 2.3.

In non-static scenarios, the temporal frequency plays an indispensable role in shaping contrast sensitivity. Not only the levels but also the shapes of the spatial CSF change with different temporal frequencies. Figure 2.3 and 2.4 illustrate a well-known spatio-temporal CSF model by Kelly [27]. As can be seen from these two figures, at low temporal frequencies, the contrast sensitivity curve holds a band-pass shape; while at high temporal frequencies, the contrast sensitivity curve holds a low-pass shape. It can also be observed that the sensitivity of the eye decreases with the increase of spatial and temporal frequencies.

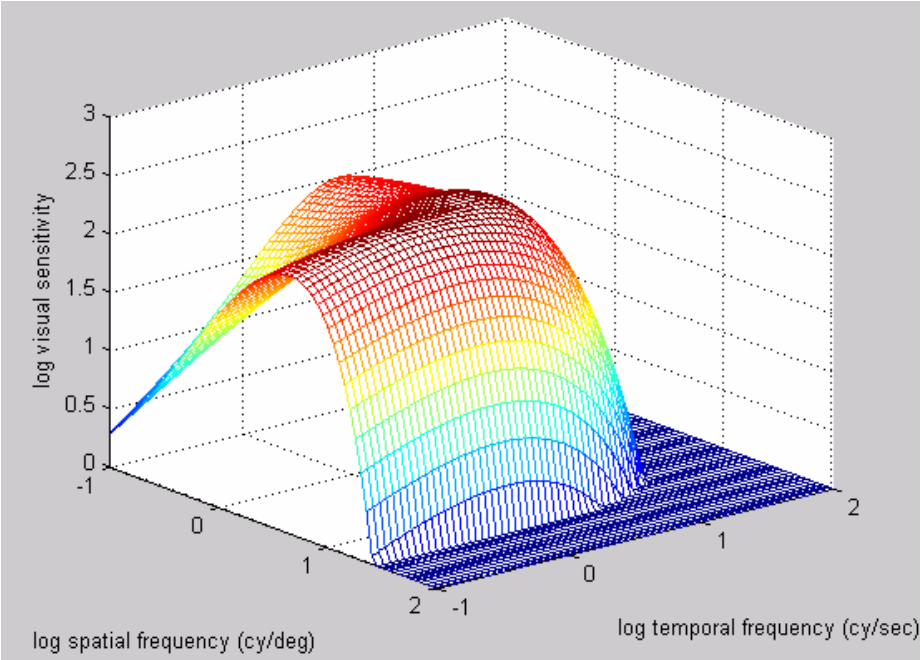


Figure 2.3 Spatio-temporal contrast sensitivity surface

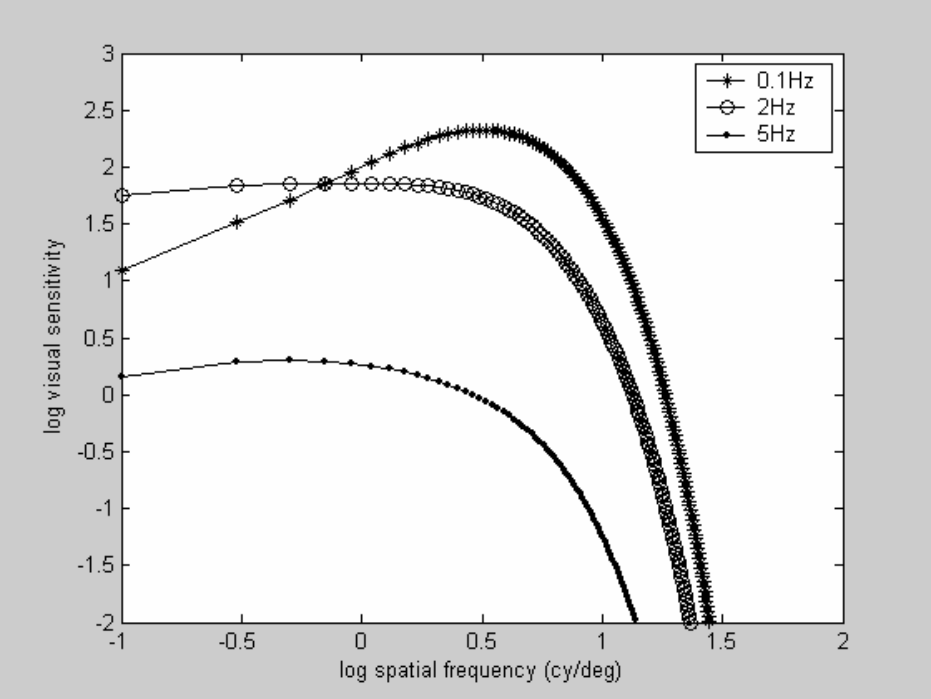


Figure 2.4 Spatial contrast sensitivity curves at different temporal frequencies

Kelly [27] measured his spatio-temporal CSF surface under the condition that eye movements were strictly controlled. However, in practice, eye movements can have important effects on the perceptual threshold and should not be ignored in the vision modeling. Based on Kelly's stabilized spatio-temporal CSF model, Daly (1998) [10] built an eye movement model and applied it to an improved CSF model which is valid for unconstrained natural viewing conditions. More details of eye movement will be explored in Section 2.5 and Daly's model will be elaborated in Chapter 3.

2.3 Luminance Adaptation

The human eye operates over a large range of light intensities. Luminance adaptation refers to the visual sensitivity adjustment for different light levels. Since the HVS is sensitive to the luminance contrast rather than the absolute luminance, the luminance

adaptation is usually modeled by measuring the increment threshold or contrast against a background of certain luminance. Figure 2.5 illustrates this mechanism.

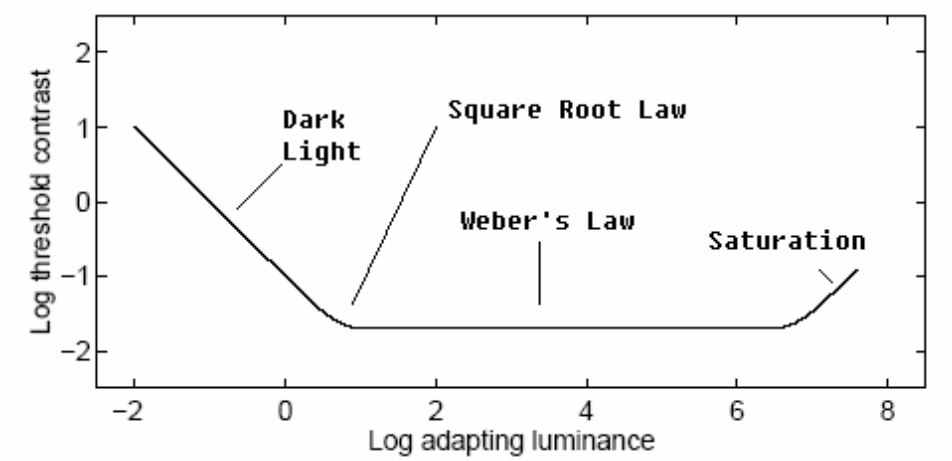


Figure 2.5 Description of luminance adaptation [28-30]

Generally, the working of the mechanism can be divided into four sections [29]:

- Dark light

- Square Root Law (de Vries-Rose Law)

- Weber's Law

- Saturation

In the “dark light” section, the sensitivity is limited by the internal noise of the retina so that the increment threshold remains the same without depending on the background luminance variance. In the “saturation” region where the background intensity is high, the slope of curve in Figure 2.5 begins to increase rapidly, which means that the eye becomes unable to detect the stimulus. The “square root law” (de Vries-Rose law) region involves a complex mechanism, the details of which can be found in [31]. Compared with the three sections above, “Weber’s law” demonstrates a more

important aspect of our visual system, because it operates at a moderate background luminance which is a more common viewing environment. Weber's law refers to the phenomenon that the threshold contrast remains the same regardless of ambient luminance. This contrast constancy property can be mathematically expressed as:

$$C = \Delta L/L \tag{2.1}$$

Where the threshold contrast C is a constant. ΔL is the luminance offset on a uniform background luminance L. Only when ΔL is greater than C·L can it be perceived by human eye.

2.4 Masking Phenomenon

In general, masking occurs where there is a significant change in luminance. For example, spatial masking is obvious at texture areas where the image activity is intense, and temporal masking can take place when there is an abrupt change of scene leading to a considerable change of intensity.

2.4.1 Contrast Masking

Contrast masking (also known as spatial masking) refers to the reduction in visibility of one image component (the target) in the presence of another image component (the masker) [24]. Generally, we consider two kinds of contrast masking phenomenon: 1. inter-band masking: accounts for the masking effect among different subband; 2. Intra-band masking: refers to the combined effect of sufficient amount of coefficients in the same subband.

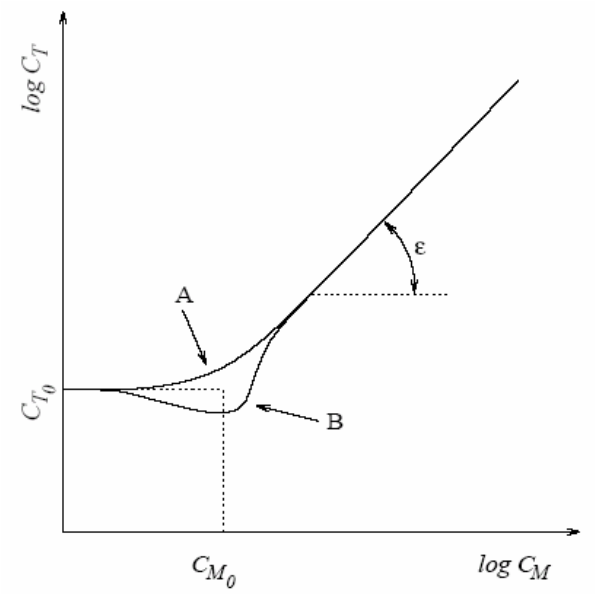


Figure 2.6 Illustration of typical masking curves.

For stimuli with different characteristics, masking is the dominant effect (case A). Facilitation occurs for stimuli with similar characteristics (case B).

In modeling contrast masking, the detection threshold for a target stimulus is measured when it is superimposed on a masker with varying contrast. Pioneer researchers have done experiments on this [32,33] and Figure 2.6 illustrates a typical masking curve [28]. The horizontal axis (logC_M) shows the logarithm of the masker contrast, and the vertical axis (logC_T) shows the log of the target contrast at detection threshold. C_{T0} denotes the detection threshold for the target stimulus without any masker. As shown in the figure, there are two cases A and B when the masker contrast is close to C_{M0}. In case A, masker and target have different characteristics and there is a smooth transition from the threshold range to the masking range. While in case B, the masker and target share similar properties and the *facilitation* effect occurs: the target is easier to be perceived due to the masker in this contrast range. Masking is strongest when the

interacting stimuli have similar characteristics, i.e. similar frequencies, orientation, colors, etc. [28].

In practical image/video applications, the extent of contrast masking depends on the local intensity activity of the image. For example, it has been found that the HVS sensitivity to error is generally high in smooth, or plain areas, and low in the texture area [34]; while the sensitivity for edge areas lies in between. Contrast masking explains the fact that similar artifacts are visible in some areas of an image but can not be detected in other places.

In the design of a vision model, contrast masking is usually locally calculated as an elevation factor for the base threshold that is determined by contrast sensitivity and luminance adaptation [3,35,36].

2.4.2 Temporal Masking

Temporal masking occurs because of the temporal discontinuities in intensity, for instance, scene cuts. It has been found that with the increase of interframe luminance difference, the error visibility threshold is increased [1,37]. Specifically, after the scene change, the perceived spatial resolution is reduced significantly immediately and this phenomenon will last up to 100ms [38]. Because of the difficulty in predicting temporal masking, very few models have taken it into account. In Watson's digital video quality metric (DVQ) model [39], temporal masking is incorporated in its

masking step with a construction of a temporally filtered masking sequence. Moreover, as indicated by Lucas etc. [40], the occurrence of temporal masking is also related to the spatial activity of the frame: the temporal masking is more applicable in areas of high details than smooth areas.

2.5 Eye Movement

As discussed in Section 2.2, the spatial CSF changes with different temporal frequencies. Because of the inconvenience of measuring the temporal frequency, the dependence of spatial acuity on temporal frequencies can be studied through exploring the relationship between the spatial sensitivity and the velocity of the image traveling across the retina [10,27,41]. It should be noted that this retina velocity of the human eye is different from the image plane velocity, due to the effect of the eye movement.

Generally, three types of eye movements are considered in the vision research [10,42]. They are the *natural drift eye movements*, the *smooth pursuit eye movements* and the *saccadic eye movements*. The natural drift eye movements are also referred to as involuntary fixation mechanism, which is responsible for perception of static imagery during fixation and helps lock the eyes on the object of interest. The saccadic eye movements (voluntary fixation mechanism) account for the behavior of the eye to rapidly relocate the fixation point on object of interest. The smooth pursuit eye movements (SPEM) occur when the eye is tracking a moving object [10]. This mechanism is especially significant in that it compensates the loss of sensitivity due to

motion. Fast moving objects tend to blur the image, however, SPEM reduces the object's velocity from the image plane to retinal so that image spatial resolution actually doesn't suffer from a substantial reduction in regions of motion. According to [41], the function of SPEM can be summarized as:

(1) maintaining the object of interest in the area of highest spatial acuity of the visual field, and

(2) minimizing the velocity of the image across the retina by matching eye velocity to image velocity.

The execution of the three types of eye movements relies on the target velocity, and the relationship between them is shown in Table 2.1.

Table 2.1 The relationship between target velocity and the type of eye movement

target velocity (deg/sec)	0.8 – 1.5	1.5 – 80	> 80
type of eye movements	the natural drift eye movements	the smooth pursuit eye movements	the saccadic eye movements

It should be noted that when the target velocity surpasses some limit (e.g. 20-30 deg/sec as reported [10]), the eye can not perform a perfect tracking (SPEM), thus a certain loss of visual sensitivity will be suffered accordingly. Considering this factor in modeling the human vision is essential in achieving consistent simulation of human perception.

In summary, the existence of eye movement leads to the consequence that spatial acuity does not directly depend on the image velocity, but on the retinal velocity which is influenced by the ability of the visual system to track objects [41].

Incorporating eye movement into modeling vision can be realized in several ways. Westen etc. (1997) [43] proposed an eye movement estimation algorithm to compensate the contrast sensitivity function, so that not more noise or blur is allowed in moderately moving object than in static objects. Daly (1998) [10] modified Kelly's stabilized CSF by inserting an eye model, through which a relationship is built between the retinal velocity and image plane velocity. The improved CSF model can fit unconstrained natural viewing conditions and is proved to be more consistent with human perception.

2.6 Pooling

The preliminary perception of human vision processes the information in various channels and then the outputs of these channels are integrated in the subsequent brain areas to form vision. The course of gathering the data from different channels according to rules of probability or vector summation and calculating them into a single number for each pixel of the image, or a single number for the whole image is known as *pooling* [28]. Two well-known mathematical models: the probability summation and the vector summation have been proposed for pooling, though the nature of this mechanism is still to be explored.

The probability summation rule can be summarized as follows:

If there are a number of independent "reason" i for an observer to view the presence of a distortion, each having probability P_i respectively, the overall probability P of the observer noticing the distortion is:

$$P = 1 - \prod_i (1 - P_i) \tag{2.2}$$

the dependence of P_i on the distortion strength x_i can be described by the psychometric function:

$$P_i = 1 - e^{-x_i^{\beta_i}} \tag{2.3}$$

If we set the homogeneity assumption that all β_i are equal, (2.2) & (2.3) can be combined to form:

$$P = 1 - e^{-\sum x_i^{\beta}} \tag{2.4}$$

Vector summation (Minkowski summmation) is used to obtain the combined effect of several mechanisms. If the individual effects of N mechanisms are represented by x_i (i=1,...,N), the combined effect x can be shown as:

$$x = \sqrt[\beta]{\frac{1}{N}\sum x_i^{\beta}} \tag{2.5}$$

β is a summation constant which can hold different values for different experiments and implements [28]. In most studies and applications, β = 2 is found to give good experimental results. In some cases, if we assume that high distortion tends to draw viewer's attention more than low distortions, β can be set to a higher value to weight

the higher distortion more.

For videos, pooling in both spatial domain and temporal domain are needed. Since the perceived distortion in an image sequence is a function of more than just one frame, temporal summation accounts for the persistence of the images on the retina and should take into account the combination of several successive frames. Commonly, 100msec is regarded as the delay time of a signal on the retina [44] and the combined effect of temporally successive frames can be regarded as imposing a low-pass time window on the image sequence. This modeling can also explain the smoothness of perceived quality recording in perceptual subjective experiments [45].

The pooling method is actually very flexible and can be determined according to individual needs. For example, in order to take into account the focus of attention of human observers, spatial summation can be operated on blocks, each of which covers two degrees of visual angle (the dimension of the fovea).

2.7 Summary

In this chapter, spatial and temporal perceptual properties of human visual systems have been particularized. We introduced the mechanisms of contrast sensitivity, luminance adaptation, masking phenomenon, eye movement and pooling, based on which their relationship with human perception are illuminated. All these characteristics discussed above are the fundamentals for deriving the perceptual models and they make the preparations for our subsequent discussion.

CHAPTER 3

Spatio-temporal Models of the Human Vision System

3.1 Introduction

Model of the human visual system (HVS) plays an essential role in perceptual visual processing system. As the pertinent and practical simulation of the human vision, the perceptual model builds a bridge between vision research and practical applications. The human vision models for images have been relatively well developed. In particular, several models for estimating the just noticeable distortion of images were proposed.

Pixel-based JND models such as the ones proposed in [37,46,47] basically take into account two components: luminance adaptation and contrast masking. In [46], the maximum effect between luminance adaptation and contrast masking is used for JND estimation, while in [37], luminance adaptation is regarded as the major factor affecting JND. The contributions of luminance adaptation and contrast masking are accumulated in [47] for a more general pixel-based JND model. In a subband domain,

spatial contrast sensitivity function (CSF), luminance adaptation, and contrast masking can be incorporated into a JND model [2,3,4,35,36]. An early scheme for the perceptual threshold was developed in [2] with DCT decomposition, based upon spatial CSF, and was improved into the DCTune model [36] after luminance adaptation effect had been added to the base threshold and contrast masking [32,48] had been calculated as the elevation factor. More recently, the DCTune model was modified [3] with a foveal region being considered instead of a single pixel. The block classification for different local structures was introduced in [34] for accounting the contrast masking effect. In [35], more realistic luminance adaptation was also considered for digital images to fit the empirical parabola curve [49] better (especially in bright and dark areas).

Compared with the effort devoted to JND estimation for images, there has not been much work on the study of JND for videos. One reason is that more knowledge of temporal mechanisms in the HVS is still to be unveiled. Another reason may come from the fact that temporal processing within the human eye is not easy to be controlled and predicted. The majority of the related work has been devoted to the evaluation of perceptual error between an original video sequence and its processed version [16,18,19,20,21,22,23], without explicit mathematical expressions for JND. In fact, JND is a property of video itself, even when no processing is performed on it. Therefore, it is meaningful to derive an explicit formula for the calculation of JND with any frame in a given video sequence, after incorporating the temporal

characteristics of the HVS. Furthermore, a *stand-alone* JND estimator for the video signal would facilitate wider and/or more convenient applications in visual processing of different nature and constraints.

The critical issue in designing a vision model for video is modeling the temporal mechanism of the HVS. Therefore, in this chapter, we will first introduce several spatio-temporal CSF models for this key task. Then JND models for the image will be discussed. In most cases, JND models for the video are actually the extensions of those models for the image with the consideration of relevant temporal properties. Finally, several practical the HVS models designed for video will be summarized. Besides,the temporal properties, these models also incorporate the spatial properties, similarly considered in the HVS models for images.

3.2 Spatio-temporal Contrast Sensitivity Models

Spatio-temporal Contrast sensitivity is very important for modeling the human visual system. Compared with the HVS models for the image, the HVS models for video sequences need to also take into account the dependence of the human sensitivity on temporal frequencies. So far, this property is best presented by the spatio-temporal CSF model. Figure 2.3 shows a classic envelope of visual sensitivity for spatiotemporal frequencies. If we cut the 3-D surface at different temporal frequencies, we can obtain the 2-D curve of different shapes (Figure 2.4). This corresponds to the experimental finding that the spatial contrast sensitivity function has its normal

bandpass shape at low temporal frequencies, whereas it gets a lowpass shape at high temporal frequencies [50]. Similarly, if we cut the 3-D surface at different spatial frequencies, it also can be seen that the temporal contrast sensitivity function has a bandpass shape at low spatial frequencies and a lowpass shape at high spatial frequencies.

3.2.1 Fredericken and Hess' two-temporal-mechanism model [53]

According to the psychophysical studies of the HVS, it is now believed that the initial stage of visual processing involves a series of spatio-temporal filters. Sensitivities with respect to the spatial frequencies were substantially explored, while less attention was given to the investigation of the temporal mechanism and how it co-varies with spatial frequency. In order to find the rationale of the spatio-temporal covariation in the human perception, R. F. Hess & R. J. Snowden [52] conducted a parametric assessment using a novel temporal masking paradigm evaluating the most sensitive temporal properties. Their experimental results suggested that the spatial dependence of the temporal surface can be adequately represented by no more than three broadband mechanisms. The evidence for the low pass mechanism and a band pass mechanism centered at 8 Hz is strong, while the second band pass mechanism is less clear-cut. A well-known best-fitting model for the multiple temporal mechanisms was proposed by Fredericksen & Hess in 1998. They used an impulse response basis set to describe the temporal mechanisms. The complete family of impulse responses is generated by taking successive temporal derivatives of a basic impulse response. After

undertaking temporal-noise-masking experiments among three subjects, two filters were selected from the basis set to give the best succinct data-fitting. Equations (3.1) and (3.2) denote the two filters h_0 and h_2, which correspond to one sustained and one transient mechanism, respectively.

$$h_0(t) = e^{-(\frac{\ln(t/\tau)}{\sigma})^2} \tag{3.1}$$

$$h_2(t) = \partial^2 h_0(t) / \partial t^2 \tag{3.2}$$

With a typical choice of parameters τ = 160 ms and σ = 0.2, the two filters can be described by Figure 3.1 and Figure 3.2 [28].

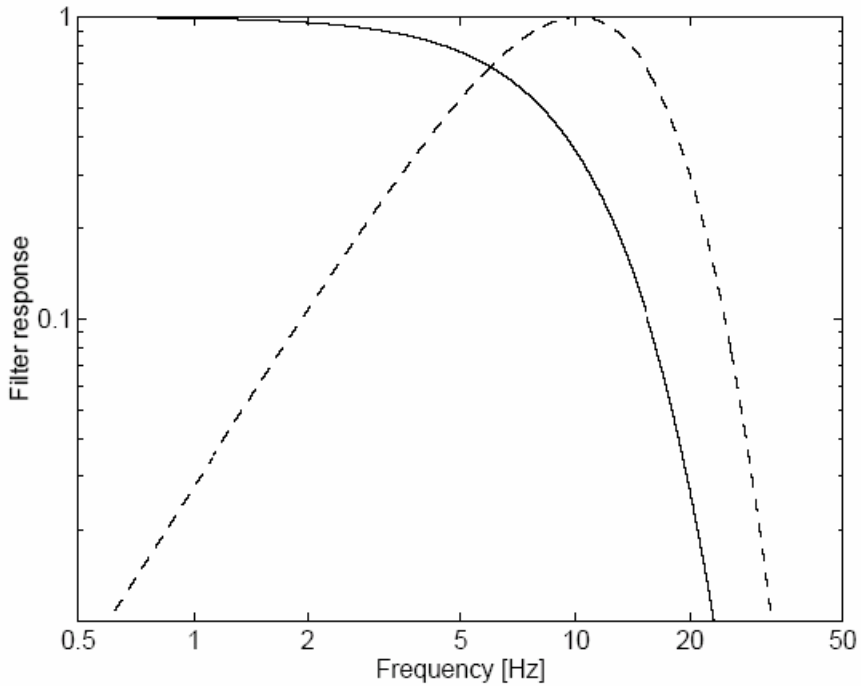


Figure 3.1 Frequency responses of sustained (solid) and transient (dashed) mechanism of vision [28,53]

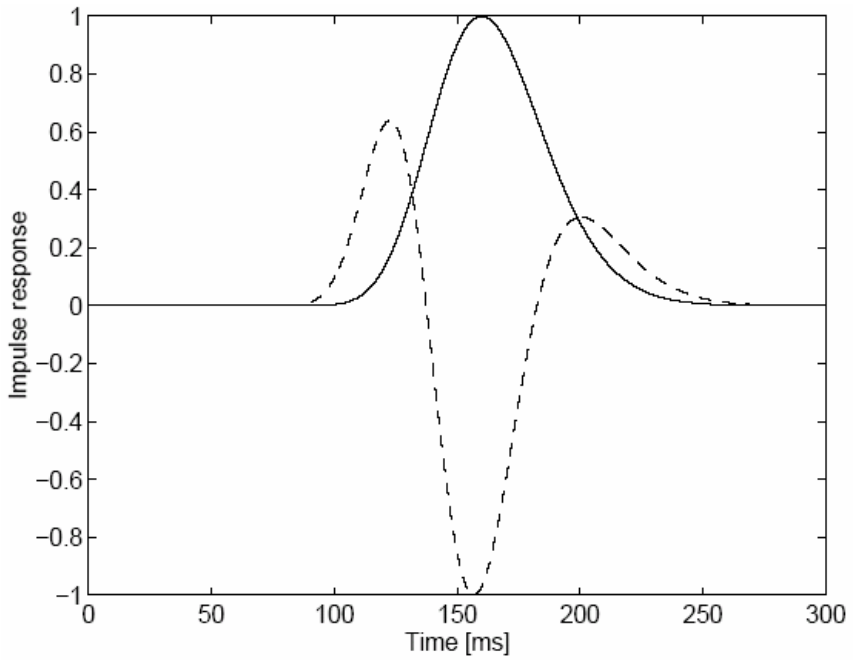


Figure 3.2 Impulse response functions of sustained (solid) and transient (dashed) mechanism of vision and its normalized second derivative [28]

The multi-channel temporal model has been used later by several perceptual video quality evaluation systems which will be summarized in Section 3.4.

3.2.2 Daly's CSF model [10]

Daly's CSF model is built upon Kelly's stabilized spatio-temporal threshold surface model, so first we will look into the theory of Kelly's model [27]. Spatio-temporal contrast sensitivity is sometimes referred to as the spatial acuity of the HVS depending on the velocity of the image traveling across the retina, where the retinal image velocity implicitly denotes the temporal frequency. In order to eliminate the influence of eye movements on the human visual sensitivity, Kelly performed the psychophysical experiments under the stabilized condition, which guaranteed that the velocity of the stimulus reflected the velocity on the retina. By measuring the contrast

sensitivity at constant velocity, Kelly proposed an expression that fits the data:

$$G(\alpha,\upsilon)=[6.1+7.3|\log(\upsilon/3)^3|]\times\upsilon\alpha^2\exp[-2\alpha(\upsilon+2)/45.9] \qquad (3.3)$$

where υ is the constant velocity in degrees per second, and $\alpha/2\pi$(cycles per degree) denotes the spatial frequency.

Since $\upsilon = \omega/\rho$, where ω represents the temporal frequency (cycle/second) and ρ represents the spatial frequency (cycle/degree), υ is actually the ratio of temporal to spatial frequency.

Although a large variation of curve shape occurs when the spatial or temporal frequency is held constant, all these constant-velocity curves have nearly the same shape according to the experiments. Each of the curves described by (3.3) is actually the 45^0 projection of the spatio-temporal threshold surface (Figure 2.3).

However, in natural viewing conditions, the velocity of the actual object is different from the retinal velocity of the perceived object because of the eye movement. The human eye tends to track the moving object so that the loss of sensitivity because of high motion can be compensated. Daly took into account this factor and developed Kelly's model into an unstabilized spatio-temporal threshold estimator. Equations (3.4) – (3.6) describe the spatiovelocity CSF model.

$$CSF(\rho,v_R)=k\bullet c_0\bullet c_2\bullet v_R\bullet(\frac{c_1\rho}{2\pi})^2\exp(-\frac{c_1\rho}{\pi\bullet\rho_{\max}}) \qquad (3.4)$$

$$k=6.1+7.3\bullet|\log(c_2v_R/3)|^3 \qquad (3.5)$$

$$\rho_{\max}=45.9/(c_2v_R+2) \qquad (3.6)$$

where ρ is spatial frequency in cycle/degree and v_R is the retina velocity in degree/second. k and ρ_{max} control the vertical shift of the sensitivity as a function of velocity and the horizontal shift of the function's peak frequency, respectively. Figure 3.3 and 3.4 give clearer descriptions of these two parameters.

As in Figure 3.4, with increasing retinal velocity, the sensitivity curve moves horizontally to the left so that the peak frequency is becoming smaller. c_0 and c_1 control the magnitude and the bandwidth of a CSF curve (Figure 3.5).

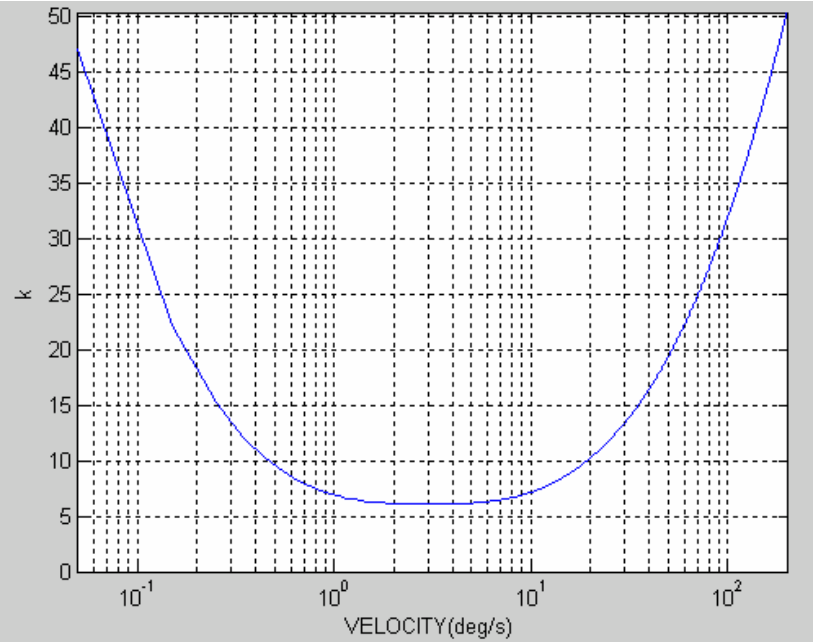


Figure 3.3 Parameter k vs. retinal velocity

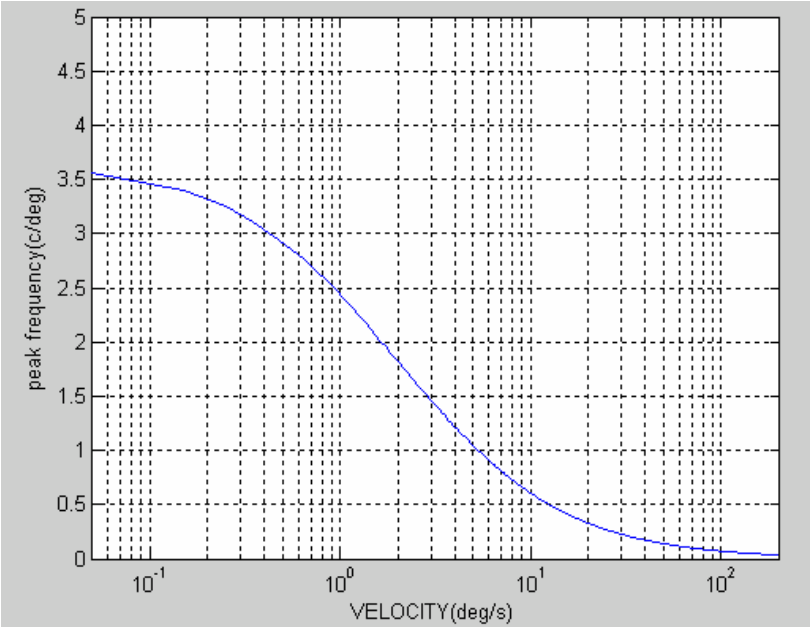


Figure 3.4 Peak frequency of spatio-temporal CSF vs. retinal velocity

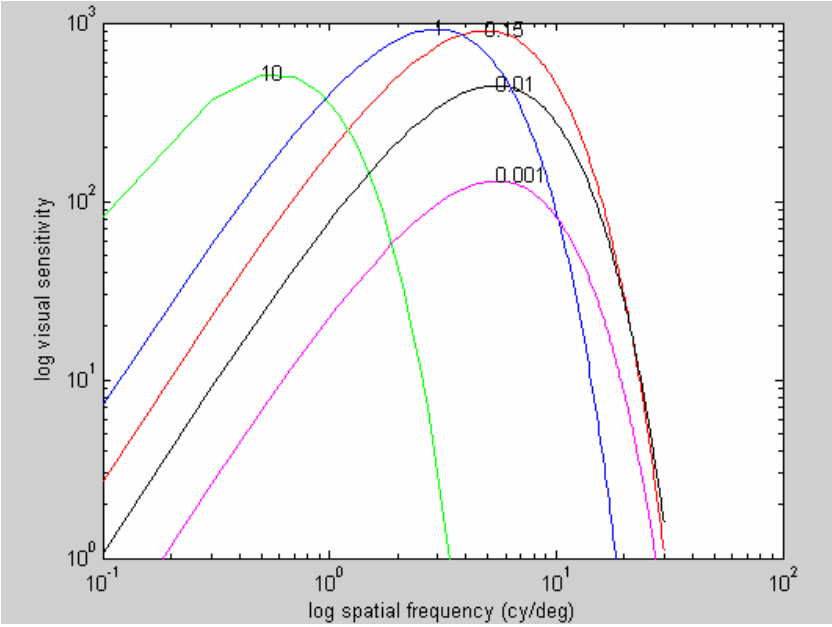


Figure 3.5 Spatial contrast sensitivity at different retinal velocities

(0.001, 0.01, 0.15, 1, 10 deg/sec) [10]

For practical application, Daly substituted the retinal velocity in (3.9) with the image plane velocity by using an eye movement model described by (3.7):

$$v_E = \min\lfloor (g \cdot v_I) + v_{\mathrm{MIN}},\ v_{\mathrm{MAX}} \rfloor \qquad (3.7)$$

Thus if the image plane velocity is v_I, the relationship between v_R and v_I can be calculated as:

$$v_R = v_I - v_E \tag{3.8}$$

(3.7) actually covers the three types of eye movements: natural drift eye movements, the smooth pursuit eye movements and the saccadic eye movements. Details about the eye movement can be referred to in Section 2.5. g is the gain of the smooth pursuit eye movement, v_{MIN} is the minimum eye velocity due to drift, and v_{MAX} is the maximum eye velocity before transitioning to saccadic movements.

With the experimental data, Daly set c_0, c_1, c_2 to 1.14, 0.67 and 1.7 respectively. v_{MIN} and v_{MAX} are assigned 0.15 and 80 deg/sec.

3.3 Just-Noticeable-Distortion Models for the image

JND models for the image are the basis for estimating JND values of the video sequence. Compared with estimating JND values in pixel domain, subband JND models are of particular interest because CSF can be more easily incorporated in subband and more images are coded in a subband scenario. A typical subband-based JND model consists of a base threshold and an elevation factor [36]. The former is determined by CSF and luminance adaptation while the latter denotes the effect of contrast masking.

3.3.1 Ahumada & Peterson's JND model [61]

Ahumada & Peterson developed their model to approximate visibility thresholds for

discrete cosine transform (DCT) coefficient quantization error. Besides considering the dependence of visibility sensitivity on spatial frequency, orientation and image luminance, the model also explored the effects of other image independent parameters such as display luminance and viewing distances. Therefore, it is applicable for display conditions other than those of the experimental measurements.

In an image, let ($n1,n2$) denote the 8×8 DCT block, (i,j) denote the location inside each block with $i, j = 0,1,\ldots 7$, the luminance threshold $T_{i,j}(n1,n2)$ can be calculated according to the following parabola equation:

$$\log T_{i,j}(n1,n2) = \log \frac{T_{\min}(n1,n2)}{r + (1-r)\cos^2 \theta_{i,j}} + K(n1,n2)(\log f_{i,j} - \log f_{\min}(n1,n2))^2 \quad (3.9)$$

where $T_{\min}$ is the luminance threshold at $f_{\min}$, the frequency where the threshold is smallest. K is the steepness of the parabola. They are all functions of the average luminance $L(n1,n2)$ in each block:

$$T_{\min}(n1,n2) = \begin{cases} (\dfrac{L(n1,n2)}{L_T})^{\alpha_T} \dfrac{L_T}{S_0}, & L(n1,n2) \le L_T \\ \dfrac{L(n1,n2)}{S_0}, & L(n1,n2) > L_T \end{cases} \quad (3.10)$$

$$f_{\min}(n1,n2) = \begin{cases} f_0 (\dfrac{L(n1,n2)}{L_f})^{\alpha_f}, & L(n1,n2) \le L_T \\ f_0, & L(n1,n2) > L_T \end{cases} \quad (3.11)$$

$$K(n1,n2) = \begin{cases} K_0 (\dfrac{L(n1,n2)}{L_K})^{\alpha_K}, & L(n1,n2) \le L_K \\ K_0, & L(n1,n2) > L_K \end{cases} \quad (3.12)$$

DCT basis functions $f_{i,j}$ are determined by two frequency components with the same spatial frequency but different orientations:

$$f_{i,j} = \frac{1}{2N_{DCT}} \sqrt{\frac{i^2}{\omega_x^2} + \frac{j^2}{\omega_y^2}} \tag{3.13}$$

θ denotes the angle between the two frequency components:

$$\theta_{i,j} = \arcsin \frac{2 f_{i,0} f_{0,j}}{f_{i,j}^2} \tag{3.14}$$

In Equation (3.9), the multiplicative factor $\frac{1}{r + (1-r)\cos^2 \theta_{i,j}}$ accounts for the effect of intermediate positions of the two Fourier components [61].

In [61], r = 0.7, N_{DCT}= 8, L_T = 13.45 cd/m^2, S_0 = 94.7, α_T = 0.649, f_0=6.78 cycles/deg, α_f = 0.649 and L_K = 300 cd/m^2. ω_x and ω_y are the horizontal width and vertical height of a pixel in degrees of visual angle, which can be determined as follows:

$$\omega = 2 \cdot \arctan(\frac{R}{2D}) \tag{3.15}$$

where R is the distance between two adjacent pixels, D is the viewing distance.

Considering JND is a property of the image and should use grey levels as the unit, we convert luminance threshold values $T_{i,j}(n1,n2)$ to the corresponding JND values as:

$$t_{DCT}(n1,n2,i,j) = \frac{MT_{i,j}(n1,n2)}{\alpha_i \alpha_j (L_{\max} - L_{\min})} \tag{3.16}$$

where M=256 is the number of gray levels for 8-bit image, $L_{\max}$ and $L_{\min}$ are the maximum and minimum display luminance. α_i and α_j account for the DCT coefficient factors, which is calculated as $\alpha_\tau = \frac{1}{\sqrt{N_{DCT}}} \begin{cases} 1, & \tau = 0 \\ \sqrt{2}, & \tau \neq 0 \end{cases}$.

3.3.2 Watson's DCTune Model [36]

Watson's DCTune Model aims at optimizing the quantization scheme for image coding

so that maximum visual quality can be achieved at a given bitrate. Luminance adaptation and contrast masking are considered for estimating the DCT-based JND values.

Compared with Ahumada and Peterson's model, Watson suggested a simpler solution to approximate the dependence of visibility threshold $t_{DCT}(n1,n2,i,j)$ on local image intensity:

$$t_{DCT}(n1,n2,i,j) = t_{ij}(\frac{C(n1,n2,0,0)}{\overline{C}})^{\alpha_T} \tag{3.17}$$

where $C(n1,n2,i,j)$ is the DCT efficient in block $(n1,n2)$, location (i,j) $(i,j = 0,1, \ldots 7)$. t_{ij} is predetermined based on the contrast sensitivity at an assumed display luminance L_0. $\overline{C}$ is a constant related to the display luminance. For a 25-graylevel image, $\overline{C}$ is set to be 128. α_T is 0.649.

Contrast masking, which accounts for the change of detection threshold for one signal at the presence of another, is simulated by a power law equation in the DCTune model. The final threshold considering both luminance adaptation and contrast masking is represented as:

$$\begin{aligned} m(n1,n2,i,j) &= \max[t_{DCT}(n1,n2,i,j), | C(n1,n2,i,j)^{\rho} t_{DCT}(n1,n2,i,j)^{(1-\rho)} |] \\ &= t_{DCT}(n1,n2,i,j) \cdot \max[1, | \frac{C(n1,n2,i,j)}{t_{DCT}(n1,n2,i,j)} |^{\rho}] \end{aligned} \tag{3.18}$$

where $\rho = 0.7$.

3.4 Human Vision Models for video

Although JND models for images are developing very fast, there are very few explicit JND model designed for videos. The majority of the human vision models for videos are devoted to perceptual video quality evaluation, which requires two input video sequences (reference and target video sequence). Basically, a large number of models for this use adopt the multi-channel scheme. The heart of this multi-channel theory lies in that stimulus are decomposed to be processed in different channels when they go into the HVS. The CSF is just the envelope of the responses of these channels. The emergence of this multi-channel structure for the HVS model was prompted by the neuronal selectivity in vision science. Electrophysiological experiments reveal that the perceptive neurons in the primary visual cortex are selectively sensitive to certain types of information. A particular cell may strongly respond to a signal of certain orientation and frequency.

Generally, in a multi-channel model for quality evaluation, the signal is first engaged in multi-channel decomposition, then luminance adaptation, inter- & intra-channel contrast masking follow, finally pooling is implemented to integrate the separate elements that contribute to the overall quality.

The most well-known DCT-based HVS models for video is the Digital Video Quality (DVQ) metric proposed by Watson etc. in 2001 [39]. Watson developed the DVQ model based on his DCT-based still image metric. The model accepts two video

sequences (one reference sequence and one test sequence), and finally gives a measure for visible difference between them. A pyramid structure for frequency and orientation analysis has been adopted by several well-developed perceptual quality metric models [23, 28, 54, 55]. The Sarnoff JND model [55] requires two input sequences and outputs a JND map sequence to indicate the visible difference between them. Then two temporal filters and Gaussian pyramid structure are used for spatio-temporal decomposition. Normalization is used to adapt the sensitivity to the temporal variation of the luminance. Winkler [28] developed his PDM model for both digital color images and videos following the earlier work by van den Branden Lambrecht [54]. It incorporates the combined effects of color perception, the multi-channel architecture of temporal and spatial mechanism, spatio-temporal contrast sensitivity, pattern masking and channel interactions. CVQE model by Masry and Hemami [23] evaluates the continuous video quality at low bit rates. Instead of producing a single pooled value for a video sequence as most models do [28, 39, 54, 55], this metric can provide a time-varying quality assessment.

3.4.1 Chou and Chen's JND model (1996) [1]

Chou and Chen built their spatio-temporal JND model as an extension of the earlier spatial JND model and the model is specifically designed for video coding. The temporal perceptual redundancy is simply modeled as an elevation factor that is a function of the interframe luminance difference. The spatio-temporal JND value $JND_{S\text{-}T}(x,y,n)$ at location (x,y) of frame n can be computed as:

$$JND_{S\text{-}T}(x,y,n) = f_3(ild(x,y,n)) \cdot JND_S(x,y,n) \tag{3.19}$$

Where $JND_S(x,y,n)$ represents the corresponding spatial JND value and $ild(x,y,n)$ represents the average interframe luminance difference between the nth and the (n-1)th frame. f_3 is empirically determined as in Figure 3.6.

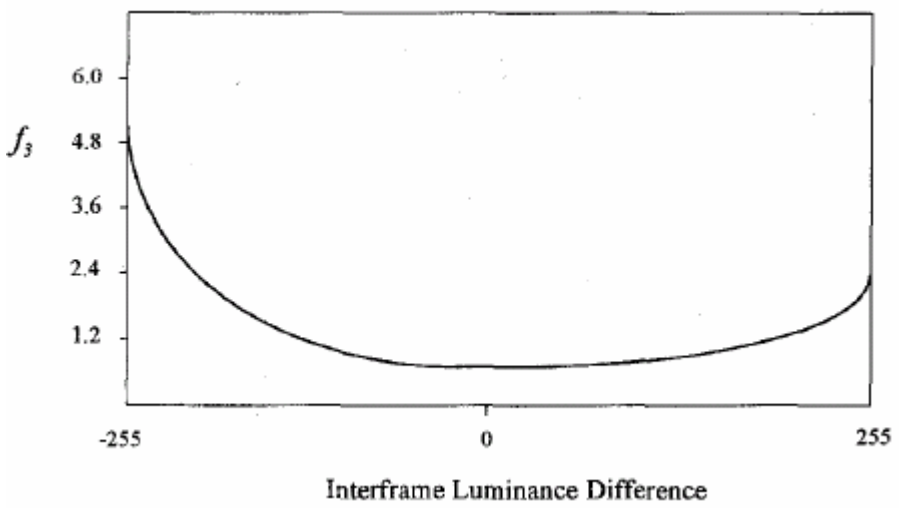


Figure 3.6 Scale factor as a function of the interframe luminance difference for modeling temporal redundancy [1]

The calculated spatio-temporal JND profile was then embedded in the video codec for perceptual bit allocation. After the target sequence is decomposed into 11 spatio-temporal subbands by a two-band temporal filter and QMF filterbanks, the perceptual weighting for each subband is determined by the spatio-temporal CSF presented by Kelly [27]. Therefore, the full-band JND defined by equation (3.19) can be distributed into subband JND profiles according to the perceptual weightings. Optimized bit-allocation and error concealment can be realized based on the perceptual importance determined by the subband JND.

3.5 Summary

In this chapter, various HVS models contributing to JND estimation in image/video have been presented. Firstly, two stand-alone schemes for modeling the spatio-temporal CSF have been discussed. Based on the human vision properties discussed in Chapter 2, two DCT-based JND models for images have been summarized. Because the majority of the human vision models for videos are applied to quality evaluation, we introduced several multi-channel models designed for this application. Finally we elaborated one spatio-temporal JND model applied for video coding.

However, as discussed in Chapter 1, most of the models (except Chou and Chen's model in 3.4.1) haven't calculated explicit JND profiles and all the models for perceptual quality evaluation discussed in Section 3.4 need both original sequence and distorted sequence for processing. Although Chou and Chen proposed a method to calculate JNDs for the video, they used a simplified equation which is lack of theoretical proof. We are going to tackle these problems in the proposed model in the next two chapters.

CHAPTER 4

DCT-based Spatio-temporal JND Model

4.1 Introduction

In this chapter, we are going to develop a spatio-temporal HVS model estimating JNDs in the DCT domain. Besides spatial frequencies, the influences of temporal frequencies and eye movements on contrast sensitivity are explored. We also incorporate luminance adaptation and an improved contrast masking estimator to make the proposed model more consistent with human perception. The same formulation in the proposed spatio-temporal model is capable of yielding JNDs for both still images and video with significant motion. The experiments (Chapter 5) conducted in this study will demonstrate that the JND values estimated with moving objects by the model are in line with the HVS perception.

We derive the JND profile in the DCT domain because the DCT is the most adopted transform for image and video compression. Therefore, our model is convenient to be applied for the existing standard coding systems such as MPEG-1/2/4 and H.26x. Figure 4.1 gives a block diagram for the model.

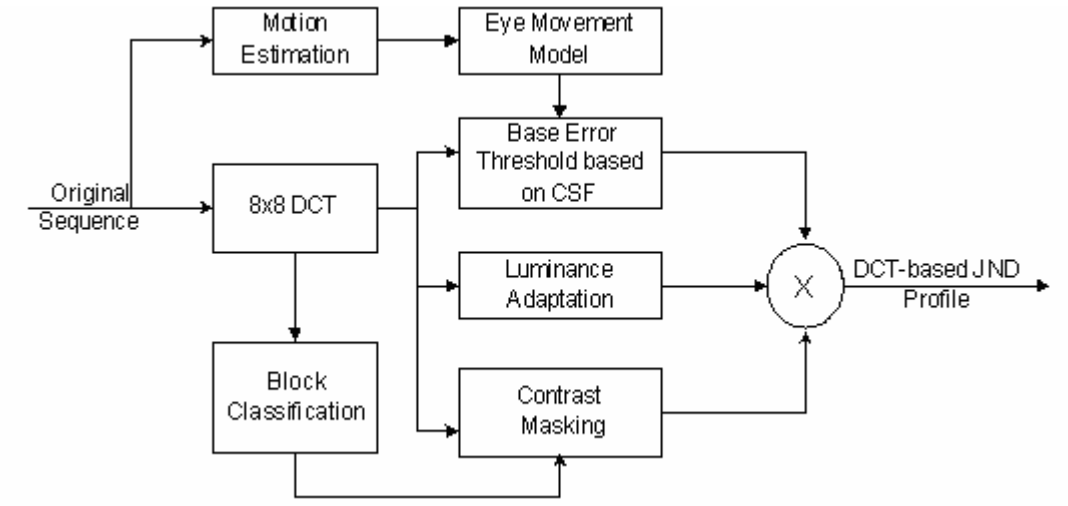


Figure 4.1 Block diagram for the proposed JND model

Within a DCT block (n,t), the JND threshold in the position (i,j) $(i, j = 0, 1, ..., 7)$ of the frame t can be calculated as:

$$JND\ (n,i,j,t) = T(n,i,j,t) \cdot a_{Lum}(n,t) \cdot a_C(n,i,j,t) \qquad (4.1)$$

where T is the base distortion threshold contributed by spatio-temporal CSF, a_{Lum} and a_C denote the respective effects of luminance adaptation and contrast masking. In the following part of this chapter, we will detail the calculation of each multiplicative factor.

4. 2 Base distortion Threshold in DCT Subbands

In this section, we compute the maximum error that can be tolerated for each DCT coefficient with consideration of spatio-temporal CSF (i.e., the relation of visibility threshold with spatial and temporal frequencies of visual signals).

4.2.1 Spatio-temporal CSF in DCT domain

As discussed in section 3.2.3, equations (3.4) – (3.6) describe the CSF. c_0 and c_1 control the magnitude and the bandwidth of a CSF curve (Figure 3.5), and can be decided with subjective viewing test data for the subband decomposition in use. If we consider the (*i,j*)-th subband in the *n*-th DCT block in the *t*-th frame, the corresponding CSF can be derived from (3.9) as follows:

$$G(n,i,j,t) = c_0 \cdot (k_1 + k_2 \,|\log(\varepsilon \cdot v(n,t)/3)|^3) \cdot v(n,t) \cdot [2\pi\rho_{i,j}]^2 \cdot \exp(-2\pi\rho_{i,j} \cdot c_1 \cdot (\varepsilon \cdot v(n,t) + 2)/k_3)$$

(4.2)

where $i, j = 0,1,\ldots,N\text{-}1$ while N is the dimension of the DCT block; *v(n,t)* depicts the associated retinal image velocity. The empirical constants k_1, k_2 and k_3 are set as 6.1, 7.3 and 23 [27], respectively, and ε = 1.7; $\rho_{i,j}$ (cycles per degree) is the spatial subband frequency [61]:

$$\rho_{i,j} = \frac{1}{2N}\sqrt{(i/\omega_x)^2 + (j/\omega_y)^2} \tag{4.3}$$

where ω_x and ω_y are the horizontal and vertical size of a pixel in degrees of visual angle, respectively. They are related to the viewing distance ℓ and the display width Λ of a pixel on the monitor as follows:

$$\omega_\hbar = 2 \cdot \arctan\left(\frac{\Lambda_\hbar}{2 \cdot \ell}\right),\ \hbar = x, y \tag{4.4}$$

In our experimental environment, ω_x and ω_y are equal to 0.0342 (degree) for ℓ=50cm and the CRT monitor (Philips 107X_2) used in the experiments.

4.2.2 Eye Movement Effect

For a practical formulation of the spatio-temporal CSF, the influence of observers' eye movement needs to be taken into account for motion imagery [41,43], and the retinal velocity in (4.2), $v(n,t)$, has to be expressed by easily measurable variables from the moving images. Because of the eye movements, the perceived velocity at the retina (retinal velocity) is different from the image plane velocity that can be usually obtained through motion estimation.

Based on the heuristic eye movement model devised by Daly (1998) [10] (Section 3.2.3), the retinal image velocity can be expressed as:

$$v(n,t) = v_I(n,t) - \min\lfloor (g \cdot v_I(n,t)) + v_{MIN}, v_{MAX} \rfloor \qquad (4.5)$$

where the definitions of $v_I(n,t)$, g, v_{MIN} and v_{MAX} can be referred to Section 3.2.3. In this study, it is assumed that the HVS tracks different parts of an image equally and g is assigned with a value of 0.92.

With a motion estimation technique (discussed in 4.2.5), we can obtain ($MV_x(n,t)$, $MV_y(n,t)$), the motion vector for the n-th block in the t-th frame, and calculate the image velocity as:

$$v_I(n,t) = f \cdot \sqrt{(MV_x(n,t) \cdot \omega_x)^2 + (MV_y(n,t) \cdot \omega_y)^2} \qquad (4.6)$$

where f is the frame rate (in *frames per second*) of the video, and $v_I(n,t)$ is measured in *degrees of visual angle per second*.

When $v(n,t) \equiv 0.15$ deg/sec (i.e., only natural drift movement occurs), the spatio-temporal CSF is equivalent to the spatial (static) CSF [10], so the formulae derived above are also applicable for images if the said value of $v(n,t)$ is used.

4.2.3 Base Distortion Threshold

When equation (4.2) is used for predicting distortion threshold due to spatio-temporal CSF, several factors need to be considered: i) the sensitivity modeled by (4.2) represents the inverse of distortion threshold; ii) the CSF-derived threshold expressed in luminance needs to be scaled to grey levels for digital images [61]; iii) since Equation (4.2) is derived from the experimental data of one-dimensional spatial frequencies (i.e., i or j is equal to 0), for an arbitrary subband ($i \neq 0$ and $j \neq 0$), the threshold is actually higher than the one given by (4.2), and therefore a compensating term (as the last term of the right-hand side in (4.7) below) needs to be introduced [61] for a DCT subband. With all considerations mentioned above, the base threshold for a DCT subband is determined as:

$$T(n,i,j,t) = \frac{1}{G(n,i,j,t)} \cdot \frac{M}{\phi_i \phi_j (L_{\max} - L_{\min})} \cdot \frac{1}{r + (1-r) \cdot \cos^2 \theta_{i,j}} \tag{4.7}$$

where $L_{\max}$ and $L_{\min}$ represent the display luminance values corresponding to the maximum and minimum grey levels (depending on the display facility); M is the number of grey levels (256 in most imagery systems); ϕ_i and ϕ_j are DCT normalization factors:

$$\phi_u = \begin{cases} \sqrt{\frac{1}{N}} & u = 0 \\ \sqrt{\frac{2}{N}} & u \neq 0 \end{cases} \qquad u = 0,1,\ldots, N\text{-}1 \tag{4.8}$$

r is set to 0.6 [61], and $\theta_{i,j}$ accounts for the effect of an arbitrary subband (the factor iii mentioned above):

$$\theta_{i,j} = \arcsin \frac{2\rho_{i,0}\rho_{0,j}}{\rho_{i,j}^{\ 2}} \tag{4.9}$$

When i=0 or j=0, $\theta_{i,j}$ =0 and $T(n,i,j,t)$ takes its smallest value since $r+(1-r)\cos^2\theta_{i,j}$=1; when i=j, $\theta_{i,j}$=90^o and $T(n,i,j,t)$ takes its biggest value since $r+(1-r)\cos^2\theta_{i,j} = r$.

4.2.4 Determination of c_0 and c_1

In equation (4.2), c_0 and c_1 regulate the value level and the bandwidth of the CSF respectively. As illustrated in [10], the spatio-temporal CSF when v = 0.15 deg/sec (i.e., only natural drift eye movement occurs) is equivalent to the spatial (static) CSF. With $v(n,t)$ = 0.15 deg/sec in (4.7), we estimated c_0 and c_1 by fitting the resultant T(i,j) to the spatial CSF model in [61]. .

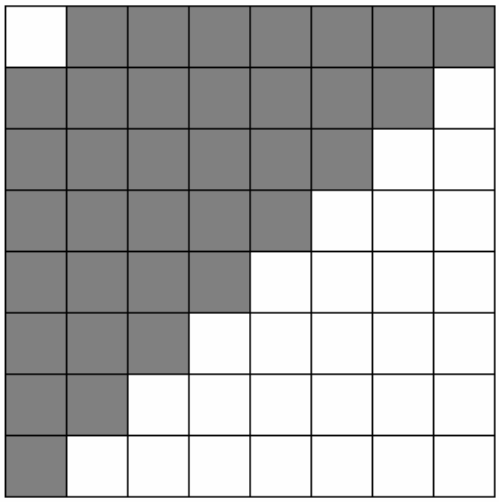


Figure 4.2 Illustration of the fitting data (gray blocks shows the data for fitting)

Only the AC components of the upper triangle in the DCT block are used for fitting because of the symmetry of the spatial frequencies expressed as (4.3) (Figure 4.2). The visibility thresholds t_s are calculated for each DCT frequency ρ_s using the JND model in [61]. We use the LMS (least mean squares) approach to get the best-fitted c_0 and c_1 for the data pair (ρ_s, t_s):

$$(c_0, c_1) = \arg\min\{\sum_s [t_s - T(\rho_s, c_0, c_1)]^2\} \tag{4.10}$$

where $T(\rho_s, c_0, c_1)$ is the threshold (derived from equation (4.7)). As a result, $c_0 = 7.126$ and $c_1 = 0.565$. Figure 4.3 shows the result of the fitted data.

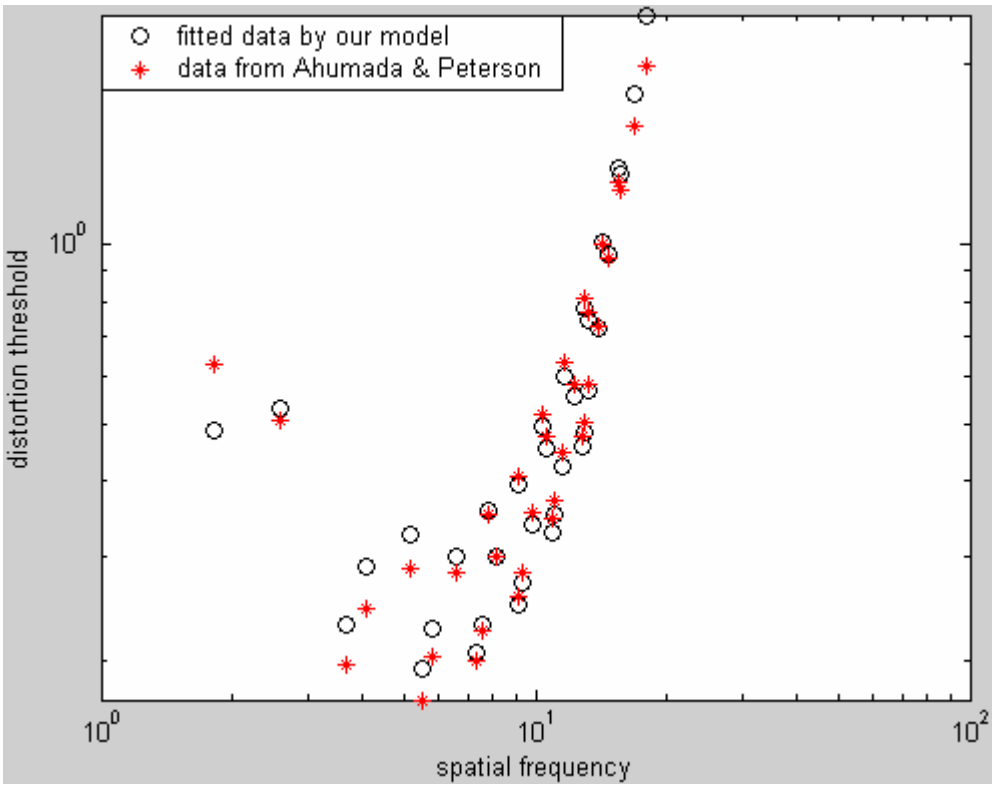


Figure 4.3 Data-fitting results from LMS

4.2.5 Motion Estimation

In order to calculate the image plane velocity as in (4.6), we need to obtain the motion vector ($MV_x(n,t)$, $MV_y(n,t)$) according to the motion estimation. Here we adopt a new three-step search algorithm (NTSS) for block motion estimation [5] to achieve the target.

NTSS has been developed from the three-step search (TSS), which is widely used as the motion estimation method in some low bit-rate video compression applications. Compared with the oldest and most reliable full search method, NTSS and TSS are superior in their simplicity and efficiency yet effectiveness. Retaining the advantages

of TSS, NTSS has more competitive features:

1. NTSS chooses a set of center-biased checking points in addition to the original checking points as in TSS in its first step, which makes the search more consistent with the motion distribution of real world image sequence (Figure 4.4).

2. NTSS has a halfway-stop policy which further reduces the computation cost.

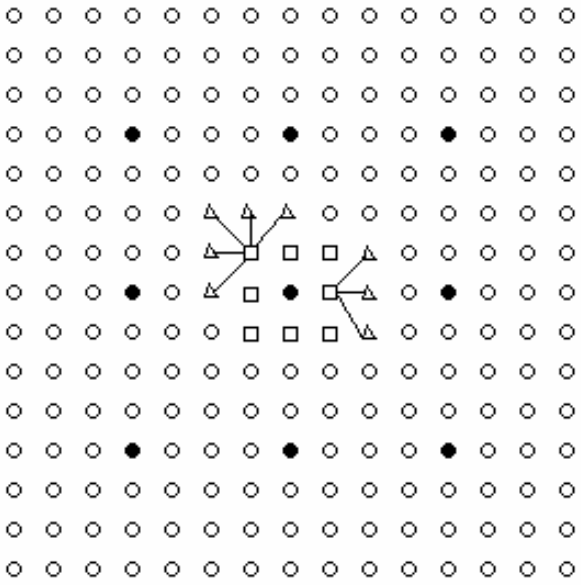


Figure 4.4 Illustration for NTSS (Filled circles are the checking points in the first step of TSS, squares are the 8 extra points added in the first step of NTSS, and triangles explain how the second step search is performed if the minimum BDM in the first step is at one of the 8 neighbors of the window center) [5]

As in Figure 4.4 (which assumes that the search window is a 15×15 pixel block), we choose 17 check points in the first step: 9 points as in TSS (filled circles) and 8 neighbor points of the center (squares). We find the point that yields the minimum

block distortion measure (BDM). If the point is the center of the search window, we can stop the search and obtain the motion vector (0, 0). If it is not, then we judge whether the point is the one of the neighbor of the center (squared points). If the answer is yes, we perform the second step of NTSS: the check points are limited to 3 or 5 neighbors of the point (illustrated in Figure 4.4 by triangles). The whole algorithm will stop at the second search. If the stand-out point in the first step is neither the center nor one of the neighbors of the center, then we need to do a complete three-step search.

4.3 Luminance Adaptation and Contrast Masking

For a complete spatio-temporal CSF model, we need to include luminance adaptation; and as a comprehensive JND estimator, different contrast masking has to be considered. For luminance adaptation, the response of the HVS depends more on the luminance against the surroundings than the absolute luminance. Intra-band masking refers to the reduction in the distortion visibility in a subband due to the signal in that subband itself, while inter-band masking refers to the reduction in the distortion visibility in a subband at the presence of other visual components in other subbands. In this work, inter-band masking is evaluated on a block basis.

The effects of luminance adaptation, intra-band masking and inter-band masking can be formulated as the multiplicative factors to the base distortion threshold estimated in the previous section; in consequence, the complete JND estimator is modeled by

formula (4.1).

4.3.1 Luminance Adaptation

As has been discussed in Section 2.3, the visual sensitivity varies with the background luminance. Since the digital images have a limited intensity range (typically 256 levels), the mean background luminance is not very different from one image to another. However, previous studies found that local intensity variations contribute to the adaptive thresholds within an image. For digital images, the HVS visibility threshold is higher in dark and light regions, i.e., the HVS is more sensitive to the noise in medium grey-grey regions, as shown in Figure 4.5.

Since the average local intensity can be represented by its DC component, $C(n,0,0,t)$, the luminance adaptation factor for the n-th DCT block is determined as [35]:

$$a_{lum}(n,t)=\begin{cases} k_1(1-\dfrac{2\cdot C(n,0,0,t)}{M\cdot N})^{\lambda_1}+1 & if \quad C(n,0,0,t)\le \dfrac{M\cdot N}{2} \\ k_2(\dfrac{2\cdot C(n,0,0,t)}{M\cdot N}-1)^{\lambda_2}+1 & otherwise \end{cases} \tag{4.11}$$

where $M = 256$ represents the range of grey levels, $N = 8$. According to the background illumination and the effect of γ-correction [2]: $k1 = 2$, $k2 = 0.8$, $\lambda_1 = 3$ and $\lambda_2 = 2$ [35].

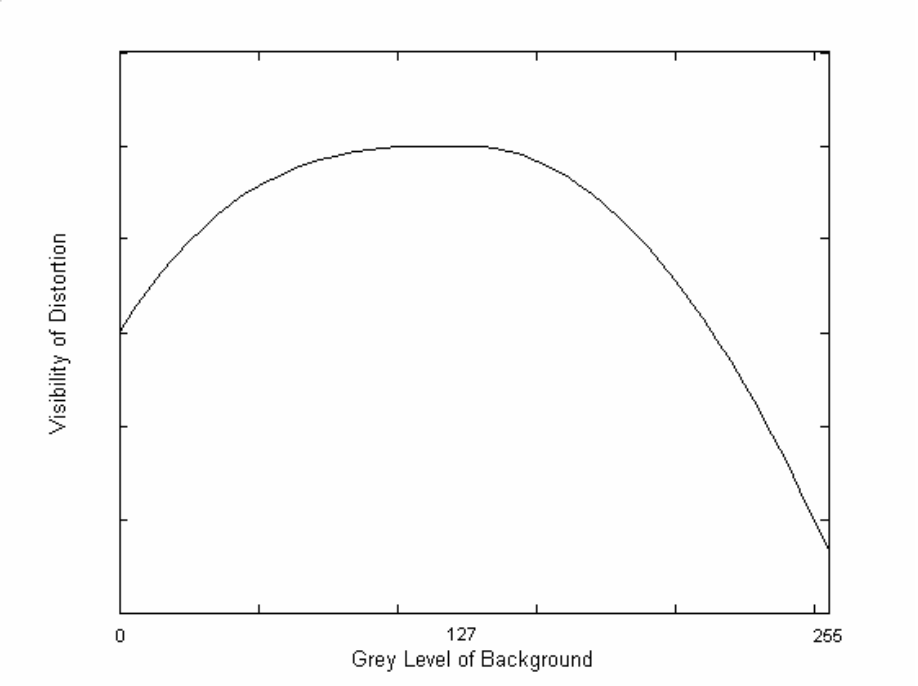


Figure 4.5 Distortion visibility as a function of background brightness [6]

4.3.2 Intra- and Inter-band Contrast Masking

The extent of contrast masking depends on the local intensity activity of the image. The HVS sensitivity to error is generally higher in smooth (or plain) region, and lower in texture region [34]; while the sensitivity for edge region lies in between. Therefore, contrast masking can be discriminated for different image context.

In computing the contrast masking adjustment, we separately estimate the intraband masking a_{intra} and interband masking a_{inter}. The combined contrast masking factor can be calculated as:

$$a_C(n1, n2, i, j, t) = a_{\text{intra}}(n1, n2, i, j, t) \cdot a_{\text{inter}}(n1, n2, t) \qquad (4.12)$$

In block-based DCT environment, local texture activity is approximately reflected by

the AC energy of the DCT coefficients [34,66]. A DCT block can be divided into four parts: DC, low-frequency (LF), medium-frequency (MF) and high-frequency (HF), as shown in Figure 4.6.

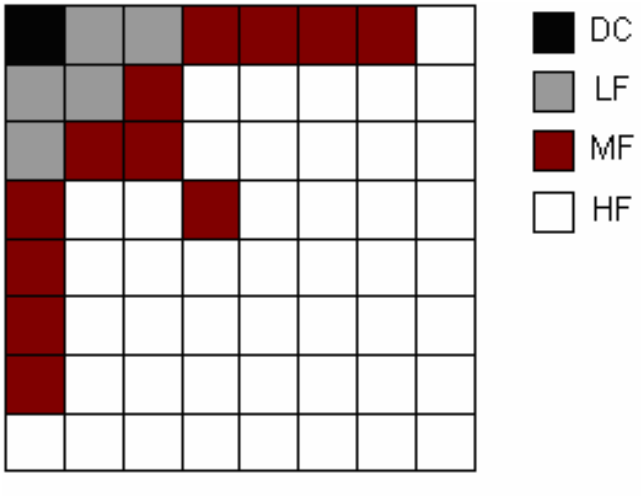


Figure 4.6 Block classification scheme for a DCT block [34,35]

If we denote the sums of the absolute DCT coefficient values in LF, MF and HF by L, M and H, then the block texture energy $TexE$ is calculated as:

$$TexE = M + H \tag{4.13}$$

Moreover, we define:

$$E_1 = (\bar{L} + \bar{M}) / \bar{H} \tag{4.14}$$

$$E_2 = \bar{L} / \bar{M} \tag{4.15}$$

where $\bar{X}$ $(X = L, M, H)$ denotes the mean of X.

Block classification is implemented according to the following rules [34]:

1. A block is assigned to be an *EDGE* class if any one of the following cases is true:

Case 1: $E_1 \geq \upsilon$

Case 2: $\max\{E_1, E_2\} \geq \alpha$ & $\min\{E_1, E_2\} \geq \beta$ where $\alpha > \beta$

Case 3: $\mu_1 \leq TexE \leq \mu_3$ and (4.14) or (4.15) is valid for $\alpha = \alpha_1$ and $\beta = \beta_1$

Case 4: $TexE > \mu_3$ and (4.14) or (4.15) is valid for $\alpha = k \cdot \alpha_1$ and $\beta = k \cdot \beta_1$

2. A block is assigned to be a *TEXTURE* class if any one of the following cases is true:

Case 1: $\mu_2 \leq TexE \leq \mu_3$ and neither (4.14) nor (4.15) is valid for $\alpha = \alpha_1$ and $\beta = \beta_1$

Case 2: $TexE > \mu_3$ and neither (4.14) nor (4.15) is valid for $\alpha = k \cdot \alpha_1$ and $\beta = k \cdot \beta_1$

3. A block is assigned to be a *PLAIN* class if any one of the following cases is true:

Case 1: $TexE \leq \mu_1$

Case 2: $\mu_1 \leq TexE \leq \mu_2$ and neither (4.14) nor (4.15) is valid for $\alpha = \alpha_1$ and $\beta = \beta_1$

Then inter-band masking can be represented as:

$$a_{inter}(n,t) = \begin{cases} 1 + [(TexE(n,t) - \mu_2)/(2\mu_3 - \mu_2)] \cdot \delta_1 & for\ TEXTURE\ block \\ \delta_1 & for\ EDGE\ block\ and\ L+M > 400 \\ \delta_2 & for\ EDGE\ block\ and\ L+M \leq 400 \\ 1 & for\ PLAIN\ block \end{cases} \quad (4.16)$$

In the above rule and equation (4.16), we set μ_1=125, μ_2 =290, μ_3 =900, α_1 = 7, β_1 = 5, κ = 0.1, υ = 16, δ_1 = 1.25 and δ_2=1.125.

As for intra-band masking, we modify the model in [34] to exclude intra-band masking from taking effect in LF and MF subbands for non-textured blocks:

$$a_{intra}(n,i,j,t) = \begin{cases} 1 & for\ (i,j) \in LF \cup MF\ in\ PLAIN\ \&\ EDGE\ block \\ \max\left\{1, \left[\frac{C(n,i,j,t)}{T(n,i,j,t) \cdot a_{Lum}(n,t)}\right]^{\varepsilon}\right\} & otherwise \end{cases} \quad (4.17)$$

where $C(n,i,j,t)$ is the DCT coefficient, and ε = 0.36.

4.4 Summary

In this chapter, we have developed a HVS model for estimating the JND profile in DCT domain. This model is applicable for both images and video sequences with the differentiation in the spatio-temporal CSF component. For image application, the retina velocity is set to a constant value; while for video application, the retina velocity is determined using motion estimation. An eye movement model is incorporated into the CSF to account for the effect of eye motion on visual sensitivity. Moreover, luminance adaptation and contrast masking are considered as elevation factors for the base threshold derived from spatio-temporal CSF. For contrast masking, in addition to the EDGE region, we also exclude the intra-band masking at the low and medium frequencies of the SMOOTH region, which is more consistent with the human perception. The experiments and model validation will be discussed in Chapter 5.

CHAPTER

5

Experiments and Model Testing

5.1 Introduction

To evaluate the performance of JND models, the generated JND profiles can be used to guide the noise injection into an image or video [35,46], and appropriate subjective viewing tests are then conducted to assess the quality of the resultant visual signal. The proposed JND estimator has been compared with Zhang, et. al.'s JND model [35] (referred as Model I hereinafter) and Daly's model[1] [10] (referred as Model II hereinafter) in the experiments. These two models are the most relevant existing models for comparison, since Model I is the recent enhancement of the well acknowledged models [36,61] in DCT subbands, while Model II is an enhanced version of [27]. The aim of our experiments is to show that the proposed model not only succeeds in the improvement of Zhang's model for spatial properties of HVS, more importantly, it also achieves an effective addition of temporal properties into

[1] with c_0 and c_1 determined in Section 4.2.4, because Daly's original model is not for DCT domain.

JND modeling.

The relevant parameters of the models are determined under the same experimental conditions: a CRT monitor (17" Philips $107X_2$) with viewing distance of 50 cm. When a model (Model II or the proposed model) is used for images in the following experiments, $v(n,t)$ is set to 0.15 deg/sec.

The noise injection can be described as:

$$C_{noise}(n,i,j,t) = C(n,i,j,t) + M_{n,i,j}^{random} \bullet J(n,i,j,t) \qquad (5.1)$$

where $C(n,i,j,t)$ is the DCT coefficient of the original image or a frame of the original video, and $C_{noise}(n,i,j,t)$ is the corresponding noise-injected coefficient; $M_{n,i,j}^{random}$ takes +1 or -1 randomly; and $J(n,i,j,t)$ represents the JND obtained via a model (Model I, Model II or the proposed model). A JND model can avoid yielding values larger than the actual HVS thresholds via adjusting its overall gain. If a model derives better JNDs, Equation (5.1) allows higher injected-noise energy (measured by PSNR) without jeopardizing picture quality.

5.2 Subjective testing

The aim of vision research is to mimic the response of HVS towards stimulus. Our perceptual study applies the results from vision research and forms objective metrics to substitute for the subjective tests. Although subjective tests are unfavorable because of their complex and time-consuming structure, they are the most reliable strategies to

predict the perception after all. Therefore, in order to evaluate the performance of our model, we need to conduct subjective tests to compare the objective estimation and the actual subjective perception. Individual subjective testing schemes in our experiments will be elaborated for image and video in the following sections.

The subjective viewing tests in this project are conducted in a room illuminated by fluorescent ceiling lights, and this is the typical conditions under which people would view digital images. The subjects' eyesight is either normal or has been corrected to be normal with spectacles.

5.3 Results and Discussions

5.3.1 Evaluation on images

For images, the proposed model is similar to Model I, and differs from Model II in that the latter does not exploit the content-based properties of an image (i.e., luminance adaptation and contrast masking).

As discussed in Section 5.1, we inject noise into the image based on different JND models to compare their performances of JND estimation. According to Equation (5.1), the amount of noise injected into the image is at the JND level. Figure 5.1 shows the noise-contaminated *Lena* images by Model I, Model II and the proposed model. We find that the noise injected in the images based on all three models is hardly noticeable, which proves that the three models yield values not larger than the actual HVS

thresholds. Comparing the PSNRs of the three images in Figure 5.1, the proposed model yields the lowest PSNR (about 2 dB lower than Model II) by inserting more noise into the image, and this reflects that the proposed model is able to exploit the HVS bounds more aggressively without introduction of noticeable visual disturbance.

(a) Model I, PSNR: 31.28 dB

(b) Model II, PSNR: 33.08 dB

(c) the proposed JND model, PSNR: 31.09 dB

Figure 5.1. Noise-injected *Lena* with Model I, Model II and the proposed JND model.

For a comprehensive evaluation, we have conducted subjective viewing tests to give quantitative scores for all images (with different visual contents and spatial complexity) shown in Figure 5.2 with JND-guided noise injection. In each viewing test, two images of the same scene (i.e., the original image and its noise-contaminated version) were juxtaposed on the screen, and six subjects (three are in the image processing field and three are naive) were employed. On each session of the experiment, subjects viewed two images, and were then given time to vote on the comparative quality of the images, using the continuous quality comparison scale shown in Table 5.1. The subjects were not allowed to respond until they had viewed the images for at least two seconds. The order of the presentation of the image pairs was randomized in each session, and the noised image appeared randomly on the left- or right-hand side of the screen. The results indicated no notable difference on whether a subject has image processing knowledge or not.

Table 5.1 Subjective rating criterion for the comparative visual quality of an image pair

Subjective score	Description
-3	the right one 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

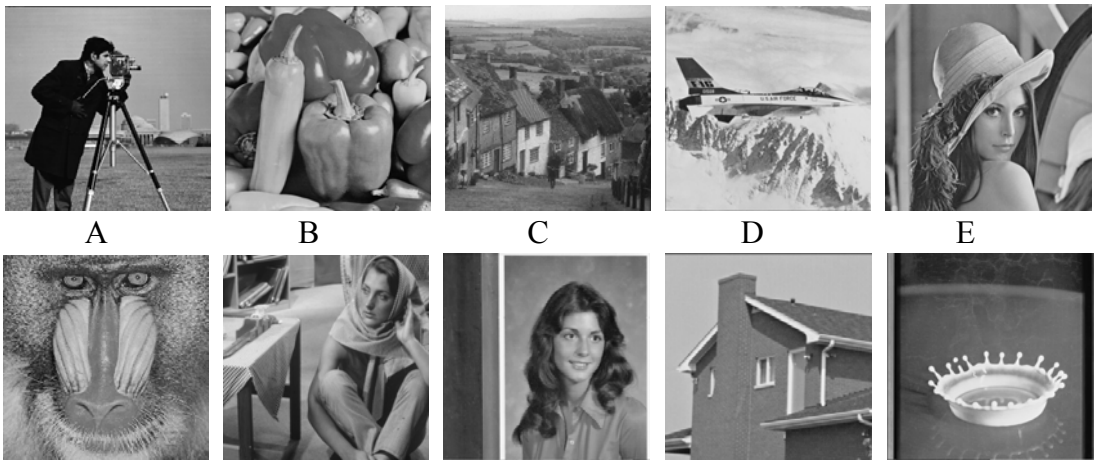


Figure 5.2 Images for the experiments. (A: Cameraman256x256; B: Pepper512x512; C: Goldhill256x256; D: Airplane512x512; E: Lena512x512; F: Baboon256x256; G.: Barbara512x512; H: Girl256x256; I: House512x512; J: Splash512x512.)

Figure 5.3 shows the mean subjective score by all subjects for each noise-injected image with one of the JND models, while Table 5.2 lists the corresponding standard deviations. In Figure 5.3, each symbol represents the average subjective score among six subjects for one image. A negative subjective score indicates that the noise-injected image has worse perceptual quality than the original image, and its magnitude represents the extent of quality degradation. Averaging the mean subjective scores of all the ten images, we get the respective scores for Model I, Model II and the proposed model. The slightly higher average quality score (in difference of 0.33 in Fig. 6) with Model II than Model I is due to the somewhat excessive intra-band masking in LF and MF subbands for non-textured blocks in Model I, the proposed model has remedied this by excluding intra-band masking from taking effect in LF and MF subbands for non-textured blocks (Section 4.3.2, Equation (4.17)). As can be observed in Figure 5.3,

the proposed model leads to similar visual quality on average as Model II in the noise-injected images, and slightly better quality on average than Model I. Overall, the 3 scores are very close to each other (the average range of subjective score variations is below 0.45) and are all close to score 0, where there is no difference between the original image and its noise-contaminated version. Therefore, noise injection into images guided by all the three models leads to very similar visual qualities and the noise-contaminated image can hardly be distinguished from its original.

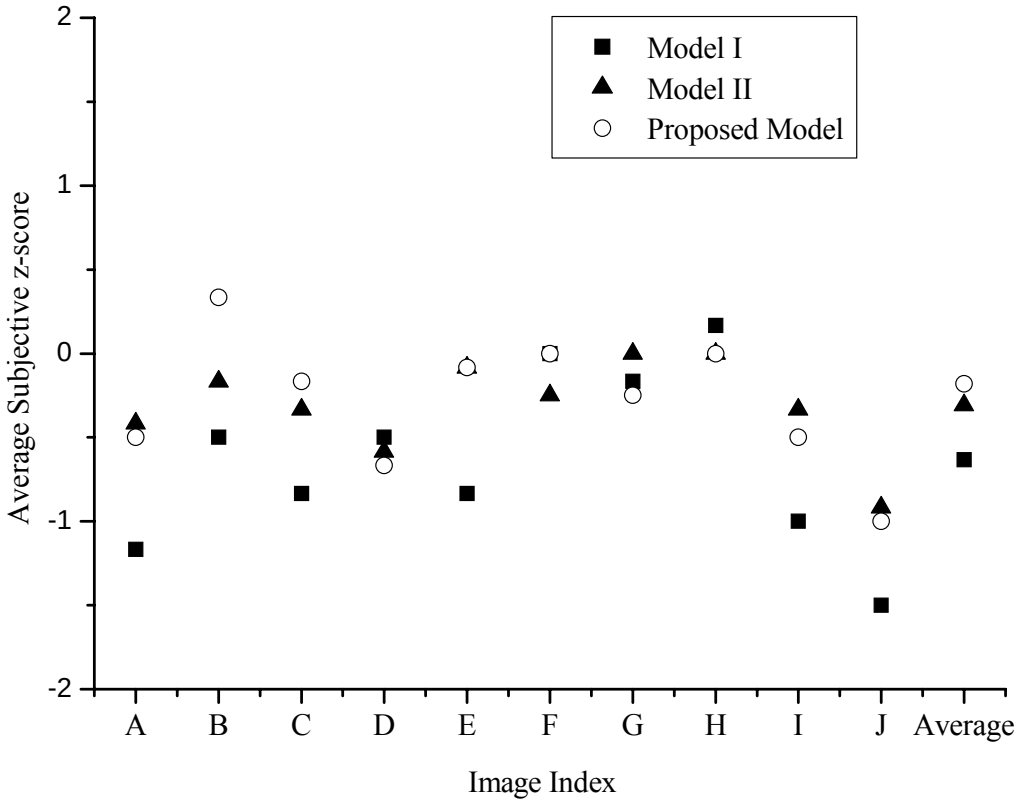


Figure 5.3. Mean subjective scores for the noise-injected images with the three JND models

Table 5.2 Standard deviations of the subjective scores

for the noise-injected images

Image	A	B	C	D	E	F	G	H	I	J
Model I	0.258	0.548	0.408	0.548	0.408	0	0.408	0.408	0.548	0.447
Model II	0.492	0.408	0.516	0.492	0.204	0.418	0	0	0.408	0.204
Proposed Model	0.548	0.516	0.408	0.516	0.492	0	0.418	0	0.548	0

The PSNR has been used to measure the visual content variations after the JND-guided noise injection. At a same level of perceived picture quality, a better JND model yields more aggressive JNDs (i.e., resulting in lower PSNR). Figure 5.4 shows the PSNRs of the noise-injected images using the three models. The PSNRs (as well as the average PSNR) for the images with the proposed model are similar to those with Model I. As aforementioned, Model II for images is image independent, and therefore the associated PSNR remains the same for every image in Figure 5.4. The propose model yields more aggressive JNDs than Model II towards the actual HVS thresholds, with the evidence of an average PSNR reduction of 2.82 dB from Model II (as shown in Figure 5.4 for all ten images), without jeopardy of visual quality (Figure 5.3).

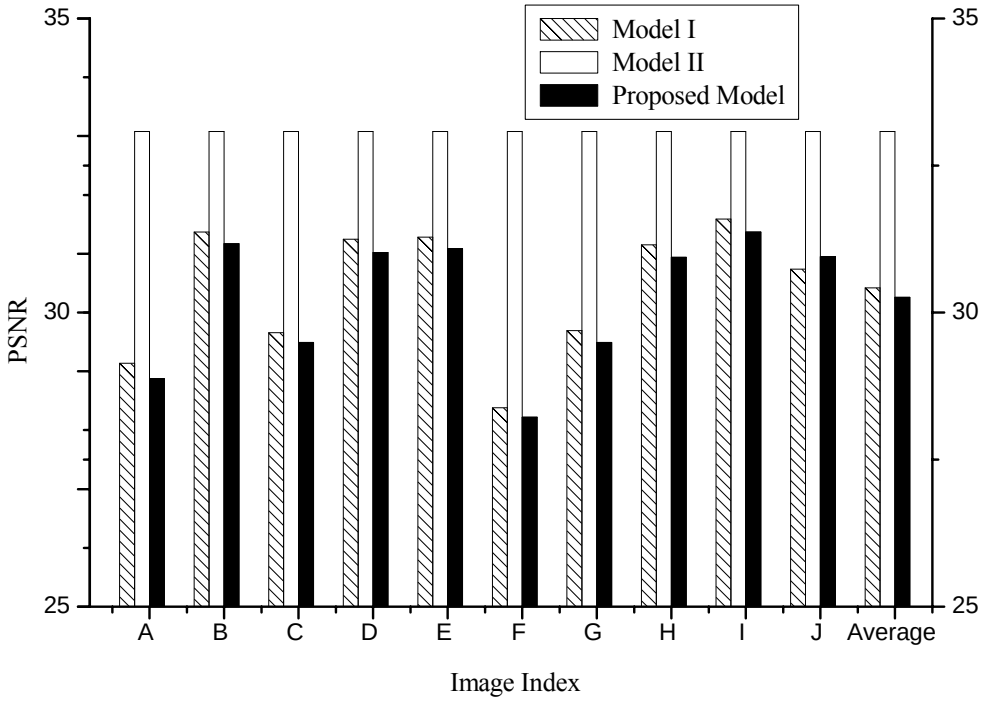
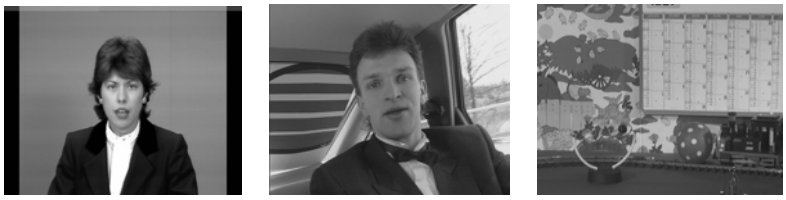


Figure 5.4 PSNRs of noise-injected images by the three models

5.3.2 Evaluation on video

We have conducted experiments to test how the models work for video sequences. Seven sequences (*Susie, Miss America, News, Bus, Claire, Carphone,* and *Caltrain,* as shown in Figure 5.5) are distorted by inserting the noise according to (5.1), for the experiments. The proposed model should outperform the other two models in video, since Model I does not exploit temporal CSF while Model II does not exploit the content-based properties of an image.

(i) (ii) (iii) (iv)



(v) (vi) (vii)

Figure 5.5 Videos for the experiments: (i) *Susie;* (ii) *Miss America.cif*; (iii) *News.qcif;* (iv) *Bus.cif;* (v) *Claire.cif;* (vi) *Carphone.qcif* ; (vii) *Caltrain.cif.*

1) Effect of Motion

Figure 5.6 illustrates how the models perform in the presence of motion, for the noise-injected *Susie* and *Miss America* sequences. First we measure the motion of each frame using the *average motion energy* which is defined as:

$$Avg\_MotionEnergy(t) = \frac{1}{N_f}\sum_n (MV_x^2(n,t) + MV_y^2(n,t)) \qquad (5.2)$$

where ($MV_x(n,t)$, $MV_y(n,t)$) represents the motion vector for the n-th block in the t-th frame, and N_f is the number of blocks in a frame.

In Figure 5.6, (a) and (b) denote the *average motion energy* over the frames for the noise-injected *Susie* and *Miss America* video sequences, with the three JND models, while (e) and (f) illustrate the zoom-in details of (a) and (b) before Frame 40. Figure 5.6 (c) and (d) show the PSNR over the frames for the two noise-injected sequences.

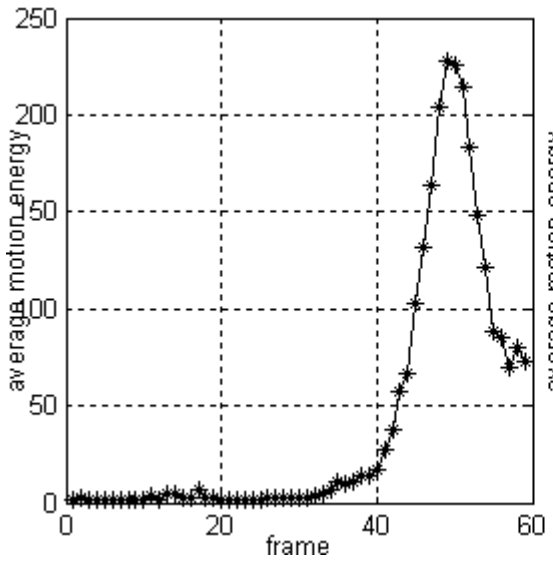
As shown in Figure 5.6 (a) and (e), small motion and big motion take place around Frames 20 and 40, respectively, in the noise-injected *Susie* sequence; as expected, Figure 5.6 (c) shows the corresponding small drop and big drop of PSNR for Model II and the proposed model. Similar phenomena occur in *Miss America* when there is small motion around Frames 3 and 18, as illustrated in Figure 5.6 (b), (d) and (f).

It is obvious that Model I is not able to respond to motion since the temporal CSF has not been incorporated; the proposed JND model predicts correctly (inheriting the temporal characteristics from Model II) that more distortions can be tolerated with higher motion.

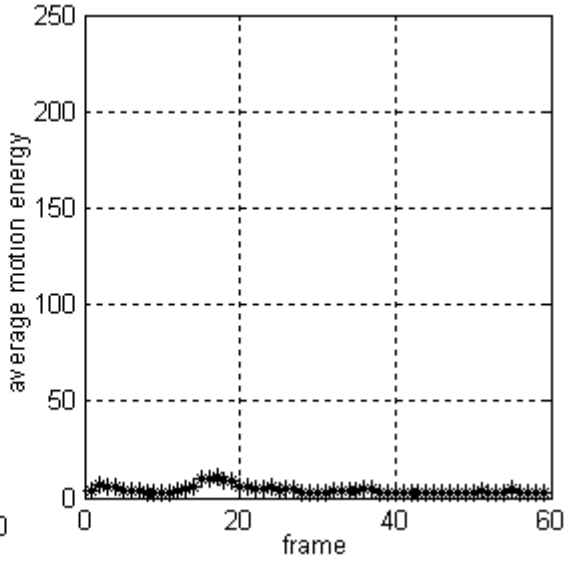
We notice that the motion energy actually fluctuates in Figure 5.6 (e) and (f) (apart from the aforementioned motion around Frames 20 and 40 in Figure 5.6 (e), and Frames 3 and 18 in Figure 5.6 (f)), while the PSNR curves by Model II and the proposed model in Figure 5.6 (c) and (d) do not reflect such changes of motion; in fact, this demonstrates the effect of eye movement: for low motion, eye movement can compensate some, if not all loss of sensitivity, and thus the JNDs derived by Model II and the proposed model do not undertake considerable elevation for these slight motion fluctuations.

In addition, the PSNR curve by the proposed model has similar shape but with a lower valuation levels than that by Model II; this demonstrates the effectiveness of

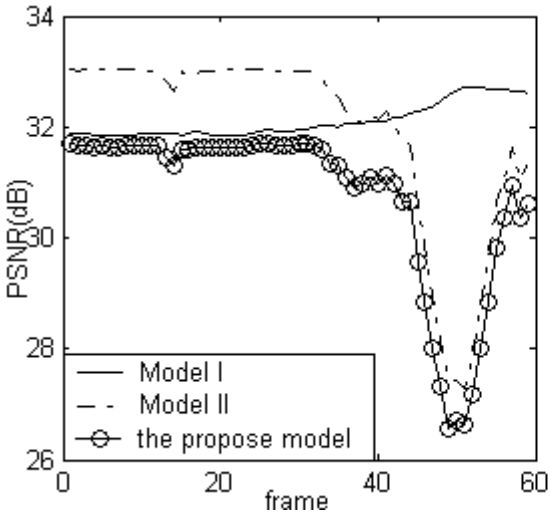
luminance adaptation and contrast masking considered in the proposed model.



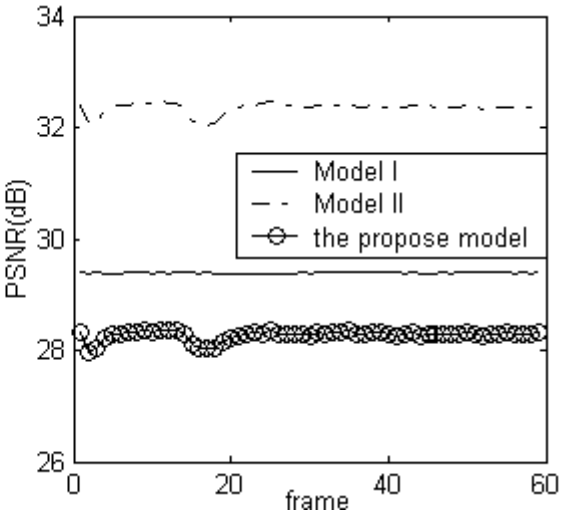
(a) *Susie*:
average motion energy vs. frame



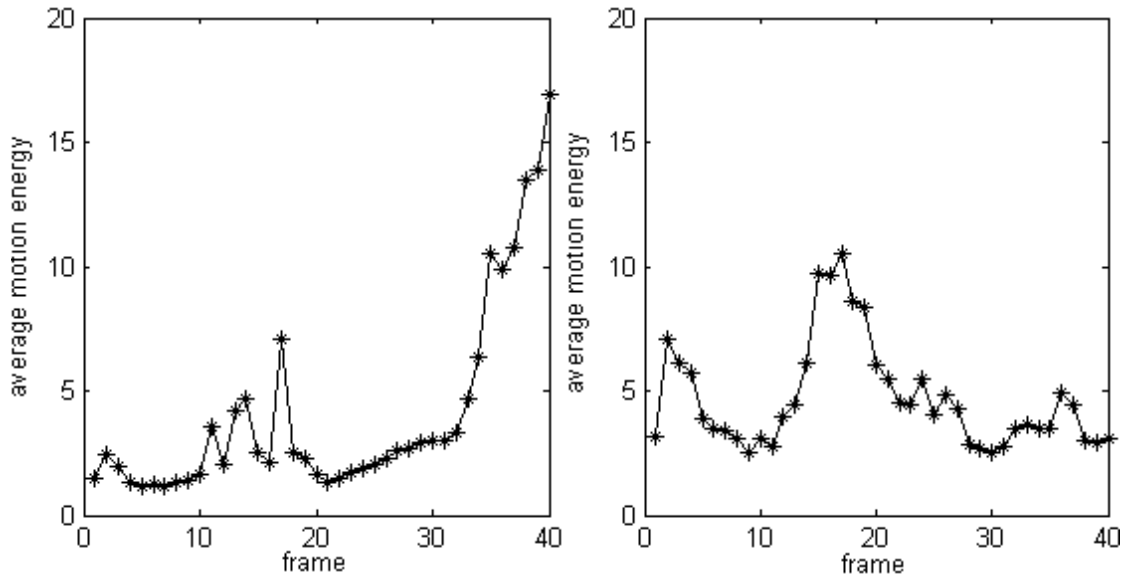
(b) *Miss America*:
average motion energy vs. frame



(c) *Susie*: PSNR vs. frame



(d) *Miss America*: PSNR vs. frame



(e) details of (c) before Frame 40 (f) details of (d) before Frame 40

Figure 5.6 Demonstration of the effect of motion.

2) Noise shaping for individual frames

In order to provide the ground to demonstrate the temporal CSF effect with the proposed model (to be discussed in Section 5.3.2-(3)), noise-injection has been performed with the first frame of the video sequences in Fig. 5.5. Since each frame is treated as a still image, $v(n,t)$ is set to be 0.15 deg/sec in Model II and the proposed model. In line with the results obtained in Section 5.3.1, the same subjective viewing tests have confirmed that the three models yield similar perceptual quality in the noise-injected frames. Figure 5.7 shows the noise-injected images for the first frame of *Bus* sequence with Model I, Model II and the proposed JND model.

Figure 5.8 illustrates the PSNRs of noise-contaminated frames of different videos by the three models (without temporal CSF effect). Similar to the results in Section 5.3.1,

the proposed model performs slightly better than Model I in PSNR reduction; and it yields an average 3.41 dB of PSNR reduction from Model II (as shown in Figure 5.8).

(a) Model I, PSNR: 28.93 dB

(b) Model II, PSNR: 33.08 dB

(c) the proposed JND model, PSNR: 28.70 dB

Figure 5.7 Noise-injection to the first frame of *Bus* sequence with Model I, Model II and the proposed JND model.

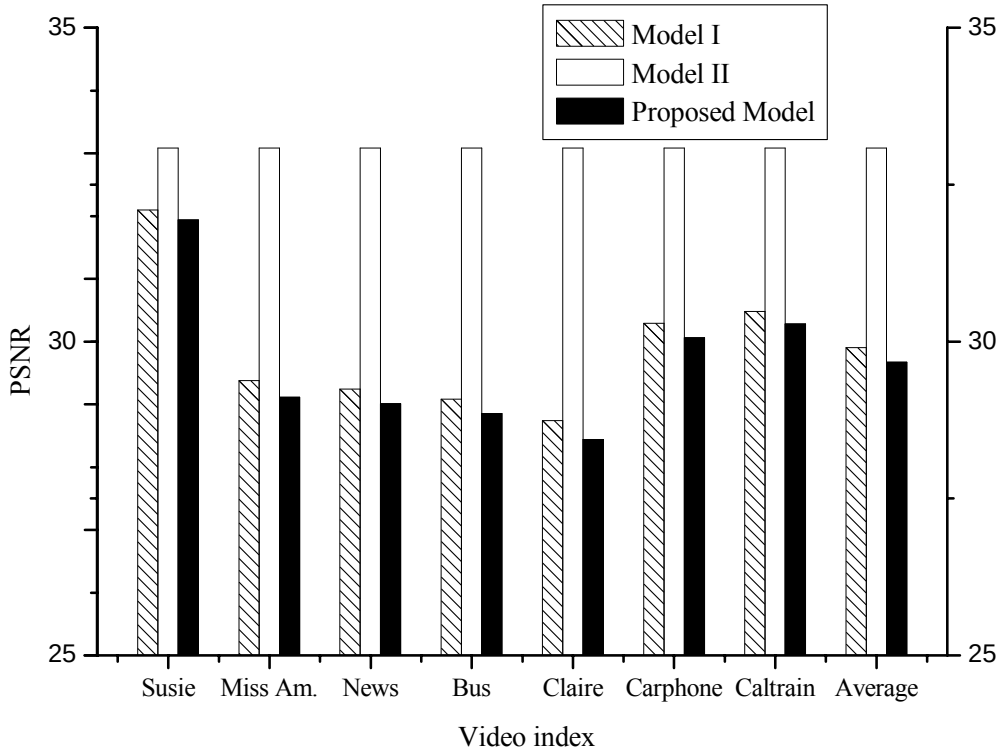


Figure 5.8. PSNRs of Noise-contaminated frames of videos by the three models (without temporal CSF effect)

3) Overall Performance

For the comprehensive performance comparison, we followed the procedures of the Double Stimulus Continuous Quality Scale (DSCQS) method in Rec. ITU-R BT.500 [67], in the subjective experiments for video. Figure 5.9 shows the presentation course. The *Mean Opinion Score* (MOS) scales are adopted for quality grading: Excellent (100-80), Good (80-60), Fair (60-40), Poor (40-20) and Bad (20-0), for both an original sequence and the associated processed sequence. Ten viewers (three were in the image processing field and seven were naive) were required to give MOS for both of the original and the processed sequences (they were not told which sequence is the original one). Then the different MOS (DMOS) is obtained by subtracting the MOS of the processed sequence from that of the original one. A higher DMOS indicates bigger quality discrepancy between the processed sequence and the original one. Again, there is no notable difference found regarding whether a subject has image processing knowledge.

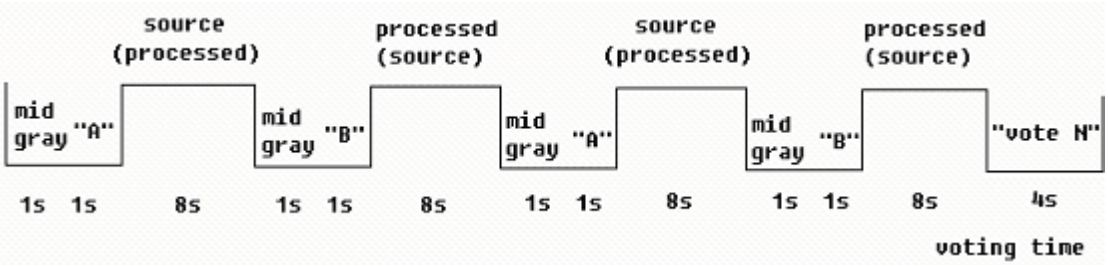


Figure 5.9. DSCQS test scheme

Figure 5.10 compares the mean DMOS with the ten viewers for each sequence, and Table 5.3 lists the corresponding standard deviations. It is shown in Figure 5.10 that

the subjective quality of the noise-injected video sequences is quite similar with the three models (the average DMOSs for Model I, Model II and the proposed model is 10.7, 10.1, 10.5, respectively). However, Figure 5.11 demonstrates that the average PSNR of noise-injected sequences by the proposed model is 0.83 dB lower than that by Model I, and 3.32 dB lower than that by Model II. Comparing the PSNRs for the proposed model in Figure 5.8 and Figure 5.11, it can be seen that the consideration of temporal CSF brings about 0.6 dB additional perceptual data redundancy on average. In summary, the proposed model is able to give more aggressive JND estimation than the other two models, without effects on the perceptual video quality, and the modeling of temporal effect is effective.

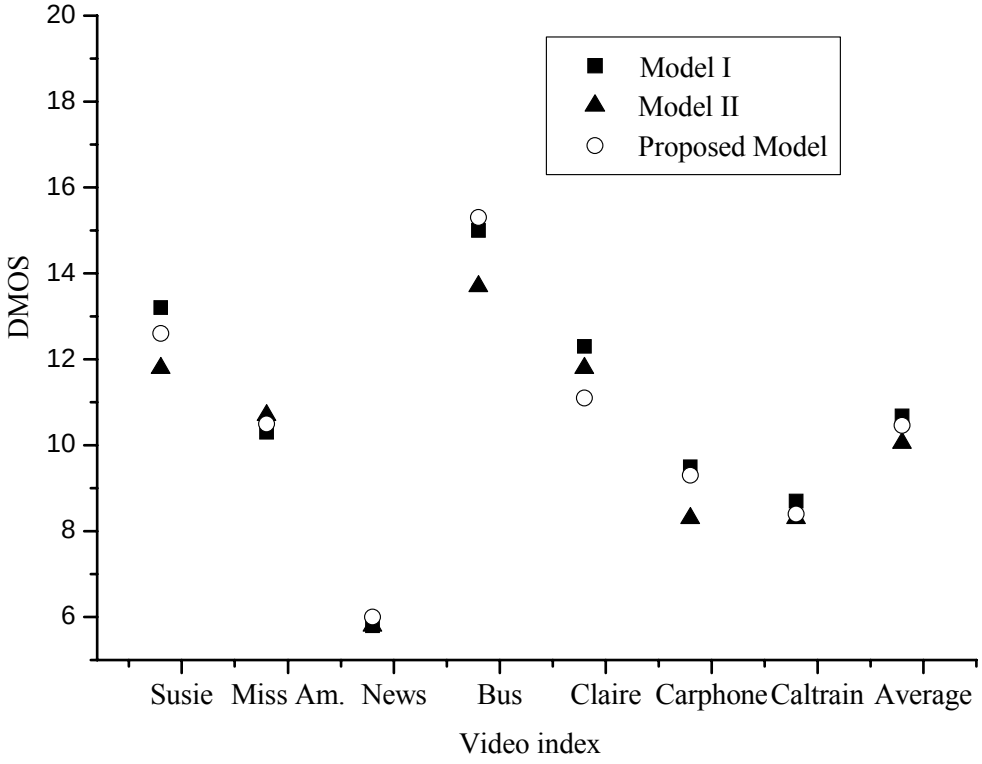


Figure 5.10 Mean DMOSs for the noise-injected videos with the three JND models

Table 5.3 Standard deviations of DMOSs for the noise-injected videos

Video	Susie	Miss Am.	News	Bus	Claire	Carphone	Caltrain
Model I	1.932	2.359	1.619	1.633	2.214	2.121	1.418
Model II	1.549	2.111	1.751	2.908	1.932	2.359	2.058
Proposed Model	2.119	2.8389	2.1089	1.947	0.995	1.703	2.011

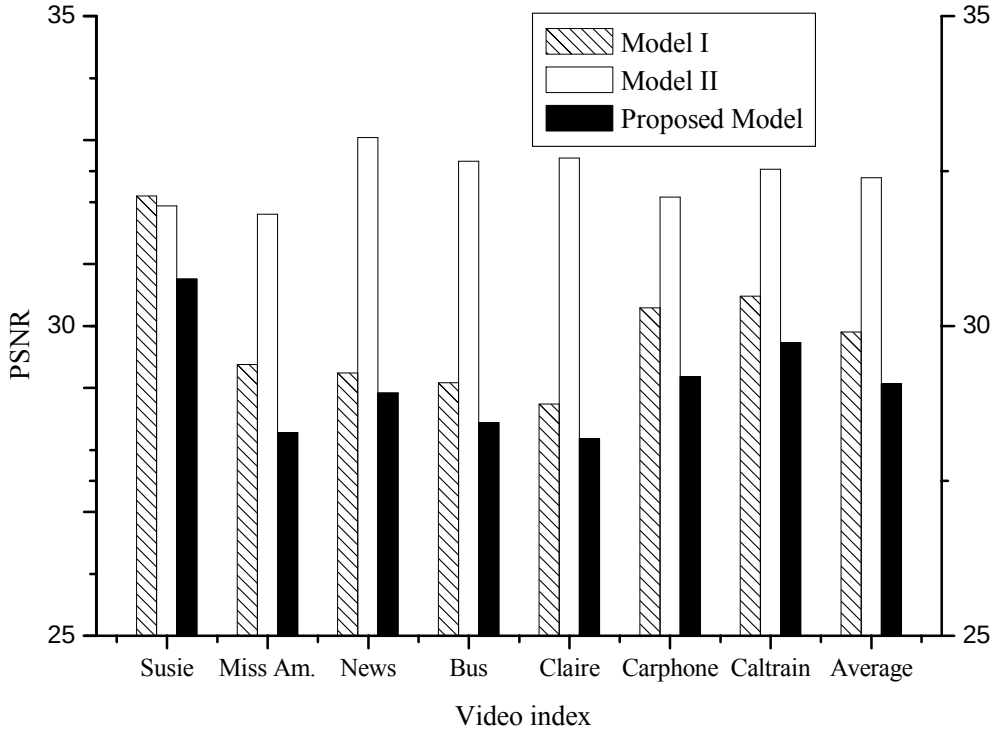


Figure 5.1. PSNRs of Noise-contaminated videos by the three models

4) Discussion

A better JND model is capable of determining a more aggressive error profile (e.g.,

with lower PSNR in the experiments described above) for an image but not causing noticeable visual distortion. Determination of the biggest unnoticeable error bounds in visual signal facilitates various processing tasks for resource savings and performance improvement.

In image and video compression, the budgeted bits can be allocated for better coding quality using a more accurate JND profile, with more bits assigned (via quantization step selection) [1,3,4,36,46] to the signal components with lower JNDs; only the DCT coefficients above the JNDs need to be coded [68], and this can result in computational complexity reduction and bit savings for more significant signal components (or alternatively, better perceptual picture quality with a same bit rate).

Accurate JND estimation benefits the non-compression processing processes as well. In watermarking, authentication, and error protection applications, the accessory data can be embedded inside the visual signal itself with the guide of the JND profile towards the minimum visual quality degradation [69,70]. For visual quality/distortion prediction, a metric can be defined or fine-tuned according to the JND [36,55] for better matching the HVS perception; a JND-based perceptual metric may be also adopted beyond the quality evaluation purpose (e.g., for image synthesis [71]).

5.4 Summary

In this chapter, we have compared the proposed JND estimation model with two

existing models (Model I and Model II defined in Section 5.1) based on a noise-injection scheme. We conducted the experiments separately for image and video. As for the image case, our model is comparable to Model I (even better performance in smooth areas of the image) and superior to Model II because it considers luminance adaptation and contrast masking. As for the video case, our model could add more invisible noise into frames of high motion compared with Model I, which shows that the addition of temporal contrast sensitivity and eye movement factor into the visual model is more consistent with human perception.

CHAPTER

6

Perceptual Image Compression Application

6.1 Introduction

In the previous chapters, we have discussed the properties of the HVS. The ultimate purpose of perception related research is actually to render more efficient and effective visual data processing. Therefore, in this chapter, the application of a JND model for image coding will be demonstrated.

As discussed earlier in Chapter 1, a JND model can play an important role in perceptual image coding. With a near-lossless compression technique, we can achieve perceptually lossless results if the coding error is below the corresponding JND. For lossy image compression, the use of JND can facilitate perceptually optimized coding. Based on the above consideration, we propose an image compression scheme for both perceptually lossless and perceptually optimized lossy compression of color images. A type of Hartley transform, an integer transform, is used for efficient decorrelation and

energy compaction before JND-based quantization. Pixel-based JND is more straightforward in this application. The JND profile accounts for the combined effect of both luminance masking and contrast masking and can be more accurately estimated after image pixel classification. Experimental results show that the proposed scheme achieves higher compression in the perceptually lossless mode and better visual quality in lossy mode compared with other related coding methods.

The chapter has been organized as follows: firstly in Section 6.2, the Hartley transform for frequency analysis will be introduced; then a JND model will be presented with finer pixel classification (Section 6.3); the coding scheme for perceptually lossless image compression as well as its extension to perceptually optimized lossy image compression will be proposed in Section 6.4. The experimental results are next presented (Section 6.5).

6.2 Hartley Transform

Discrete Hartley transform (DHT), an efficient integer transform at length N=4, can be extended to 2-D *lossless Hartley transform* (L-HT) [72]. According to [74], the L-HT for a 4x4 array $x(p, q)$ is defined as:

$$X(m,n) = \sum_{p=0}^{3}\sum_{q=0}^{3} x(p,q) \bullet cas(\pi mp/2)cas(\pi nq/2) \tag{6.1}$$

where $cas(\alpha) = \sin(\alpha) + \cos(\alpha)$, $m, n, p, q = 0, 1, 2, 3$.

In the above 4x4 L-HT, the basis vectors take only the binary values +1 and -1, which makes it computationally less complex. Furthermore, the forward and inverse HT share the same transform kernel, and this facilitates economical hardware/software implementation. According to (6.1), the coefficients of HT are obtained via a linear combination of the pixel values in the original image, and therefore, the distortion due to compression can be controlled by the quantization step, d. For the n-th block, the reconstructed error $E_n(p, q)$ for any pixel (p, q) within the block is constrained by:

$$E_n(p,q) \le (d/2) \tag{6.2}$$

for $p, q = 0, 1, 2, 3$.

6.3 JND in Pixel Domain

In real-life images, the JND value of each pixel of the image is associated with the inter-relevance of two factors: luminance masking and contrast masking [46]. The combination of these two masking effects for JND calculation in image domain can be modeled by the following nonlinear additivity model [47]:

$$T_{JND_\theta}(x,y) = T^l(x,y) + T^c_\theta(x,y) - C^{lc}_\theta \bullet \min\left\{T^l(x,y), T^t_\theta(x,y)\right\} \tag{6.3}$$

where $\theta = Y, Cb, Cr$, denotes the three channels for a color image; $T^l(x, y)$ and $T_\theta^c(x, y)$ are the visibility thresholds due to luminance masking and contrast masking, respectively; and the last term of equation (6.3) represents the interactional effect of

the two masking types.

According to [46], $T^l(x, y)$ can be described by:

$$T_l(x,y) = \begin{cases} T_0(1-\sqrt{\frac{\overline{I(x,y)}}{127}})+3 & if \quad \overline{I(x,y)} \le 127 \\ \gamma(\overline{I(x,y)}-127)+3 & otherwise \end{cases} \tag{6.4}$$

where

$$\overline{I(x,y)} = (1/32)\sum_{i=1}^{5}\sum_{j=1}^{5} I(x-3+i, y-3+j) \bullet B(i,j) \tag{6.5}$$

accounts for the average background luminance, and $B(i, j)$ is a weighted low pass operator shown in Figure 6.1. T_0 and γ are set to 17 and 3/128 respectively based on the experimental results in [46]. For color images, only the information of the luminance channel is used to estimate luminance masking.

$$\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 2 & 0 & 2 & 1 \\ 1 & 2 & 2 & 2 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

Figure 6.1 The low pass operator *B*

Contrast masking depends on the local texture activity of the image. The HVS sensitivity to error is generally higher in smooth, or plain areas, and lower in the

texture area [34]; while the sensitivity for edge areas lies in between. Therefore, we discriminately calculate contrast masking for different image pixel classes. We use the Canny method [75] to detect the edges in the image and classify a block as an *edge class* if there are more than two edge pixels in the 4x4 block. We then classify non-edge blocks as *smooth class* or *texture class* based on the block-based standard deviation (STD). We give different weights to different block classes and obtain a classification map W_θ after a Gaussian low-pass filtering:

$$W_\theta = C_\theta * h \tag{6.6}$$

where $C_\theta(n) = \begin{cases} 1 & when\ \ block\ \ n \subset texture\ \ class \\ 0.3 & when\ \ block\ \ n \subset edge\ \ class \\ 0.1 & when\ \ block\ \ n \subset smooth\ \ class \end{cases}$

and h is a $k \times k$ Gaussian low pass filter with standard deviation σ (k=7 and σ=0.8).

The contrast masking is then determined as:

$$T_\theta^t(x,y) = G_\theta(x,y) \bullet \beta_\theta \bullet W_\theta(x,y) \tag{6.7}$$

where G_θ denotes the maximal weighted average of gradients [46] around the pixel at (x, y). β_θ is the empirical weights for each color channel and is determined according to the subjective experiment in [47]. Table 6.1 shows the values of the parameters in the model.

Table 6.1 Empirical experimental parameters for the JND model

Parameter	C_θ^{lc}			β_θ		
	Y	*Cb*	*Cr*	*Y*	*Cb*	*Cr*
Value	0.3	0.25	0.2	0.117	0.65	0.45

6.4 JND Guided Image Compression

6.4.1 Perceptually Lossless Compression

The proposed coding scheme can be described by the block diagram in Figure 6.2.

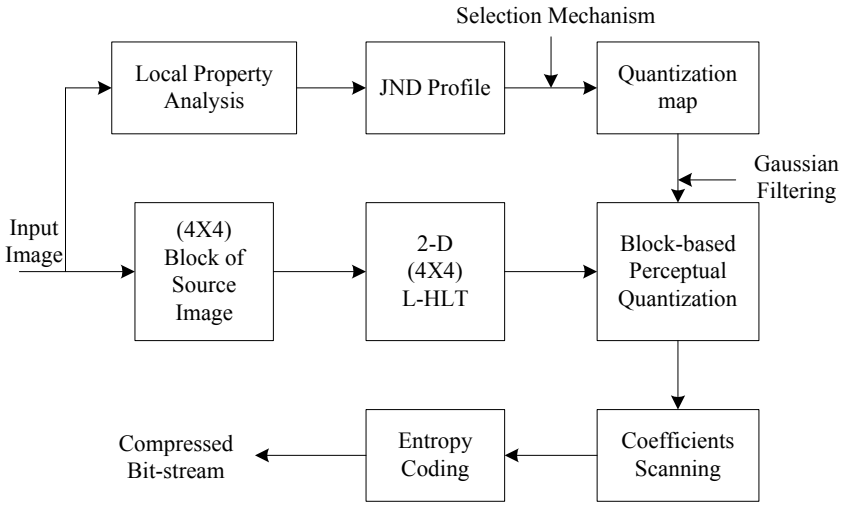


Figure 6.2 Block diagram for the proposed encoding process

In the selection mechanism, we set the quantization step for each block as:

$$Q_\theta(n) = 2 \bullet \left\lfloor \sqrt{\frac{\sum_{p=0}^{3}\sum_{q=0}^{3} T_{JND_\theta}(p,q)^2}{16}} \right\rfloor \qquad (6.8)$$

where $T_{JND\theta}$ (p, q) $(p, q=0, 1, 2, 3)$ denotes the JND values within the 4x4 block n in channel θ. Operator $\lfloor\rfloor$ gets the maximum integer that is less than or equal to the inside. We call $T_{JND\theta}(x, y)^2$ *JND energy* [46] for pixel (x, y). According to equation (6.2), equation (6.8) guarantees that the total error energy of each block in a reconstructed image is below the JND energy in the corresponding block. In this way, we realize the perceptually lossless coding. The scanning order for HT coefficients is depicted in Figure 6.3**.**

$$\begin{bmatrix} 1 & 2 & 4 & 6 \\ 3 & 8 & 9 & 11 \\ 5 & 10 & 13 & 14 \\ 7 & 12 & 15 & 16 \end{bmatrix}$$

Figure 6.3 The scanning order of HLT coefficients

6.4.2 Perceptually-Optimized Lossy Compression

With a lower bit-rate than that required by perceptually lossless coding, we aim at achieving the best possible perceptual quality of compressed images at a given bit-rate budget. In order to do this, we extend the above perceptually lossless coding scheme to

optimized lossy image coding by dispensing the distortion among the image pixels according to the perceptual importance information given by the JND profile. Here we use minimally noticeable distortion (MND) [6] instead of JND for the perceptual consideration. Different from the JND leading to a visually transparent processing, MND renders non-transparent but still visually optimum image/video under a bit rate.

In our project, the MND is essentially a scaled version of the JND profile depending on the bit rate budget:

$$T_{MND_\theta}(x,y) = \alpha \bullet T_{JND_\theta}(x,y) \qquad (6.9)$$

where α is the adjustable parameter for different bit-rate budgets. Then the quantization step is set:

$$Q_\theta(n) = 2 \bullet \left\lfloor \sqrt{\frac{\sum_{p=0}^{3}\sum_{q=0}^{3}(T_{MND_\theta}(p,q))^2}{16}} \right\rfloor \qquad (6.10)$$

The rest of the lossy image compression scheme is similar to the perceptually lossless compression described in section 6.4.1.

6.5 Experimental Results

6.5.1 Perceptually Lossless Compression

We compare our perceptually lossless compression scheme with the near lossless scheme described in [72] (uniform quantization with quantization step d =5), based on the same L-HT. Table 6.2 lists the *bit per pixel* (bpp) comparison of the two schemes

for different (grey-level and color) images from image database (Table 6.3). As can be seen, the proposed scheme significantly lowered the bit rate compared with the near lossless scheme.

Table 6.2 Comparison of bit-rates for the proposed compression scheme and the near lossless compression scheme (with uniform quantization)

Image	Near lossless coding (d=5) (bpp) [72]	Proposed perceptually lossless coding (bpp)
A	3.31	2.31
B	3.01	2.07
C	4.20	2.83
F	11.05	7.37
G	6.25	4.06
H	4.27	2.29

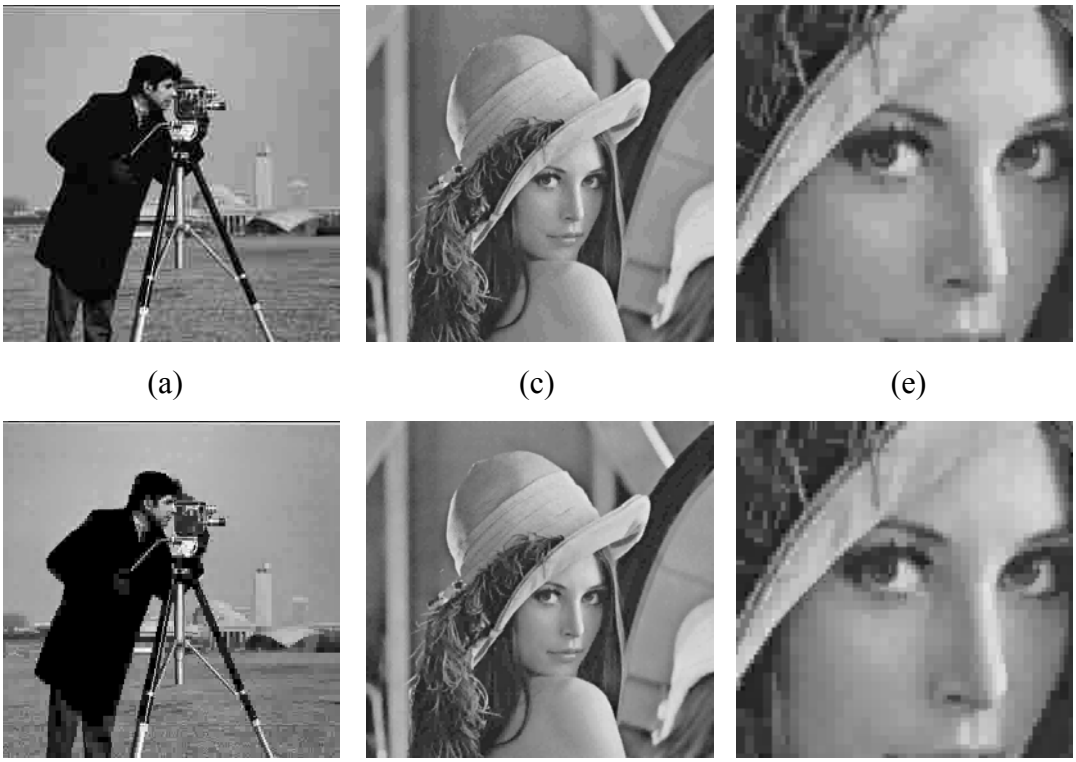
Table 6.3 Image database for the experiments

Gray-level Image		Color Image	
Image Index	Image Description	Image Index	Image Description
A	cameraman256x256	F	mandrill512x512
B	pepper512x512	G	peppers256x256
C	goldhill256x256	H	splash512x512
D	airplane512x512	I	lena512x512
E	lena512x512	J	house256x256

6.5.2 Perceptually-Optimized Lossy Compression

We compare the visual quality of reconstructed images obtained by MND-based compression with that obtained using uniform quantization and the standard JPEG at an equivalent bit rate in Figure 6.4, where we see that image (b,d,f) by the proposed scheme appears to have less visible distortion than images (a,c,e) by the other two

scheme. The difference between the schemes is more obvious in the smooth areas, due to our block classification in the JND model.



(a) (c) (e)

(b) (d) (f)

Figure 6.4 Comparison of visual quality between other coding methods and the proposed MND-quantization-based coding method. (a) image A (Table 6.3) by JPEG scheme at 0.4 bpp; (b) image A (Table 6.3) by our proposed scheme at 0.4 bpp; (c) image E (Table 6.3) by uniform quantization at 0.35 bpp; (d) image E (Table 6.3) by our proposed scheme at 0.35 bpp; (e) zoom-in image of (c); (f) zoom-in image of (d)

We conducted subjective viewing experiments to give quantitive scores for all the images concerned. Under the same experimental environment as in [47], we performed 10 trials on the images from the database (Table 6.3). In each trial, two reconstructed images (from the uniform-based coding method and the proposed method at equivalent bit rate) of a same image were juxtaposed on the screen and 6 subjects (3 are in the

image processing field and 3 are naive) were asked to rate the comparative visual quality of the pair according to Table 6.4. The rating results are listed in Table 6.5, where the mean subjective scores and the standard deviation are computed based on the 10 trials. In Table 6.5, the positive average mean with the average standard deviation of 0.761 shows that overall subjective rating favors the proposed scheme. Thus, we can say that our perceptual coding scheme helps better distribute the reconstruction distortion so as to optimize the perceived quality of decoded images.

Table 6.4 Subjective rating table for comparing the visual quality of a pair of images

-3	the left one[1] much better than the right one[2]
-2	the left one better than the right one
-1	the left one slightly better than the right one
0	the same
+1	the left one slightly worse than the right one
+2	the left one worse than the right one
+3	the left one much worse than the right one

[1.] the left one: decoded image by uniform-quantization-based coding method
[2.] the right one: decoded image by MND-quantization-based coding

Table 6.5 Results for subjective evaluation

Subject index	Mean	Standard Deviation
1	+1.2	0.919
2	+0.9	1.135
3	+1.3	0.823
4	+1.1	0.568
5	+1.2	0.422
6	+0.6	0.699
Average	+1.05	0.761

6.6 Summary

In this chapter, we give an example of using the JND to facilitate image coding. A unified scheme for both perceptually lossless image compression and perceptually optimized lossy image compression based on L-HT and JND estimation in pixel domain has been proposed. The experiments show that in perceptually lossless mode, the reconstructed error is controlled below the visual threshold of the human perception, so that better compression performance can be achieved without jeopardizing the visual quality of the decoded image. While in lossy mode, we optimize the compression by distributing more distortion to image regions of less perceptual importance.

CHAPTER

7

Conclusion and Future Work

Recent developments in vision research have been contributing significantly to the advancement of perception-related research. How to effectively apply the characteristics of the human visual system to optimize digital imaging systems becomes increasingly important. For applications such as image/video coding and quality evaluation, pertinent understanding and proper modeling of human vision is essential.

In this thesis, the main properties of the human visual system are first explored. Most of these properties can be simulated and represented by mathematical models. Appropriately combining these separated one-fold models leads to a rounded vision model, which mimics the human perception to certain extent for practical applications. Several existing perceptual models have been reviewed in the work to set a background for the proposed model.

7.1 Concluding remarks

The major contribution of this thesis is the design of a DCT-based spatio-temporal JND (just noticeable distortion) estimation model, because a stand-alone JND estimation model can hardly be found. In comparison with the image case, estimation of JND for video needs to take the temporal HVS properties into account, in addition to the spatial properties. The temporal factor is considered in the model with a spatio-temporal CSF (contrast sensitivity function) model. Since eye motions may change the shape of spatial CSF, an eye movement model is incorporated into the spatio-temporal CSF to compensate for this mechanism. Similar to the model for images, luminance adaptation and contrast masking are inserted to account for the spatial properties of each frame in the video sequence. Compared to the related work [35], we exclude smooth blocks from the intra-band masking because we find that human vision is quite sensitive to the noise in the smooth areas even when motion takes place.

Experimental results with subjective viewing confirm the improved performance of the proposed model. The model is capable of predicting more aggressive JND values without introducing noticeable distortion for both images and videos, and therefore outperforms the relevant existing models.

We finally give an example of applying the JND model into the image coding scheme. A JND model estimating visual thresholds in pixel domain for images has been

introduced. In order to better estimate the contrast masking phenomenon, a blocking classification method has been adopted to separate the image blocks into the *smooth*, *edge* and *texture* group. Luminance adaptation has also been incorporated for the complete construction of the JND model. Based on Hartley transform and JND estimation in pixel domain, a unified scheme for both perceptually lossless image compression and perceptually optimized lossy image compression has been proposed.

The experiments show that in perceptually lossless mode, the reconstructed error is controlled below the visual threshold of the human perception, so that better compression performance can be achieved without jeopardizing the visual quality of the decoded image. While in lossy mode, we optimize the compression by distributing more distortion to image regions of less perceptual importance, so that a tradeoff between visual quality and bit rate budget is achieved.

7.2 Future work

Though the proposed JND model has already considered many spatial and temporal properties of the human visual system, there are still more to be added. For example, higher processing in the human perception related to visual attention and foveal property are also very important yet not well developed for modeling the HVS and for quality evaluation. In a video sequence, the foreground object and motion tend to draw more attention from the observer. Therefore, we can further enhance the JND model for background and visually unnoticed areas.

In our model, we have made several assumptions to simplify modeling. We can exploit these in the future for more accurate modeling. For instance, it has been assumed that the HVS tracks different parts of an image equally (Section 4.2.2), but this is just an approximation, especially for a large-size image.

Moreover, our model is designed only for gray-level images and video. Although achromatic factors play more important roles than chromatic factors in terms of perception, it should not be ignored when a more thorough and accurate model is desired.

As for practical applications, the more accurate JND estimation towards the actual visibility bounds can facilitate resource savings (e.g., for bandwidth/storage, computation) and performance improvement (for perceived quality, etc.) in video coding, as well as improvement in various other visual processing tasks (such as perceptual quality evaluation, visual signal restoration/ enhancement, watermarking, authentication, and error protection).

Bibliography

[1] C. -H. Chou and Y.-C. Li, "A perceptually optimized 3-D subband codec for video communication over wireless channels," in *IEEE Trans. Circuits Syst. Video Technol.*, vol.6, no.2, pp. 143- 156, 1996.

[2] H. A. Peterson, A. J. Ahumada Jr. and A. B. Watson, "Improved detection model for DCT coefficient quantization," in *Proceedings of the SPIE International Conference on Human Vision, Visual Processing, and Digital Display IV*, vol. 1913, 1993.

[3] I. Hontsch, and L. J. Karam, "Adaptive image coding with perceptual distortion control", in *IEEE Trans. on Image Processing*, vol. 11, No. 3, pp. 213-222, 2002.

[4] R. J. Safranek & J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression", in *Proceedings International. Conference on Accoustics, Speech and Signal Processing*, New York, NY, vol. 3, pp. 1945-8, May 1989.

[5] Renxiang Li, Bing Zeng and Ming L. Liou, "A new three step search algorithm for block motion estimation", in *IEEE Transactions on Circuit and Systems for Video Technology*, Vol. 4, No. 4, pp. 438-442, 1994.

[6] N. Jayant, J. Johnston, and R. Safranek, "Signal Compression Based on Models of Human Perception", in *Proceedings of the IEEE*, 81(10), October 1993.

[7] X. K. Yang, W. Lin, Z.K. Lu, E.P. Ong and S.S.Yao, "Perceptually-Adaptive

Hybrid Video Encoding Based On Just-noticeable-distortion Profile", in *SPIE 2003 Conference on Video Communications and Image Processing (VCIP)*, Vol.5150, pp.1448-1459, Lugano, Switzerland, July 2003.

[8] X. K. Yang, W. Lin, Z.K. Lu, X. Lin, R. Susanto, E.P. Ong and S.S.Yao, "Rate Control for Videophone Using Local Perceptual Cues", in *IEEE Transactions on Circuit and Systems for Video Technology*, Vol. 15, No. 4, pp. 496-507, 2005.

[9] Wilfried. Osberger, Anthony. J. Maeder, "Automatic Identification of Perceptually Important Regions in an Image", in *Proceedings of the Fourteenth International Conference on Pattern Recognition,* Vol.1, pp. 701-704, Australia, 1998.

[10] Scott Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models", in *Proc. of SPIE Human Vision and Electronic Imaging III*, Vol.3299, pp180-191, San Jose, California, January 1998.

[11] E.P. Ong, W. Lin, Z.K. Lu, S.S.Yao, X. K. Yang and F. Moschetti, "Low bit rate video quality assessment based on perceptual characteristics", in *IEEE International Conference on Image Processing*, Vol. 3, pp.189-192, Singapore, 2003.

[12] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies", in *SPIE Video Communications and Image Processing Conference*, Lugano, Switzerland, 2003.

[13] Mahesh Ramasubramanian, Sumanta N. Pattanaik, Donald P. Greenberg, "A Perceptually Based Physical Error Metric for Realistic Image Synthesis", in *Proceedings of SIGGRAPH 99 Conference*, pp73-82, Los Angeles, CA, 1999.

[14] Scott Daly, Kristine Matthews and Jordi Ribas-Corbera, “Face-Based Visually-Optimized Image Sequence Coding”, in *IEEE Proc. Int. Conf. Image Processing (ICIP),* pp. 443-447, Chicago, IL, Oct. 1998.

[15] Christian J. van den Branden Lambrecht, “A Working Spatio-temporal Model of the Human Visual System for Image Restoration and Quality Assessment Applications”, in *IEEE Proceedings of the Intl. Conf. on Acoustics, Speech, and Signal Processing*, pp. 2293-2296, Atlanta, GA, May 1996.

[16] Zhenghua Yu and H. R. Wu, “Human Visual System based Objective Digital Video Quality Metrics”, in *Proceedings of the International Conference on Signal Processing of IFIP World Computer Conference*, 2, pp. 1088–1095, August 2000.

[17] Stefan Winkler, “Issues in Vision Modeling for Perceptual Video Quality Assessment”, *Signal Processing*, 78(2):231–252, Oct.1999.

[18] R. E. Fredericksen, R. F. Hess, “Temporal Detection in Human Vision: Dependence on Spatial Frequency”, *Opt. Soc. Am. A*, Vol.16, No. 11, pp2601-2611, November 1999.

[19] R. E. Fredericksen, R. F. Hess, “Temporal Detection in Human Vision: dependence on stimulus energy”, *Opt. Soc. Am. A*, Vol.14, No. 10, pp2557-2569, October 1997.

[20] Stefan J.P. Westen, Reginald L. Lagendijk and Jan Biemond, “A Quality Measure for Compressed Image Sequences Based on an Eye Movement Compensated Spatio-temporal Model”, in *IEEE International Conference on Image Processing*, Vol. 1, pp.279-282, 2003.

[21] Michael P. Eckert, Gershon Buchsbaum, and Andrew B. Watson, "Separability of Spatiotemporal Spectra of Image Sequences", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 12, Dec. 1992

[22] Andrea Cavallaro, Stefan Winkler, "Segmentation Driven perceptual Quality Metric", in *IEEE International Conference on Image Processing*, pp.3543-3546, Singapore, 2004.

[23] Mark A. Masry and Sheila S. Hemami, "CVQE: A Metric for Continuous Video Quality Evaluation at Low Bit Rates", in *SPIE Conf. on Human Vision and Electronic Imaging*, 2002.

[24] A. Bovik, *handbook of image and video processing*, Academic Press, San Diego, May 2000.

[25] Arun N. Netravali and Barry G. Haskell, *Digital Pictures: Representation, Compression, and Standards*, Second Edition, Plenum Press, New York and London.

[26] VQEG (Video Quality Expert Group), *Final report from the video quality expert group on the validation of objective models of video quality assessment*, March 2000, http://www.vqeg.org.

[27] D. H. Kelly, "Motion and vision II: Stabilized spatiotemporal threshold surface", *J. Opt. Soc. Amer.*, Vol. 69, no. 10, pp.1340-1349, 1979.

[28] Stefan Winkler, *Vision models and Quality Metric for Image Processing Application*, Lausanne, EPFL, Dec. 21, 2000

[29] Davson H (1990) *Physiology of the Eye, 5th ed*. London: Macmillan Academic and

Professional Ltd.

[30] T. N. Cornsweet, *Visual Perception*, New York, Academic Press, 1970.

[31] Rose A., "The sensitivity performance of the human eye on a absolute scale", *Journal of the Optical Society of America,* 38:196-208, 1948.

[32] G. E. Legge and J. M. Foley, "Contrast masking in human vision", *Journal of the Optical Society of America*, Vol. 70, pp. 1458-1471, 1980

[33] C. Carlson and R. Cohen, "A simple psychophysical model for predicting the visibility of displayed information", in *Proc. of the Society for Information Display*, vol. 21, pp. 229-245, 1980.

[34] H. H. Y. Tong, A. N. Venetsanopoulos, "A perceptual model for JPEG applications based on block classification, texture masking, and luminance masking,"in *IEEE Int'l Conf. Image Processing*, Chicago, Oct. 1998.

[35] X. Zhang, W.S. Lin and P. Xue, "Improved estimation for just-noticeable visual distortion", *Signal Processing*, Vol. 85, Issue 4, pp. 795-808, April 2005.

[36] A. B. Watson, "DCTune: A technique for visual optimization of DCT quantization matrices for individual images", *Society for Information Display Digest of Technical Papers XXIV*, pp. 946-949, 1993.

[37] Yi-Jen Chiu, Toby Berger, "A Software-Only Videocodec Using Piexelwise Conditonal Differential Replenishment and Perceptual Enhancements", in *IEEE Trans. on Circuits and Systems for Video Tech.*, Vol. 9, No. 3, April 1999.

[38] K.T.Tan, M. Ghanbari and D.E.Pearson, "An objective measurement tool for MPEG video quality", *Signal Processing*, 70, pp. 279-294, 1998.

[39] A. B. Watson, James Hu, John F McGowan III, “Digital Video Quality Metric based on Human Vision”, *Journal of Electronic Imaging*, 10(1), 20-29,2001.

[40] Lukas, F. X. J. and Z.L.Budrikis, “Picture quality prediction based on a visual model”, in *IEEE Trans. on Communications*, vol. 30, no.7, pp1679-1692, 1982.

[41] Michael P. Eckert, Gershon Buchsbaum, “the Significance of Eye Movements and Image Acceleration for Coding Television Image Sequences”, *Digital Images and Human Vision*, MIT press, Cambridge, MA, USA, pp.89-98, 1993.

[42] P.E.Hallett, Chapter 10 in *Handbook of perception and human performance*, John Wiley and Sons, New York, 1986.

[43] S.J.P Westen, R.L.Lagendijk, J. Biemond, “Spatio-Temporal Model of Human Vison for Digital Video Compression”, in *IEEE International Conference on Image Processing*,Volume: 1 , 26-29 Oct. 1997

[44] Christian J van den Branden Lambrecht, “Color Moving Pictures Quality Metric”, 1996, in *IEEE International Conference on Image Processing,* vol. 1, pp. 885-888, 1996.

[45] M. Masry and S.S. Hemami, "Models for the perceived quality of low bit rate video", in *IEEE International Conference on Image Processing*, Rochester, NY, Sept. 2002.

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

[47] X. K. Yang, W. Lin, Z.K. Lu, E.P. Ong and S.S.Yao, “Just-noticeable-distortion

profile with nonlinear additivity model for perceptual masking in color images", in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2003)*, vol. 3, Hong Kong, pp.609-612, April 2003.

[48] G. E. Legge, "A power law for contrast discrimination", *Vision Research*, Vol.21, pp. 457-467, 1981.

[49] M.J.Nadenau, *Integration of human color vision models into high quality image compression*, PhD thesis, Lausanne, EPFL 2000.

[50] Peter G. J. Barten, *Contrast Sensitivity of the Human Eye and Its Effects on Image Quality*, SPIE Optical Engineering Press, Bellingham, Washington USA.

[51] Yang Li Hector Yee, *Spatiotemporal Sensitivity and Visual Attention for Efficient Rendering of Dynamic Environments*, Master thesis, 2000,Cornell University

[52] R. F. Hess, R. J. Snowden, "Temporal Properties of Human Visual Filters: Number, Shapes and Spatial Covariation", *Vision Research*, Vol .32, No.1, pp. 47-59, 1992.

[53] R. E. Fredericksen, R. F. Hess, "Estimating Multiple Temporal Mechanisms in Human Vision", *Vision Research*, Vol. 38, No. 7, pp1023-1040, 1998.

[54] Par Lindh and Christian J van den Branden Lambrecht, "Efficient Spatio-temporal Decomposition for Perceptual processing of Video Sequences" in *Proceedings of the International Conference on Image Processing*, vol. 3, pp. 331--334, Lausanne, Switzerland, September 16--19, 1996.

[55] Sarnoff Corp., *Sarnoff JND Vision Model Algorithm Description and Testing*, VQEG, Aug. 1997.

[56] Patrick C. Teo and David J. Heeger, "Perceptual Image Distortion", in

Proceedings of the International Conference on Image Processing, pp. 982-986, Austin, TX, November 13-16, 1994.

[57] Poirson and Wandell, "Pattern-color separable pathways predict sensitivity to simple colored patterns", *Vision Research*, 36(4), 515-526, 1996.

[58] A.B.Watson and J. A. Solomon, "A model of visual contrast gain control and pattern masking", *Journal of the Optical Society of America A,* 14, pp. 2379-2391, Sept. 1997.

[59] D. Pearson, "Viewer response to time-varying video quality", in *Proceedings of the SPIE – Human Vision and Electronic Imaging*, 3299, pp. 16-25, San Jose, CA, Jan. 1999.

[60] Jesus Malo, Juan Gutierrez, I. Epifanio, Francesc J. Ferri and Jose M. Artigas, "Perceptual Feedback in Multigrid Motion Estimation Using an Improved DCT Quantization", in *IEEE Trans. on Image Processing*, Vol. 10, No. 10, Oct. 2001.

[61] Albert J. Ahumada Jr.& Heidi A. Peterson, "Luminance-model-based DCT quantization for color image compression", in *SPIE Proceedings*, 1666, 365-374, 1992.

[62] P. E. Hallett, Chapter 10 in *Handbook of Perception and Human Performance*, John Wiley and Sons, New York, 1986

[63] R.W.Ditchburn, *Eye Movements and Perception*, Clarendon Press, Oxford, UK, 1973.

[64] G. C. Philips, H. R. Wilson, "Orientation bandwidths of spatial mechanisms measured by masking", *Journal of the Optical Society of America A*, vol. 1, pp.

226-232, 1984.

[65] A. B. Watson, “Detection and recognition of simple spatial forms”, in *O.J.Braddick, A.C. Sleigh, eds., Physical and Biological Processing of Images, Springer-Verla*g, Berlin, 1983.

[66] J. Park et al., “Some adaptive quantizers for HDTV image compression”, in L. Stenger et al., editors, *Signal processing of HDTV*, V. 1994.

[67] ITU-R, Recommendation BT.500-8, *Methodology for the subjective assessment of the quality of television pictures*, September 1998.

[68] R. J. Safranek, ``A JPEG compliant encoder utilizing perceptuallybased quantization", in *Proc. SPIE Human Vision, Visual Proc., and Digital Display V*, Vol. 2179, pp. 117-126, Feb. 1994.

[69] R. B. Wolfgang, C. I. Podilchuk, and E. J. Delp, ``Perceptual watermarks for digital images and video", in *Proc. IEEE*, 87(7), pp.1108-1126, July 1999.

[70] W. Zeng, ``Visual optimization in digital image watermarking", in *Proc. ACM Mulitimedia Workshop on Multimedia and Security*, 1999.

[71] M. Ramasubramanian, S. N. Pattanaik, and D. P. Greenberg, ``A perceptual based physical error metric for realistic image synthesis", *Computer Graphics (SIGGRAPH'99 Conference Proceedings)*, 33(4), pp. 73-82, August 1999.

[72] P. K. Meher, T. Srikanthan, J. Gupta, and H. K. Agarwal**,** “Near lossless image compression using lossless hartley like transform,” *Proc. of The Fourth IEEE Pacific-Rim Conf. on Multimedia*, SG, Dec. 2003.

[73] W. Lin, L. Dong, and P. Xue, “Discriminative analysis of pixel difference towards picture quality prediction,” *IEEE Int’l Conf. on Image Processing*, Barcelona, Spain, Sept. 2003.

[74] C. H. Paik, and M. D. Fox, “Fast hartley transform for image processing,” *IEEE Trans. on Med. Image*, Vol. 7, No. 6, pp. 149-153, 1988.

[75] Canny John, “A computational approach to edge detection,” *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 8, No. 6, pp. 679-698, 1986.