Perceptual Criteria for Image Quality Evaluation

Thrasyvoulos N. Pappas

Bell Laboratories
Lucent Technologies
Murray Hill, NJ 07974
Phone: (908) 582-2933 FAX: (908) 582-7308
Email: pappas@bell-labs.com

Robert J. Safranek

Bell Laboratories
Lucent Technologies
Murray Hill, NJ 07974
Phone: (908) 582-4949 FAX: (908) 582-7308
Email: rjs@bell-labs.com

1 Introduction

Recent advances in digital imaging technology, computational speed, storage capacity, and networking have resulted in the proliferation of digital images, both still and video. As the digital images are captured, stored, transmitted, and displayed in different devices, there is a need to maintain image quality. In this chapter we examine objective criteria for the evaluation of image quality that are based on models of visual perception.

An image is a two (or three) dimensional reproduction of the real world with an additional dimension for moving images. In this chapter we are only interested in images that are intended to be seen by humans. Such images could include a number of different imaging modalities, e.g. infrared, x-ray images, CAT scans, etc., as long as the ultimate destination is the human eye. The tasks could vary from looking at a photograph or watching a movie to reading text and trying to detect a target or a medical condition. However, the problem of image quality changes fundamentally when the ultimate user or interpreter of an image is not the human eye. For example, a typical quality metric for a range[1] image is a maximum (percent) deviation from the true value. On the other hand, the human eye's sensitivity to luminance variations depends on a number of factors, including light level, spatial frequency, and signal content.

Even though we use the term image quality, we are primarily interested in image fidelity, i.e. how close an image is to a given original or reference image. It is very hard to develop objective metrics that evaluate image quality without a reference image, even though the human visual system is very good at doing that. Also, we are not considering image enhancement (see Chapters 3.1–3.4) which can improve the quality of an image by modifying it, e.g. by increasing the contrast or changing the colors.

In the following sections, we will examine objective criteria for image quality that are based on models of the human visual system (HVS). Mean squared error based metrics (such as peak signal to noise ratio) are still widely used for performance evaluation, and despite their well-known limitations, they can be quite helpful if used carefully. However, they fail when one compares different kinds of artifacts [1] (e.g. artifacts of block-based vs. subband or wavelet coders).

[1] The values of a range image represent the distances from a point or planar surface to each point on the surface of an object or collection of objects.

The perceptual metrics we will discuss were developed for different applications. Even though each metric was influenced by the particular application it was developed for, some of the metrics have general applicability. Such metrics are more elaborate and computationally intensive. We will take a closer look at the metrics developed by Daly [2], Lubin [3], and Teo and Heeger [4]. We also examine metrics that were designed for specific applications, e.g. compression, halftoning, printing, and displays. These metrics are simpler and computationally efficient. Our main focus in this second category will be on metrics for image compression.

Even though storage capacity and transmission bandwidth have been increasing, so have the demand for and the resolution of digital images. Thus, there is an ever increasing need for image compression. In order to achieve high compression ratios, image compression techniques make use of the properties of the human visual system. The amount of compression that can be attained with lossless compression techniques is limited, typically to a factor of two (see Chapters 5.1 and 5.6). Most of the traditional lossy compression techniques make implicit use of the HVS characteristics to achieve much higher compression ratios (by a factor of eight and higher) without significantly sacrificing image quality. (Chapters 5.2–5.5, 5.7, and 6.1–6.5 discuss lossy compression techniques for still image and video compression.) Recently, however, a number of algorithms have appeared that make use of explicit perceptual models [5, 6, 7, 8, 9, 10, 11]. The idea is to make the distortions introduced by the compression scheme invisible to the eye, i.e. perceptually lossless. Typically, the signal is analyzed into components (e.g. spatial and/or temporal subbands), and the role of the perceptual model is to provide the maximum amount of distortion that can be introduced to each component without resulting in any perceived distortion. This is usually referred to as the *just noticeable distortion* level or *JND.* We will look closely at the metrics developed by Safranek and Johnston for subband coders [5], by Watson for DCT coders [6], and by Watson *et al.* for wavelet-based coders [12].

Most of the existing models for image quality and compression deal with the threshold of perception. In an increasing number of applications, however, there is a need to achieve very high compression ratios, and in such cases, a certain amount of perceived distortion is unavoidable. In supra-threshold image compression, i.e. when the coding distortions exceed the threshold of visibility, there is a need to derive quantitative objective measures of perceived distortion. In general, it is easier to obtain models for the perceptually transparent

case; it is much more difficult to quantify perceived distortion, especially across different types of artifacts.

The perceptual models we examine are based on properties of the visual system and measurements of the eye characteristics, e.g. the contrast sensitivity function (CSF), light adaptation, masking, etc. However, some models, especially those that are designed for specific applications can be obtained empirically (experimentally). The disadvantage of such models is that it is very difficult to adapt them to different conditions.

All of the metrics we examine share a similar basic structure, which includes a calibration stage, linear filters tuned to different spatial frequencies and orientations, contrast sensitivity adjustments, and nonlinear mechanisms that account for masking. A final stage involves error pooling to obtain, either a single number that describes image quality, or a map of distortions or detection probabilities. Simpler (linear) models like the SQRI (square root integral) model [13] have been quite successful but are limited to specific applications. Similarly, simple linear models of the HVS have been very successful in image halftoning [14, 15, 16, 17, 18].

The advantage of objective quality metrics is that they are relatively easy to use. However, they are no substitute for subjective evaluations which are accepted to be the most effective and reliable, albeit quite cumbersome and expensive, way to assess image quality. A significant effort has been dedicated for the development of subjective tests for the image quality. A considerable part of this effort has come from IPO, the Center for Research on User-System Interaction, in the Netherlands. Some recent contributions can be found in [19, 20]. There has also been standards activity on subjective evaluation of image quality [21]. The study of the topic of subjective evaluation of image quality is beyond the scope of this chapter.

In the following sections, we review the fundamentals of human perception and image quality, and consider different metrics for image quality based on models of human perception.

2 Fundamentals of Human Perception and Image Quality

In this section we will introduce some of the concepts from the psychophysics of human perception that apply to image and video quality metrics. A more in depth coverage of this topic can be found in [22] and Chapter 1.3.

2.1 Psychophysics of Vision

The psychophysics of vision is a vast topic full of interesting experiments and illusions. The HVS models that are used for quality measurement are based on the lower order processing of the visual system, i.e. the modeling of the function of the optics, retina, lateral geniculate nucleus, and striate cortex. Higher level processing, such as attentive vision, Gestalt, and figure/ground effects are, either too local in their effect, or not understood well enough to be effectively utilized.

The approach taken by most visual quality models is to determine how the lower level physiology of the visual system limits visual sensitivity. In general these models incorporate four types of processes that introduce sensitivity variations: light level, spatial frequency, signal content, and in the case of video, temporal variation. These limits are then converted to *masking thresholds* which determine the level of distortion to which an image can be exposed to before the alteration is apparent to a human observer. If the magnitude of the distortion is less than the masking threshold (also called the *Just Noticeable Distortion* or *JND* level), the original and distorted image will be indistinguishable.

Light level adaptation is caused primarily by the retina and is referred to as the *amplitude nonlinearity* of the visual system. Imperfect optics coupled with neural interactions produce a non-uniform frequency response that is called the *contrast sensitivity function.* Sensitivity variation due to signal content is due to post-receptor neural circuitry and gives rise to *masking* (also referred to as *texture masking*). Adaptation of neural state to an input signal gives rise to *temporal masking.* The next several sections will provide an introduction to each of these phenomena.

2.1.1 Amplitude Nonlinearity

It is well known that the perception of lightness is a nonlinear function of luminance. Consider the following experiment: create a series of images consisting of a background

of uniform intensity, I, each with a square of a different intensity, $I + \delta I$ inserted into its center. Show these to an observer in order of increasing δI. Ask the observer to determine the point at which they can first detect the square. Then, repeat this experiment for a large number of different values of background intensity. For a wide range of background intensities, the ratio of the threshold value δI divided by I is a constant. This equation

$$\frac{\delta I}{I} = k \tag{1}$$

is called *Weber's Law*. The value for k is roughly 0.33.

Most implementations of this aspect of visual sensitivity treat it as a point process using the value of a single pixel for the central square and the average of the surrounding pixels for the background. Various instantiations of the amplitude nonlinearity include [23, 24, 3, 25, 26].

2.1.2 Contrast Sensitivity Function

The human visual system's contrast sensitivity function (also called the modulation transfer function) provides a characterization of its frequency response. The contrast sensitivity function can be thought of as a bandpass filter. There have been several different classes of experiments used to determine its characteristics which are described in detail in [22, Ch. 12] and Chapter 1.3.

One of these methods involves the measurement of visibility thresholds of sine-wave gratings in a manner analogous to the experiment described in the previous section. For a fixed frequency, a set of stimuli consisting of sine waves of varying amplitudes are constructed. These stimuli are presented to an observer and the detection threshold for that frequency is determined. This procedure is repeated for a large number of grating frequencies. The resulting curve is called the contrast sensitivity function and is illustrated in Fig. 3.

Note that these experiments used sine-wave gratings at a single orientation. To fully characterize the contrast sensitivity function, the experiments would need to be repeated with gratings at various orientations. This has been accomplished and the results show that the HVS is not perfectly isotropic. However, for the purposes of general quality measurement, it is close enough to isotropic that that assumption is normally used.

It should also be noted that the spatial frequencies are in units of cycles per degree of

visual angle. This implies that the visibility of details at a particular frequency is a function of viewing distance. As an observer moves away from an image, a fixed size feature in the image takes up fewer degrees of visual angle. This action moves it to the right on the contrast sensitivity curve, possibly requiring it to have greater contrast to remain visible.

On the other hand moving closer to an image can allow previously imperceivable details to rise above the visibility threshold. Given these observations, it is clear that the minimum viewing distance is where distortion is maximally detectable. Therefore quality metrics need to specify a minimum viewing distance and evaluate its distortion metric at that point. Several "standard" minimum viewing distances have been established for subjective quality measurement and have generally been used with objective models as well. These are six times image height for standard definition television and three times image height for high definition television.

2.1.3 Contrast Masking

In the previous two sections we have dealt with stimuli that are either constant or contain a single frequency. In general, this is not characteristic of natural scenes. They have a wide range of frequency content over many different scales. Consider for a moment the following thought experiment: Consider two images, a constant intensity field and an image of a sand beach. Take a random noise process whose variance just exceeds the amplitude and contrast sensitivity thresholds for the flat field image. Add this noise field to both images. By definition, the noise will be detectable in the flat field image. However, it will not be detectable in the beach image. The presence of the multitude of frequency components in the beach image hides or *masks* the presence of the noise field.

Contrast masking refers to the reduction in visibility of one image component caused by the presence of another image component with similar spatial location and frequency content. As will be discussed further in a later section, the HVS can be thought of as a spatial frequency filter-bank with octave spacing of subbands in radial frequency, and angular bands of roughly 30 degree spacing. The presence of a signal component in one of these subbands will raise the detection threshold for other signal components in the same subband [27, 28, 29] and Chapter 1.3.

The base masking threshold for each spatial frequency band is determined by a combination of the amplitude nonlinearity and the contrast sensitivity function. For the DC band

(which corresponds to flat field images), this is the entire masking threshold. For the other bands, the amount by which the masking threshold is elevated is a nonlinear function of the energy in that band. Up to a point the masking threshold remains constant. Once that threshold is exceeded, the masking threshold rises with signal energy as shown in Fig. 4.

2.1.4 Temporal Masking

Temporal masking of time varying stimuli is extremely important in determining the quality of a video signal. This effect is complicated by the fact that the perception of a moving object depends heavily on whether or not the object is being tracked by the eye. Since the purpose of the metrics we will be discussing is to provide a quality measure for the *entire* image sequence, we will assume the worst case, i.e. all objects in a scene that can be tracked will be. There are two major forms of temporal masking that have been utilized in the literature; scene change and the temporal contrast sensitivity function.

In video, a *scene change* occurs when there is an abrupt change in the the content of the entire image. This large change in the overall visual field induces a dramatic increase in the masking levels for period of up to 100ms after the scene change [30, 31, 32, 33].

The temporal contrast sensitivity function can be thought of as an extension to the spatial contrast sensitivity function. Experiments similar to those used to measure the spatial sensitivity function have been performed by Kelly [34, 35, 36]. In this case the stimulus was a uniform disk whose intensity was varied sinusoidally in time. For a given temporal frequency, subjects were asked to adjust the amplitude of the sine wave so that it was just at the the threshold of visibility. The results of these experiments are shown in Fig. 5. They show that distortion with a period of 4 to 8 HZ is most visible. Combining these results with those from the spatial contrast sensitivity function provides a means of visualizing the HVS spatio-temporal transfer function. This surface is illustrated in Fig. 6.

3 Visual Models for Image Quality and Compression

In this section, we consider various metrics for image quality based on models of the human visual system. First, we will consider perceptual metrics that are intended for general applicability. Such metrics are quite elaborate, computationally intensive, and require careful calibration and parameter selection. Then, we will consider metrics that are designed specifically for image compression. Such metrics are simpler, more efficient computationally, and sometimes are designed with specific compression algorithms in mind. For example, they could apply to specific viewing conditions and display devices, and may be intended for specific classes of coders (e.g. block-based or wavelet coders). Finally, we briefly consider perceptual metrics for other applications, e.g. halftoning, displays, and printing.

3.1 Basics of Perceptual Image Quality Metrics

The goal of a perceptual metric for image quality is to determine the differences between two images that are visible to the human visual system. Usually one of the images is the reference which is considered to be "original," "perfect," or "uncorrupted." The second image has been modified or distorted in some sense. It is very difficult to evaluate the quality of an image without a reference. Thus, a more appropriate term would be image "fidelity" or "integrity," or alternatively, image "distortion" [37]. However, for historical reasons we adopt the term image quality.

In addition to the two digital images, an image quality metric requires a few other parameters, e.g. viewing distance, image size, display parameters, etc. The output is a number that represents the probability that a human eye can detect a difference in the two images or a number that quantifies the perceptual dissimilarity between the two images. Alternatively, the output of an image quality metric could be a map of detection probabilities or perceptual dissimilarity values.

The very first stage of an image quality metric is calibration.

Calibration

The array of numbers that represents an image may have come from a number of different devices and could have gone through different transformations (conversion to densities, gamma correction, etc.) before it is displayed for observation by a human eye. Many

quality metrics require that the input image values be converted to physical luminances[2] before they enter the HVS model. Alternatively, the quality metric could be designed for a specific set of conditions, and thus, all of the calibration could be incorporated in the model, but then, the model would have to be rederived for each new set of conditions.

Registration

Registration, i.e. the point-by-point correspondence between two images, is necessary for any quality metric to make any sense. Otherwise, we could arbitrarily modify the value of a metric by arbitrarily shifting one of the images. The shift does not change the images but changes the value of the metric.

Display Model

An accurate model of the display device is an essential part of any image quality metric (e.g. [18]), as the human visual system can only see what the display can reproduce. In some cases, when the perceptual model is obtained empirically (e.g. [5]), the effects of the display are incorporated in the model. The obvious disadvantage of this approach is that when the display changes, a new set of the model parameters must be obtained (e.g. [8]). The study of display models is beyond the scope of this chapter.

3.2 General Models

A number of different quality metrics have been proposed that are based on models of the low-level processing of the HVS, such as the optics, the retina, the lateral geniculate nucleus, and the striate cortex [2]. Such metrics are quite general and are intended for a variety of image processing applications such as compression, halftoning, display evaluation, etc.

We discuss the following models in detail: the "Visible Differences Predictor" by Daly [2], the model proposed by Lubin [3], and the "Perceptual Image Distortion" metric by Teo and Heeger [4]. All such models have a similar basic structure, shown in Fig. 1. The front end includes the calibration, display model, and registration.

[2]In video practice, the term luminance is sometimes, incorrectly, used to denote a nonlinear transformation of luminance [38, p. 24].

3.2.1 Frequency Analysis

The frequency analysis comprises a hierarchy of filters that decompose the image into several components or channels (usually called subbands) with different spatial frequencies and orientations. Some examples of such decompositions are shown in Fig. 14. The range of each axis is from -$u_s/2$ to $u_s/2$ cycles per degree, where u_s is the sampling frequency. Fig. 14(a) shows the Cortex transform that was proposed by Watson [39]. The Cortex transform consists of two classes of filters applied sequentially which decompose the image, first into different radial frequency bands, and then into different orientation bands. A variation of the Cortex transform was used by Daly [2] and is shown in Fig. 14(b). One of the differences is that Daly used six orientation bands to better approximate the orientation selectivity of the human visual system. The radial filters of both decompositions have octave bandwidths.

A similar decomposition was used by Lubin [3]. He used the Laplacian pyramid of Burt and Adelson [40] to decompose the image into seven radial frequency bands, and the steerable filters of Freeman and Adelson [41] to decompose each pyramid level into four different orientations. Finally, Teo and Heeger [4] adopted the steerable pyramid transform by Simoncelli *et al.* [42] which also has octave radial bandwidths and six orientation bands. In an earlier paper [43] they considered a hexagonally sampled quadrature mirror filter transform, and concluded that its orientation selectivity was not adequate for matching the HVS data.

Both Daly and Lubin convert the filtered images to units of contrast. Daly [2] proposes two alternatives: local contrast, which uses the value of the baseband at any given location to divide the values of all the other bands, and global contrast, which divides all subbands by the average value of the input image. Lubin [3] converts to local contrast by dividing each point in each level of the Laplacian pyramid by the corresponding point obtained from the Gaussian pyramid two levels down in resolution. The conversion to contrast is necessary for general models to account for contrast variations that are not visible. As we will see below, this is not necessary for specific applications such as image compression, where the contrast of the original and coded image remains basically unchanged.

In the remainder of this paper, we will use $f(\mathbf{n})$ to denote the value (intensity, grayscale, etc.) of an image pixel at location $\mathbf{n}$. Usually the image pixels are arranged in a Cartesian

grid and $\mathbf{n} = (n_1, n_2)$. The value of the $\mathbf{k}$-th image subband at location $\mathbf{n}$ will be denoted by $b(\mathbf{k}, \mathbf{n})$. The subband indexing $\mathbf{k} = (k_1, k_2)$ could be in Cartesian or polar or even scalar coordinates. The same notation will be used to denote the $\mathbf{k}$-th coefficient of the $\mathbf{n}$-th DCT block (both Cartesian coordinate systems). This notation underscores the similarity between the two transformations, even though we traditionally display the subband decomposition as a collection of subbands and the DCT as a collection of block transforms: A regrouping of coefficients in the blocks of the DCT results in a representation very similar to a subband decomposition. A more careful discussion of the relationship between the DCT, subband, and wavelet decompositions can be found in Chapter 5.4.

3.2.2 Baseline Contrast Sensitivity (Base Sensitivity)

The baseline contrast sensitivity determines the amount of energy in each subband that is required in order to detect the target in an (arbitrary or) flat mid-gray image. As we discussed earlier, this is sometimes referred to as the just noticeable difference or JND. We will use $t_b(\mathbf{k})$ to denote the baseline sensitivity of the $\mathbf{k}$-th band or DCT coefficient. Note that the base sensitivity is independent of the location $\mathbf{n}$.

The base sensitivities can be obtained from the contrast sensitivity function (CSF), as in [3], or can be derived empirically and listed in a table, as in [5]. The base sensitivities are then adjusted to account for variations in luminance and texture masking to obtain the overall sensitivity $t(\mathbf{k}, \mathbf{n})$. An alternative approach, implemented in Daly's model [2], is to filter the image by the contrast sensitivity function before the frequency decomposition.

The key parameters for the contrast sensitivity are the viewing distance (in inches) and the resolution of the display device (in pixels per inch). Alternatively, one can specify the viewing distance in image heights and the image height in pixels (assuming the same horizontal and vertical display resolution). In either case, one must derive the "display visual resolution" in pixels per degree [12].

Since the contrast sensitivity function has a band-pass characteristic (e.g. see [2]), if we assume a single fixed viewing distance, the metric may show a degradation in image quality as we move away from the image. To avoid this, one can assume a range of viewing distances [2], or a minimum viewing distance. This will result in a flattening of the CSF. This flattening is commonly assumed in image halftoning applications [14, 18] because it is the low-pass characteristic of the eye that is critical for halftoning.

3.2.3 Luminance Masking

Most of the perceptual models assume that the input image values have been converted to physical luminances. The human visual system's sensitivity to variations in luminance depends on (is a nonlinear function of) the local mean luminance. Some authors call this "light adaptation." Others prefer the term "luminance masking" (e.g. [6]), which groups it together with the other types of masking we will see below. It is called masking because the luminance of the original image signal masks the variations in the distortion signal. Thus, one approach to account for light adaptation or luminance masking is by including a modification of the base sensitivities. The luminance masking adjustment is a function of the local luminance (or gray level). A common simplification is to assume that it is independent of the subband index $\mathbf{k}$.

An alternative way to account for light adaptation is by including a nonlinear transformation before the frequency analysis. Commonly used transformations include conversion to density (log) and various power laws (e.g. cube root). Daly [2] includes a simple point-by-point amplitude nonlinearity that lies between these two. Daly even allows for an adaptation to absolute luminance levels. However, this is a second order effect.

For simplicity, and since most image compression algorithms work on gray levels that are gamma-corrected[3] luminances (or sometimes image densities), many applied models are calibrated for gamma-corrected input images [12]. In a well-calibrated display system, the gray levels represent relatively uniform steps in perceived lightness and the luminance masking component of the perceptual model does not have a significant effect (i.e. it is relatively flat as in [8]). The disadvantage of this approach is that it is difficult to account for variations in gamma correction.

3.2.4 Contrast/Texture Masking

Masking is the reduction in the visibility of one image component (the target) due to the presence of another (the masker). It is strongest when both components have the same or similar frequency, orientation, and location. Sometimes the term *contrast masking* is used to denote the case where both the target and the masker have the same frequency and orientation [43], and the term *texture masking* is used to refer to the more general case.

[3]Gamma correction is a nonlinear transformation (power law) of image luminance to compensate for the display characteristics, and at the same time, to obtain an efficient image representation [38, Ch. 6].

As we saw in Section 2, several experiments can be conducted to determine masking. For example, in the simplest case, both the signal and the mask may consist of single, possible different, frequencies. Alternatively, the two signals could be noise fields with different bandwidths, and in particular bandwidths that correspond to the components of the frequency decomposition of the visual model at hand.

Most of the metrics model contrast masking only. Watson [6] uses the following model for the contrast masking adjustment

$$\tau_c(\mathbf{k},\mathbf{n}) = \max\left\{1, |b(\mathbf{k},\mathbf{n})|^{w_c(\mathbf{k})} t_l(\mathbf{k},\mathbf{n})^{-w_c(\mathbf{k})}\right\} \tag{2}$$

where $t_l(\mathbf{k},\mathbf{n})$ is the base sensitivity threshold adjusted for luminance masking and $w_c(\mathbf{k})$ is a number between 0 and 1. The overall sensitivity threshold $t(\mathbf{k},\mathbf{n})$ is then equal to $\tau_c(\mathbf{k},\mathbf{n})\ t_l(\mathbf{k},\mathbf{n})$. A typical empirical value is $w_c(\mathbf{k}) = 0.7$ for $\mathbf{k} \neq 0$ and $w_c(\mathbf{0}) = 0$. Note that in the case of the DC coefficient, the contrast masking adjustment coincides with the luminance masking adjustment, which is applied separately. Lubin [3] uses a sigmoid[4] nonlinearity to account for contrast masking. Daly uses a similar nonlinearity for contrast masking [2]. He also allows for *mutual* masking which uses both the original and the distorted image to determine the degree of masking.

Safranek and Johnston [5] define as texture any deviation from a flat field within a subband and use the following texture masking adjustment:

$$\tau_t(\mathbf{k},\mathbf{n}) = \max\left\{1, \left[\sum_{\mathbf{k}} w_{MTF}(\mathbf{k}) e_t(\mathbf{k},\mathbf{n})\right]^{w_t}\right\} \tag{3}$$

where $e_t(\mathbf{k},\mathbf{n})$ is the "texture energy" of subband $\mathbf{k}$ at location $\mathbf{n}$, $w_{MTF}(\mathbf{k})$ is a weighting factor for subband $\mathbf{k}$ determined empirically from the MTF of the HVS, and w_t is a constant equal to 0.15. The subband texture energy is given by:

$$e_t(\mathbf{k},\mathbf{n}) = \begin{cases} \text{local variance}_{\mathbf{m}\in \mathrm{N}(\mathbf{n})}(b(\mathbf{0},\mathbf{m})), & \text{if } \mathbf{k} = \mathbf{0} \\ b(\mathbf{k},\mathbf{n})^2, & \text{otherwise} \end{cases} \tag{4}$$

where $\mathrm{N}(\mathbf{n})$ is the neighborhood of the point $\mathbf{n}$ over which the variance is calculated. In the

[4] A sigmoid function starts out flat, its slope increases to a maximum, and then decreases back to zero, i.e. it changes curvature like the letter S.

Safranek-Johnston model, the overall sensitivity threshold is the product of three terms

$$t(\mathbf{k},\mathbf{n}) = \tau_t(\mathbf{k},\mathbf{n})\,\tau_l(\mathbf{k},\mathbf{n})\,t_b(\mathbf{k}) \tag{5}$$

where $\tau_t(\mathbf{k},\mathbf{n})$ is the texture masking adjustment, $\tau_l(\mathbf{k},\mathbf{n})$ is the luminance masking adjustment, and $t_b(\mathbf{k})$ is the baseline sensitivity threshold. The Safranek-Johnston texture adjustment does not differentiate between random and structured texture. The contrast masking models used by Watson, Daly, and Lubin are more effective [44]. As we will see below, Teo and Heeger propose a more elaborate texture model that accounts for contrast masking as well as masking that occurs when the orientations of the target and the masker are different.

3.2.5 Error Pooling

The final step of an image quality metric is to combine the errors (normalized by the sensitivity thresholds, and therefore expressed in JNDs), that have been computed for each spatial frequency and orientation band and each spatial location, into a single number for each pixel of the image, or a single number for the whole image. Some metrics convert the JNDs to detection probabilities. Daly [2] converts to probabilities before error pooling, and then uses probability summation to obtain a spatial map of detection probabilities for each point of the image. Lubin [3] converts to probabilities after error pooling. Watson [6] simply expresses the metric in JNDs.

An example of error pooling is the following Minkowski metric

$$E(\mathbf{n}) = \frac{1}{M}\left\{\sum_{\mathbf{k}}\left|\frac{b(\mathbf{k},\mathbf{n}) - \hat{b}(\mathbf{k},\mathbf{n})}{t(\mathbf{k},\mathbf{n})}\right|^{Q}\right\}^{1/Q} \tag{6}$$

where $b_{\mathbf{k}}(\mathbf{n})$ and $\hat{b}_{\mathbf{k}}(\mathbf{n})$ are the $\mathbf{n}$-th element of the $\mathbf{k}$-th subband of the original and coded image, respectively, $t(\mathbf{k},\mathbf{n})$ is the corresponding sensitivity threshold, and M is the total number of subbands. In this case, the errors are pooled across frequency to obtain a distortion measure (expressed in JNDs) for each spatial location. The value of Q varies from 2 (energy summation) to infinity (maximum error). Lubin [3] uses $Q = 2.4$. Teo and Heeger [4] use $Q = 2$. Watson [6] pools over DCT blocks first to obtain a "perceptual error matrix," then pools across frequency to obtain a single valued perceptual error metric. He

considers different values of Q, possibly different for each stage.

In order to be consistent with traditional error metrics, Van den Branden Lambrecht and Verscheure [45] suggest expressing the perceptual metric in terms of "visual decibels." They define the "masked peak signal-to-noise ratio (MPSNR)" as

$$\text{MPSNR} = 10\, log_{10}\, \frac{255^2}{E^2} \tag{7}$$

where E is the value of the metric. In fact, Safranek and Johnston [5], even though they did not define a perceptual metric explicitly, also expressed the perceptual threshold of their coder as a PSNR in decibels.

3.2.6 The Daly and Lubin Models

Daly's model [2] was developed for the evaluation of high quality imaging systems. It is the most general and elaborate image quality metric. As we saw above, it accounts for variations in sensitivity due to light level, spatial frequency (CSF), and signal content (contrast masking).

Lubin's model [3] was developed for display evaluation. It is also quite general and elaborate and accounts for sensitivity variations due to spatial frequency and masking. In addition, it accounts for fixation depth and eccentricity of the images in the visual field.

We discussed most of the components of these models in the sections above. The main drawback of the metrics based on these models is that they are computationally intensive. Also, they are so elaborate and general that they are difficult to implement and to match to a given set of conditions.

3.2.7 The Teo and Heeger model

As we saw above, the Teo and Heeger metric [4] uses the steerable pyramid transform [42] which decomposes the image into several spatial frequency and orientation bands.[5] However, unlike the other two models we saw above, it does not attempt to separate the base sensitivity and the other masking effects. Instead, Teo and Heeger propose a *normalization model* that explains baseline contrast sensitivity, contrast masking, as well as masking that occurs when the orientations of the target and the masker are different. The normalization

[5] A more detailed discussion of this model, with a different transform, can be found in [43].

model has the following form:

$$R(\mathbf{k},\mathbf{n},i) = R(\rho,\theta,\mathbf{n},i) = \kappa_i \frac{[b(\rho,\theta,\mathbf{n})]^2}{\sum_\phi [b(\rho,\phi,\mathbf{n})]^2 + {\sigma_i}^2} \tag{8}$$

where $R(\mathbf{k},\mathbf{n},i)$ is the normalized response of a sensor corresponding to the transform coefficient $b(\rho,\theta,\mathbf{n})$, $\mathbf{k} = (\rho,\theta)$ specifies the spatial frequency and orientation of the band, $\mathbf{n}$ specifies the location, and i specifies one of four different contrast discrimination bands characterized by different scaling and saturation constants, κ_i and ${\sigma_i}^2$, respectively. The scaling and saturation constants κ_i and ${\sigma_i}^2$ are chosen to fit the experimental data of Foley and Boynton. Thus, this model is tailored to a specific set of conditions, and would require a lot more work to be adapted to a new set of conditions.

3.3 Coder-Specific Models

We now look at models that have been developed specifically for image compression applications. While still based on the properties of the HVS, these models adopt the frequency decomposition of a given coder, which is chosen to provide high compression efficiency as well as computational efficiency. They are considerably simpler than the general models we saw above, as they only have to consider the properties of the HVS that are relevant for this application. For example, as we saw above, since the contrast of the original and coded image is basically the same, there is no need to account for local contrast variations. Another difference is that frequency decompositions used for compression are usually critically sampled, which means that the number of samples of the original image and the frequency decomposition is the same. Critical sampling is advantageous for compression, but it is not necessary for image quality metrics with general applicability.

The block diagram of a generic perceptually based coder is shown in Fig. 2. The frequency analysis decomposes the image into several components (subbands, wavelets, etc.) which are then quantized and entropy coded. The frequency analysis and entropy coding are virtually lossless; the only losses occur at the quantization step. The perceptual masking model is based on the frequency analysis and regulates the quantization parameters to minimize the visibility of the errors.

The visual models can be incorporated in a compression scheme to minimize the visibility of the quantization errors, or they can be used independently to evaluate its performance.

While the coder-specific image quality metrics are quite effective in predicting the performance of the given coder, some of them may not be as effective in predicting performance across different coders [1].

3.3.1 The Safranek-Johnston Perceptual Image Coder (PIC) and Metric

One of the first image coders to incorporate an elaborate perceptual model was the Safranek-Johnston perceptual subband image coder (PIC) [5]. It uses an empirically derived perceptual masking model that was obtained for a given CRT display and viewing conditions (six times image height). The PIC coder has the basic structure shown in Fig. 2. It uses a separable generalized quadrature mirror filter (GQMF) bank for subband analysis/synthesis. The base band is coded with DPCM while all other subbands are coded with PCM. All subbands use uniform quantizers with sophisticated entropy coding. The perceptual model specifies the amount of noise that can be added to each subband of a given image so that the difference between the output image and the original is just noticeable. The model contains the following components: The base sensitivity determines the noise sensitivity in each subband given a flat mid-gray image and was obtained using subjective experiments. The results are listed in a table. The second component is a brightness adjustment. In general this would be a two dimensional lookup table (for each subband and gray value). Safranek and Johnston made the reasonable simplification that the brightness adjustment is the same for all subbands. The final component is the texture masking adjustment described in the previous subsection.

In contrast to the general models we saw above, the PIC coder uses a decomposition into subbands of equal spacing and bandwidth as shown in Fig. 14(d). This decomposition is very efficient for compressing high quality images with a lot of high-frequency content. It is also separable in Cartesian coordinates, which makes it a lot more efficient computationally than the decompositions used in the general models, which are separable in polar coordinates.

A simple metric based on the PIC coder can be defined as follows

$$E = \left\{ \frac{1}{N} \sum_{\mathbf{n},\mathbf{k}} \left[\frac{b(\mathbf{k},\mathbf{n}) - \hat{b}(\mathbf{k},\mathbf{n})}{t(\mathbf{k},\mathbf{n})} \right]^{Q} \right\}^{\frac{1}{Q}} \tag{9}$$

where $b_{\mathbf{k}}(\mathbf{n})$ and $\hat{b}_{\mathbf{k}}(\mathbf{n})$ are the $\mathbf{n}$-th element of the $\mathbf{k}$-th subband of the original and coded image, respectively, $t(\mathbf{k},\mathbf{n})$ is the corresponding perceptual threshold, and N is the number

of pixels in the image. A typical value for Q is 2. If the error pooling is done over the subband index $\mathbf{k}$ only, as in (6), we obtain a spatial map of perceptually weighted errors This map is downsampled by the number of subbands in each dimension. A full resolution map can also be obtained by doing the error pooling on the upsampled and filtered subbands.

Figs. 7–13 demonstrate the performance of the PIC metric. Fig. 7 shows an original 512×512 image. The gray-scale resolution is 8 bits/pixel. Fig. 8 shows the image coded with the SPIHT coder [46] at 0.52 bits/pixel; the PSNR is 33.3 DB. Fig. 8 shows the same image coded with the PIC coder [5] at the same rate. The PSNR is considerably lower at 29.4 DB. This is not surprising, as the SPIHT algorithm is designed to minimize the mean-squared error (MSE) and has no perceptual weighting. The PIC coder assumes a viewing distance of six image heights or 21 inches. Depending on the quality of reproduction (which is not known at the time this chapter is written), at a close viewing distance, the reader may see ringing near the edges of the PIC image. On the other hand, the SPIHT image has considerable blurring, especially on the wall near the left edge of the image. However, if the reader holds the image at the intended viewing distance (approximately at arm's length), the ringing disappears, and all that remains visible is the blurring of the SPIHT image. Figs. 10 and 11 show the corresponding perceptual distortion maps provided by the PIC metric. The resolution is 128×128 and the distortion increases with pixel brightness. Observe that the distortion is considerably higher for the SPIHT image. In particular, the metric picks up the blurring on the wall on the left. The perceptual PSNR (pooled over the whole image) is 46.8 DB for the SPIHT image and 49.5 DB for the PIC image, in contrast to the PSNR values. Fig. 12 shows the image coded with the standard JPEG algorithm at 0.52 bits/pixel and Fig. 13 shows the PIC metric. The PSNR is 30.5 DB and the perceptual PSNR is 47.9 DB. At the intended viewing distance, the quality of the JPEG image is higher than the SPIHT image and worse than the PIC image as the metric indicates. Note that the quantization matrix provides some perceptual waiting which explains why the SPIHT image is superior according to PSNR and inferior according to perceptual PSNR.

For the PIC coder and metric we used a 4×4 subband decomposition. We used the base sensitivity thresholds and texture masking adjustment but no brightness adjustment, i.e. we assumed that the gray levels represent relatively uniform steps in perceived lightness. We are hoping the this turns out to be the case when the book is printed. Even though the metric is matched to the PIC coder, we believe that the above examples illustrate the power

of image quality metrics. In [1], Pappas *et al.* tested various metrics on different coders over many coding rates and several images, and found that the PIC metric provides better correlation with subjective evaluations than the MSE metric, the MTF weighted MSE, and Watson's DCT-based metric that we discuss below.

3.3.2 Watson's DCT-based Metric

Many current compression standards are based on a discrete cosine transform (DCT) decomposition. Watson [6] presented a model that computes the visibility thresholds for the DCT coefficients, and thus provides a metric for image quality. Watson's model was developed as a means to compute the perceptually optimal image dependent quantization matrix for DCT-based image coders like JPEG. It has also been used to further optimize JPEG-compatible coders [7, 44, 9]. The JPEG compression standard is discussed in Chapter 5.5.

The DCT decomposition is similar to the subband decomposition and is shown in Fig. 14(f). However, the filters are quite different and the coding artifacts that DCT-based coders produce are very different from those of subband coders. The DCT decomposition is separable and computationally efficient. Because of the popularity of DCT-based coders and computational efficiency of the DCT, we will give a more detailed overview of Watson's DCT-based perceptual model and its implementation, and how it can be used to obtain a metric of image quality.

The first step in the implementation is to convert the original reference and degraded images into a luminance/chrominance color space such as YC_rC_b. The luminance components are then partitioned into 8×8 pixel blocks and transformed to the frequency domain using the forward DCT. Next, the perceptual thresholds are computed from the DCT coefficients of the original image. These thresholds are computed for each block of data from the image. For each coefficient $b(\mathbf{k}, \mathbf{n})$, where $\mathbf{k}$ identifies the DCT coefficient and $\mathbf{n}$ denotes the block within the reference image, we will compute a threshold $t(\mathbf{k}, \mathbf{n})$ which accounts for the contrast sensitivity, luminance masking, and contrast masking.

The baseline contrast sensitivity thresholds $t_b(\mathbf{k})$ are determined by the Peterson, Ahumada, Watson method [47]. The threshold matrices for viewing distance equal to six image heights are provided in Table 1. Note that the quantization matrices can be obtained from the threshold matrices by multiplying by 2. These baseline thresholds are then modified to account, first for luminance masking, and than for contrast masking, in order to obtain the

overall sensitivity thresholds.

Since luminance masking is a function of only the average value of a region, it depends only on the DC coefficient $b(\mathbf{0}, \mathbf{n})$ of each DCT block. The luminance-masked threshold is given by

$$t_l(\mathbf{k}, \mathbf{n}) = t_b(\mathbf{k}) \left(\frac{b(\mathbf{0}, \mathbf{n})}{\bar{b}(\mathbf{0})} \right)^{a_T} \tag{10}$$

where $\bar{b}(\mathbf{0})$ is the DC coefficient corresponding to average luminance of the display (1024 for an 8 bit image using a JPEG compliant DCT implementation) and a_T has a suggested value of 0.649. This parameter controls the amount of luminance masking that takes place. Setting it to zero turns off luminance masking. Note that, since this is a power law, the effect of non-unity display Gamma can be accounted for by multiplying a_T by the Gamma exponent.

The Watson model of contrast masking assumes that the visibility reduction is confined to each coefficient in each block. As we saw in Section 3.2.4, the contrast masking adjustment $\tau_c(\mathbf{k}, \mathbf{n})$ is a function of the coefficient $b(\mathbf{k}, \mathbf{n})$ and the luminance-masked threshold $t_l(\mathbf{k}, \mathbf{n})$

$$\tau_c(\mathbf{k}, \mathbf{n}) = \max \left\{ 1, |b(\mathbf{k}, \mathbf{n})|^{w_c(\mathbf{k})} t_l(\mathbf{k}, \mathbf{n})^{-w_c(\mathbf{k})} \right\} \tag{11}$$

where $w_c(\mathbf{k})$ has values between 0 and 1. The exponent may be different for each frequency, but is typically set to a constant in the neighborhood of 0.7. If $w_c(\mathbf{k})$ is 0, no contrast masking occurs and the contrast masking adjustment is equal to 1. The overall sensitivity threshold is given by

$$t(\mathbf{k}, \mathbf{n}) = \tau_c(\mathbf{k}, \mathbf{n})\, t_l(\mathbf{k}, \mathbf{n}) \tag{12}$$

At this point, for each coefficient in each block we have a distortion visibility threshold. For each coefficient, we now need to determine the amount of distortion in terms of JND units. This is done by computing the error at each location (the difference between the DCT coefficients in the original and distorted images) weighted by the visibility threshold

$$d(\mathbf{k}, \mathbf{n}) = \frac{b(\mathbf{k}, \mathbf{n}) - \hat{b}(\mathbf{k}, \mathbf{n})}{t(\mathbf{k}, \mathbf{n})} \tag{13}$$

where $b(\mathbf{k}, \mathbf{n})$ and $\hat{b}(\mathbf{k}, \mathbf{n})$ are the reference and distorted images, respectively. Note that $d(\mathbf{k}, \mathbf{n}) < 1$ implies the distortion at that location is not visible, while $d(\mathbf{k}, \mathbf{n}) > 1$ implies the distortion is visible.

At this point, we have an array of distortion visibilities. These need to be combined into a single value denoting the quality of the image. This combination process is performed in two steps. First, error pooling is done spatially. Then the pools of spatial errors are pooled across frequency. Both pooling processes utilize the same probability summation framework.

$$p(\mathbf{k}) = \left\{ \sum_{\mathbf{n}} |d(\mathbf{k}, \mathbf{n})|^{Q_s} \right\}^{Q_s} \tag{14}$$

From psychophysical experiments, a value of 4 has been observed to be a good choice for Q_s.

The matrix $p(\mathbf{k})$ provides a measure of the degree of visibility of artifacts at each frequency. In order to determine a single valued metric for perceptual error, we need to pool these visibility measurements. This is accomplished using a procedure similar to the spatial pooling.

$$P = \left\{ \sum_{\mathbf{k}} p(\mathbf{k})^{Q_f} \right\}^{Q_f} \tag{15}$$

Q_f again can have many values depending on if average or worst case error is more important. Low values emphasize average error, while setting Q_f to infinity reduces the summation to a maximum operator thus emphasizing worst case error.

Alternative means of representing image quality is to use a distortion map instead of a single number. A distortion map can be computed by not performing the spatial error pooling. By performing the frequency pooling on each block independently, and treating the result as an image, the distribution of error across the image is maintained. This approach has the advantage of providing a visual indication of quality. For example, using visibility maps it is easy to distinguish between an image where the peak distortion was confined to a single DCT block and one where that peak level appears in a large percentage of blocks, but both would have the same quality indicator if both the spatial and frequency pooling were performed.

Quality measures for the chrominance channels can be computed in a similar fashion turning off the luminance and contrast masking portions of the model.

This metric has been shown to be very effective in predicting the performance of block-based coders. However, it is not as effective in predicting performance across different coders. In [1], Pappas *et al.* tested this and other metrics on different coders over many coding rates and several images, and found that the metric predictions (they used $Q_f =$

$Q_s = 2$) are not always consistent with subjective evaluations when comparing different coders. They found that this metric is strongly biased towards the JPEG algorithm. This is not surprising since both the metric and JPEG are based on DCTs.

3.3.3 Watson's Wavelet Metric

Many recent coders are based on the discrete wavelet transform (DWT). Popular examples are Shapiro's EZW algorithm [48] and the SPIHT algorithm [46]. A more detailed description of wavelet-based compression algorithms can be found in Chapter 5.4. The DWT is a separable, hierarchical subband decomposition. It has octave bandwidths in each dimension as shown in Fig. 14(e). It is also computationally efficient.

Watson *et al.* [12] measured the baseline sensitivity thresholds for the wavelet decomposition. They used the linear-phase 9/7 biorthogonal filters [49]. Note that the threshold values depend on the filterbank that is used for the wavelet decomposition. Even though [12] does not provide any detailed light adaptation and texture masking models, models such as those presented earlier in this chapter can be combined with the baseline sensitivity thresholds to obtain a perceptual image quality metric similar to the PIC and DCT-based metrics we discussed above.

3.3.4 Other Models

A number of other perceptual models that share many of the features of the ones we discussed above have been proposed. We mention some of them briefly. Silverstein and Klein [50] applied a DCT perceptual metric to a text-based scheme for image display. Westen *et al.* [51]. Horowitz and Neuhoff [52] developed a coder based on an image indistinguishability criterion based on a cortex transform similar to that of Watson [39]. Finally, some metrics have been designed to specific types of artifacts, e.g. blocking [53, 54], and [55].

In all of the models in this section the threshold values were determined by subjective experiments. Such experiments are time consuming. In [56], Hahn and Mathews present an analytical method for estimating the baseline contrast sensitivity values as well as the luminance and contrast masking correction factors.

3.4 Other Applications

We briefly mention image quality models that have been developed for other specific applications. In such cases the models do not have to be as complicated and can be very successful.

A good example of such a model is Barten's SQRI (square root integral) model [13] that was developed for the evaluation of displays. It is a simple one dimensional filter that has been very successful. However, it cannot be adapted to more general and complex applications.

An other application area where simple linear models of the HVS have proven to be quite successful is image halftoning [14, 15, 16, 17, 18]. Recently, Qian and Kimia have proposed more complicated multiscale models of the HVS for halftoning [57].

3.5 Video Quality Metrics

The basic principles we discussed above for still image metrics can be extended to video. The basic structure of the video quality metric is basically the same as that of Fig. 1. The frequency analysis decomposes the signal into channels with different spatial frequencies, orientations, and temporal frequencies. The baseline spatio-temporal contrast sensitivities are incorporated into the metric and modified by the luminance and contrast masking adjustments. Van den Branden Lambrecht and Verscheure [45] describe such a video quality metric. They use four spatial frequency bands with octave bandwidths, four orientation bands, and two temporal frequency bands. They compute a global metric that is a combination (Minkowski summation) over image blocks whose spatial dimensions depend on the focus of attention (size of the fovea) and temporal dimension depends on the persistence of images on the retina. They also segment the image into uniform regions, regions of texture, and regions of contours, and compute distortion metrics for each region type. Lindh and Van den Branden Lambrecht [58] developed a similar metric that extends Teo and Heeger's still image metric [4].

The above video quality metrics are examples of metrics of general applicability like the still metrics we saw above. Watson [59] proposed a coder-specific video quality metric that is based on the discrete cosine transform (DCT). It is an extension of Watson's DCT-based still image metric [6] and is designed with computational efficiency in mind. For temporal

filtering, the metric uses a first-order discrete IIR low-pass filter to minimize the number of frames that must be stored in memory. Watson *et al.* [60] evaluated the metric on DCT-based video coders and showed that the metric results correlate well with subjective evaluations.

Finally, Rohaly *et al.* [61] used a software version of a commercial video metric and compared various temporal pooling strategies. The metric is based on the Lubin and Bergen model [3]. The temporal pooling methods they included an exponentially weighted Minkowski summation to model the recency effect described in [19, 62].

4 Conclusions

In this chapter we presented objective criteria for the evaluation of image quality that are based on models of visual perception. We considered two major classes of models for image quality: general models and models that designed specifically for image compression applications. The general models are more elaborate and computationally intensive. The coder-specific models are simpler and computationally efficient. The coder-specific models have been shown to be quite effective when used to predict the performance of the given coder. However, the real challenging task is to predict the performance of coders with different structure, i.e. wavelet vs. DCT-based coders. The general models are expected to be more effective in this case.

Most of the existing models for image quality and compression are designed for applications where the distortions are near the threshold of perception. When the distortions exceed the threshold of visibility, it is considerably more difficult to derive quantitative measures of perceived distortion, especially across different types of artifacts. The models we discussed in this chapter can be used is such cases but their predictions of subjective image quality may not be as accurate.

As the demand for high quality still and video images increases, so does the need for their efficient storage and transmission. The existence of reliable and efficient metrics for image quality is critical in the development of algorithms that maximize the use of the available transmission bandwidth and storage capacity and produce the highest quality images. We have shown that an understanding of the properties of human visual system is critical in the development of such metrics. In recent years there has been a lot of activity in this field and we have attempted to summarize it in this chapter. However, it is still a very active and exciting field. The basic principles and ideas we discussed in this chapter should be a valuable tool in following its evolution.

References

[1] T. N. Pappas, T. A. Michel, and R. O. Hinds, "Supra-threshold perceptual image coding," in *Proc. ICIP-96, vol. I*, (Lausanne, Switzerland), pp. 237–240, Sept. 1996.

[2] S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision* (A. B. Watson, ed.), pp. 179–206, Cambridge, MA: The MIT Press, 1993.

[3] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision* (A. B. Watson, ed.), pp. 163–178, Cambridge, MA: The MIT Press, 1993.

[4] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. ICIP-94, vol. II*, (Austin, TX), pp. 982–986, Nov. 1994.

[5] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Proc. ICASSP-89*, vol. 3, (Glasgow, Scotland), pp. 1945–1948, May 1989.

[6] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Human Vision, Visual Proc., and Digital Display IV* (J. P. Allebach and B. E. Rogowitz, eds.), vol. Proc. SPIE, Vol. 1913, (San Jose, CA), pp. 202–216, Feb. 1993.

[7] R. J. Safranek, "A JPEG compliant encoder utilizing perceptually based quantization," in *Human Vision, Visual Proc., and Digital Display V* (B. E. Rogowitz and J. P. Allebach, eds.), vol. Proc. SPIE, Vol. 2179, (San Jose, CA), pp. 117–126, Feb. 1994.

[8] D. L. Neuhoff and T. N. Pappas, "Perceptual coding of images for halftone display," *IEEE Trans. Image Proc.*, vol. 3, pp. 341–354, July 1994.

[9] R. Rosenholtz and A. B. Watson, "Perceptual adaptive JPEG coding," in *Proc. ICIP-96, vol. I*, (Lausanne, Switzerland), pp. 901–904, Sept. 1996.

[10] I. Höntsch and L. J. Karam, "Apic: Adaptive perceptual image coding based on subband decomposition with locally adaptive perceptual weighting," in *Proc. ICIP-97, vol. I*, (Santa Barbara, CA), pp. 37–40, Oct. 1997.

[11] I. Höntsch, L. J. Karam, and R. J. Safranek, "A perceptually tuned embedded zerotree image coder," in *Proc. ICIP-97, vol. I*, (Santa Barbara, CA), pp. 41–44, Oct. 1997.

[12] A. B. Watson, G. Y. Yang, J. A. Solomon, and J. Villasenor, "Visibility of wavelet quantization noise," *IEEE Trans. Image Proc.*, vol. 6, pp. 1164–1175, Aug. 1997.

[13] P. G. J. Barten, "The SQRI method: A new method for the evaluation of visible resolution on a display," in *Proc. Society for Information Display*, vol. 28, pp. 253–262, 1987.

[14] J. Sullivan, L. Ray, and R. Miller, "Design of minimum visual modulation halftone patterns," *IEEE Trans. Sys., Man, Cyb.*, vol. 21, pp. 33–38, Jan./Feb. 1991.

[15] M. Analoui and J. P. Allebach, "Model based halftoning using direct binary search," in *Human Vision, Visual Proc., and Digital Display III* (B. E. Rogowitz, ed.), vol. Proc. SPIE, Vol. 1666, (San Jose, CA), pp. 96–108, Feb. 1992.

[16] J. B. Mulligan and A. J. Ahumada, Jr., "Principled halftoning based on models of human vision," in *Human Vision, Visual Proc., and Digital Display III* (B. E. Rogowitz, ed.), vol. Proc. SPIE, Vol. 1666, (San Jose, CA), pp. 109–121, Feb. 1992.

[17] T. N. Pappas and D. L. Neuhoff, "Least-squares model-based halftoning," in *Human Vision, Visual Proc., and Digital Display III* (B. E. Rogowitz, ed.), vol. Proc. SPIE, Vol. 1666, (San Jose, CA), pp. 165–176, Feb. 1992.

[18] T. N. Pappas and D. L. Neuhoff, "Least-squares model-based halftoning," *IEEE Trans. Image Proc.* to appear.

[19] R. Hamberg and H. de Ridder, "Continuous assessment of time-varying image quality," in *Human Vision and Electronic Imaging II* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3016, (San Jose, CA), pp. 248–259, Feb. 1997.

[20] H. de Ridder, "Psychophysical evaluation of image quality: from judgement to impression," in *Human Vision and Electronic Imaging III* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3299, (San Jose, CA), pp. 252–263, Jan. 1998.

[21] "ITU/R Recomendation BT.500-7, 10/1995," Internet address http://www.itu.ch.

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

[23] C. F. Hall and E. L. Hall, "A nonlinear model for the spatial characteristics of the human visual system," *IEEE Trans. Sys., Man, and Cyber.*, vol. SMC-7, pp. 162–170, Mar. 1977.

[24] T. J. Stockham, "Image processing in the context of a visual model," *Proc. IEEE*, vol. 60, pp. 828–842, July 1972.

[25] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 525–536, July 1974.

[26] J. J. McCann, S. P. McKee, and T. H. Taylor, "Quantitative studies in the retinex theory," in *Vision Research*, vol. 16, pp. 445–458, 1976.

[27] J. G. Robson and N. Graham, "Probability summation and regional variation in constrast sensitivity across the visual field," in *Vision Research*, vol. 21, pp. 419–418, 1981.

[28] G. E. Legge and J. M. Foley, "Contrast masking in human vision," in *Journal of the Optical Society of America*, vol. 70.

[29] G. E. Legge, "A power law for constrast discrimination," in *Vision Research*, vol. 21, pp. 457–467, 1981.

[30] B. G. Breitmeyer, *Visual Masking: An Integrative Approach.* New York: Oxford University Press, 1984.

[31] A. J. Seyler and Z. L. Budrikas, "Detail perception after scene change in television image presentations," in *IEEE Transactions on Information Theory*, vol. IT-11.

[32] Y. Ninomiya, T. Fujio, and F. Namimoto, "Perception of impairment by bit reduction on cut-changes in television pictures. (in japanese)," in *Electrical Communication Association Essay Periodical*, vol. J62-B.

[33] W. J. Tam, L. Stelmach, L. Wang, D. Lauzon, and P. Gray, "Visual masking at video scene cuts," in *Proceedings of the SPIE Conference on Human Vision, Visual Processing and Digital Display VI* (B. E. Rogowitz and J. P. Allebach, eds.), vol. Proc. SPIE, Vol. 2411, (San Jose, CA), pp. 111–119, Feb. 1995.

[34] D. H. Kelly, "Visual response to time-dependent stimuli," in *Journal of the Optical Society of America*, vol. 51, pp. 422–429, 1961.

[35] D. H. Kelly, "Flicker fusion and harmonic analysis," in *Journal of the Optical Society of America*, vol. 51, pp. 917–918, 1961.

[36] D. H. Kelly, "Flickering patterns and lateral inhibition," in *Journal of the Optical Society of America*, vol. 59, pp. 1361–1370, 1961.

[37] D. A. Silverstein and J. E. Farrell, "The relationship between image fidelity and image quality," in *Proc. ICIP-96, vol. II*, (Lausanne, Switzerland), pp. 881–884, Sept. 1996.

[38] C. A. Poynton, *A Technical Introduction to Digital Video*. New York: Wiley, 1996.

[39] A. B. Watson, "The cortex transform: Rapid computation of simulated neural images," *Computer Vision, Graphics, and Image Processing*, vol. 39, pp. 311–327, 1987.

[40] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Comm.*, vol. 31, pp. 532–540, 1983.

[41] W. T. Freeman and E. H. Adelson, "The design and use of steerable filters," *IEEE Trans. Pattern Analysis and Machine Intel.*, vol. 13, pp. 891–906, Sept. 1991.

[42] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Trans. Inf. Theory.*, vol. 38, pp. 587–607, Mar. 1992.

[43] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Human Vision, Visual Proc., and Digital Display V* (B. E. Rogowitz and J. P. Allebach, eds.), vol. Proc. SPIE, Vol. 2179, (San Jose, CA), pp. 127–141, Feb. 1994.

[44] R. J. Safranek, "A comparison of the coding efficiency of perceptual models," in *Proc. SPIE, vol. 2411, Human Vision, Visual Proc., and Digital Display VI*, (San Jose, CA), Feb. 1995.

[45] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Digital Video Compression: Algorithms and Technologies* (F. S. Vasudev Bhaskaran and S. Panchanathan, eds.), vol. Proc. SPIE, Vol. 2668, (San Jose, CA), pp. 450–461, Jan./Feb. 1996.

[46] A. Said and W. A. Pearlman, "A new fast and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Tech.*, vol. 6, pp. 243–250, June 1996.

[47] H. A. Peterson, A. J. Ahumada, Jr., and A. B. Watson, "An improved detection model for DCT coefficient quantization," in *Human Vision, Visual Proc., and Digital Display IV* (J. P. Allebach and B. E. Rogowitz, eds.), vol. Proc. SPIE, Vol. 1913, (San Jose, CA), pp. 191–201, Feb. 1993.

[48] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Proc.*, vol. SP-41, pp. 3445–3462, Dec. 1993.

[49] A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 45, pp. 485–560, 1992.

[50] D. A. Silverstein and S. A. Klein, "A DCT image fidelity metric for application to a text-based scheme for image display," in *Human Vision, Visual Proc., and Digital Display IV* (J. P. Allebach and B. E. Rogowitz, eds.), vol. Proc. SPIE, Vol. 1913, (San Jose, CA), pp. 229–239, Feb. 1993.

[51] S. J. P. Westen, R. L. Lagendijk, and J. Biemond, "Perceptual image quality based on a multiple channel hvs model," in *Proc. ICASSP-95, vol. 4*, (Detroit, MI), pp. 2351–2354, May 1995.

[52] M. J. Horowitz and D. L. Neuhoff, "Image coding by perceptual pruning with a cortical snapshot indistinguishability criterion," in *Human Vision and Electronic Imaging III* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3299, (San Jose, CA), pp. 330–339, Jan. 1998.

[53] C. Fenimore, B. Field, and C. V. Degrift, "Test patterns and quality metrics for digital video compression," in *Human Vision and Electronic Imaging II* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3016, (San Jose, CA), pp. 269–276, Feb. 1997.

[54] J. M. Libert and C. Fenimore, "Visibility thresholds for compression-induced image blocking: measurement and models," in *Human Vision and Electronic Imaging IV* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3644, (San Jose, CA), Jan. 1999.

[55] E. M. Yeh, A. C. Kokaram, and N. G. Kingsbury, "A perceptual distortion measure for edge-like artifacts in image sequences," in *Human Vision and Electronic Imaging III* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3299, (San Jose, CA), pp. 160–172, Jan. 1998.

[56] P. J. Hahn and V. J. Mathews, "An analytical model of the perceptual threshold function for multichannel image compression," in *Proc. ICIP-98, vol. III*, (Chicago, IL), pp. 404–408, Oct. 1998.

[57] W. Qian and B. Kimia, "On the perceptual notion of scale for halftone representations: Nonlinear diffusion," in *Human Vision and Electronic Imaging* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3299, (San Jose, CA), pp. 473–481, Jan. 1998.

[58] P. Lindh and C. J. van den Branden Lambrecht, "Efficient spatio-temporal decomposition for perceptual processing of video sequences," in *Proc. ICIP-96, vol. III*, (Lausanne, Switzerland), pp. 331–334, Sept. 1996.

[59] A. B. Watson, "Toward a perceptual video quality metric," in *Human Vision and Electronic Imaging III* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3299, (San Jose, CA), pp. 139–147, Jan. 1998.

[60] A. B. Watson, J. Hu, J. F. McGowan, and J. B. Mulligan, "Design and performance of a digital video quality metric," in *Human Vision and Electronic Imaging IV* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3644, (San Jose, CA), Jan. 1999.

[61] A. M. Rohaly, J. Lu, N. R. Franzen, and M. K. Ravel, "Comparison of temporal pooling methods for estimating the quality of complex video sequences," in *Human Vision and Electronic Imaging IV* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3644, (San Jose, CA), Jan. 1999.

[62] D. Pearson, "Viewer response to time-varying video quality," in *Human Vision and Electronic Imaging III* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE, Vol. 3299, (San Jose, CA), pp. 16–25, Jan. 1998.

Y Channel Thresholds

5	3	4	7	11	16	24	34
3	4	4	6	8	12	18	25
4	4	8	9	11	15	20	28
7	6	9	14	16	20	26	33
11	8	11	16	26	28	34	42
16	12	15	20	28	41	46	54
24	18	20	26	34	46	63	71
34	25	28	33	42	54	71	95

C_r Channel Thresholds

7	9	18	28	43	63	91	128
9	8	17	23	33	48	68	94
18	17	31	34	43	58	78	105
28	23	34	55	63	77	98	126
43	33	43	63	98	108	128	157
63	48	58	77	108	154	174	204
91	68	78	98	128	174	239	255
128	94	105	126	157	204	255	255

C_b Channel Thresholds

14	18	46	71	109	161	232	255
18	17	43	59	85	122	173	240
46	43	80	87	111	148	200	255
71	59	87	142	160	196	249	255
109	85	111	160	251	255	255	255
161	122	148	196	255	255	255	255
232	173	200	249	255	255	255	255
255	240	255	255	255	255	255	255

Table 1: Perceptual threshold matrices for DCT-based coders (for viewing distance equal to six image heights)

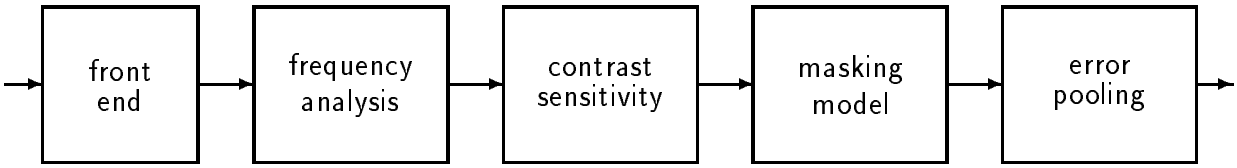


Figure 1: Perceptual metric

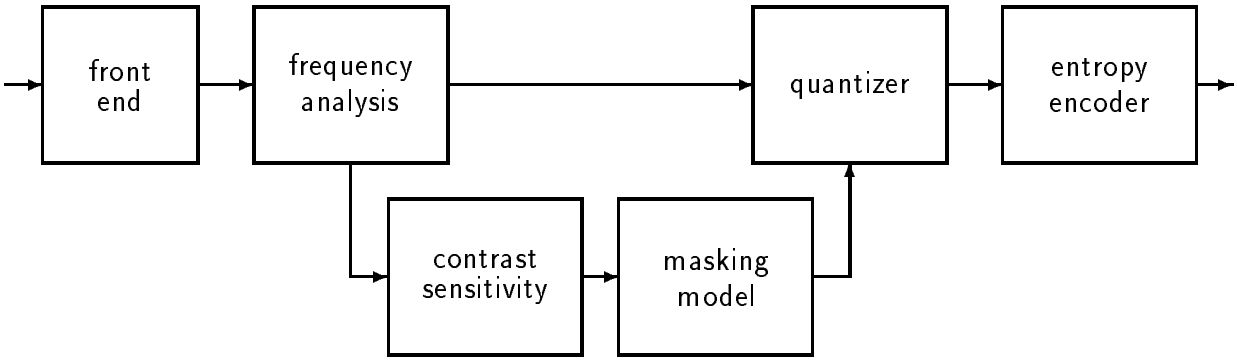


Figure 2: Perceptual coder

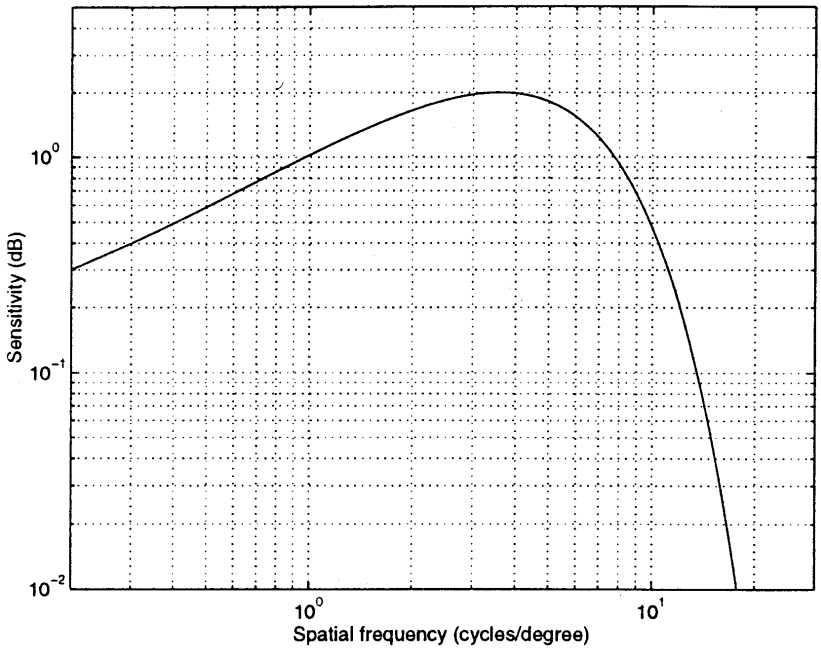


Figure 3: Spatial contrast sensitivity function

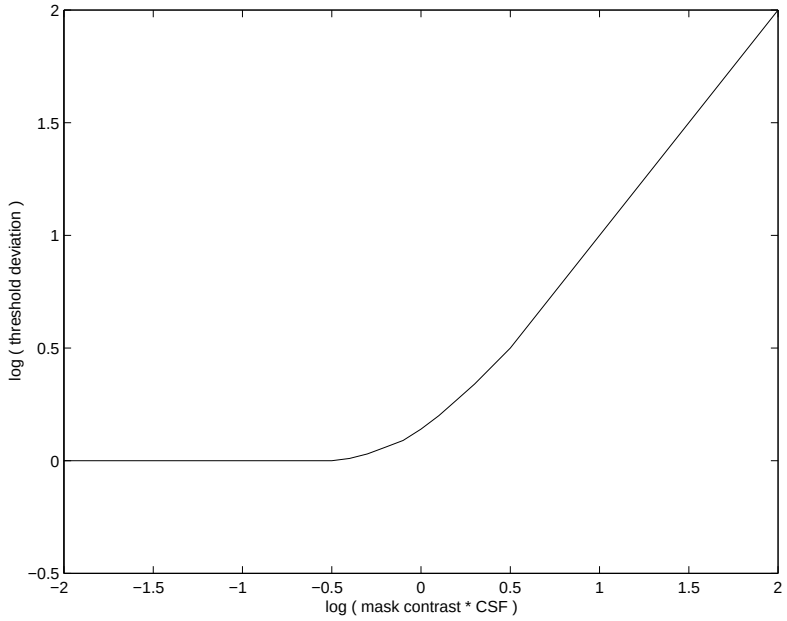


Figure 4: Contrast masking function

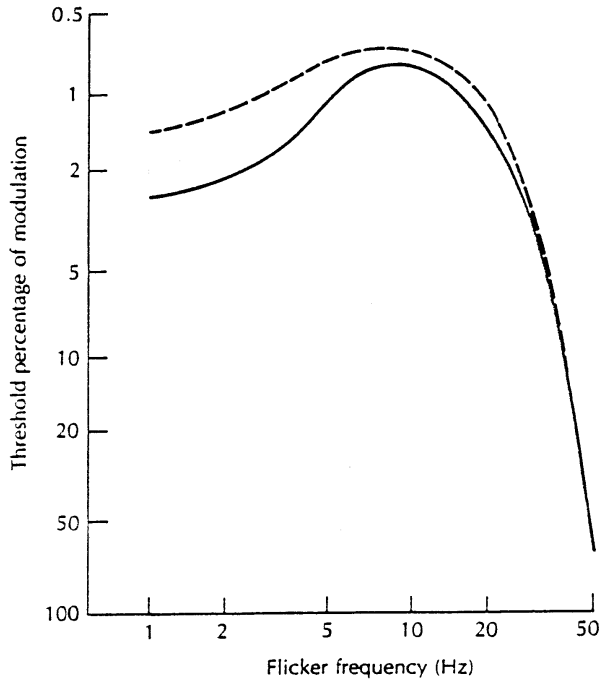


Figure 5: Temporal contrast sensitivity function (from [Kelly, 1969])

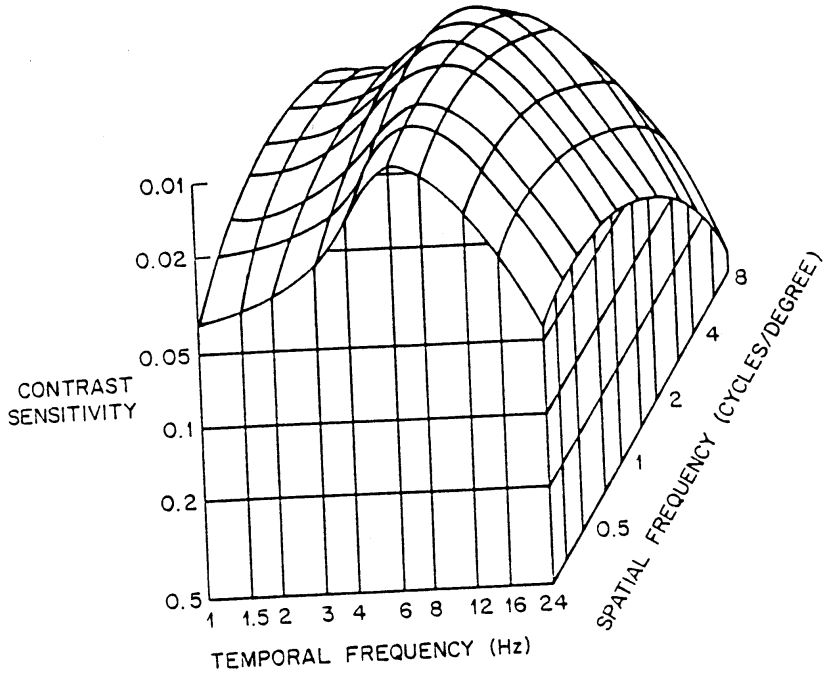


Figure 6: Spatio-temporal contrast sensitivity function

Figure 7: Original 512×512 image

Figure 8: SPIHT coder at 0.52 bits/pixel, PSNR = 33.3 DB

Figure 9: PIC coder at 0.52 bits/pixel, PSNR = 29.4 DB

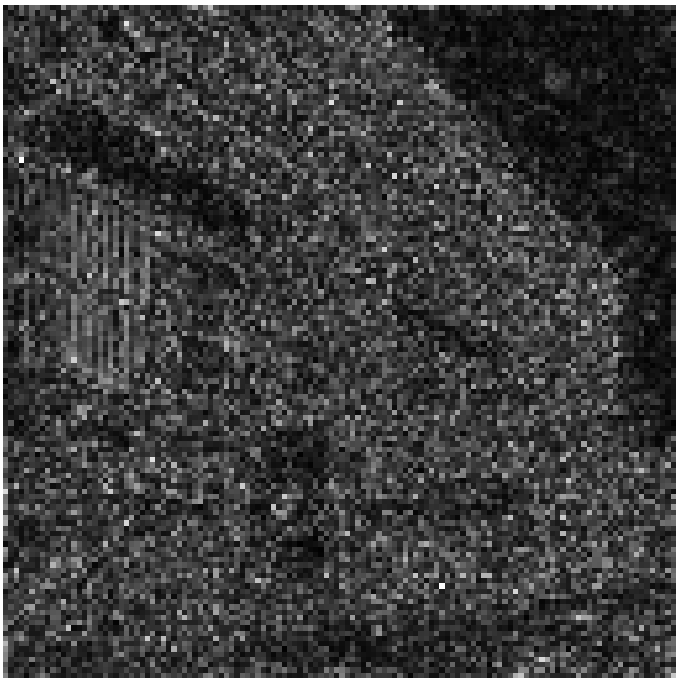


Figure 10: PIC metric for SPIHT coder, perceptual PSNR = 46.8 DB

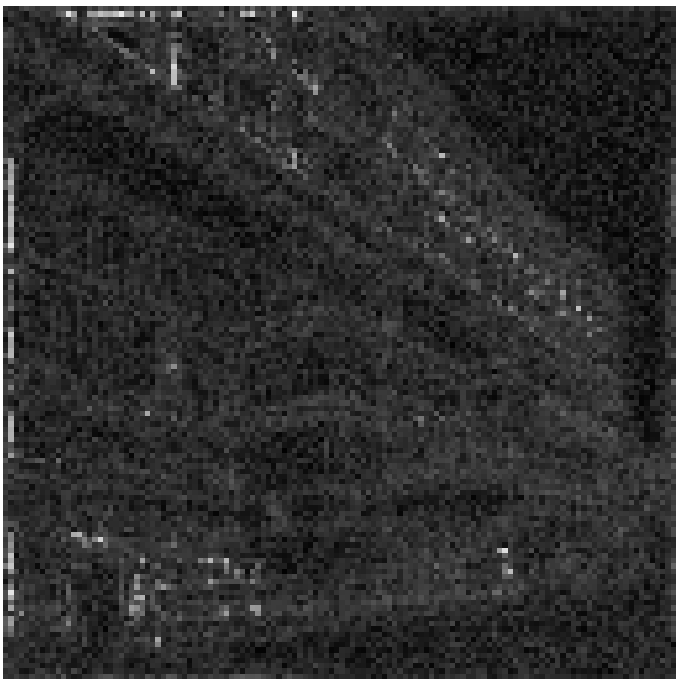


Figure 11: PIC metric for PIC coder, perceptual PSNR = 49.5 DB

Figure 12: JPEG coder at 0.52 bits/pixel, PSNR = 30.5 DB

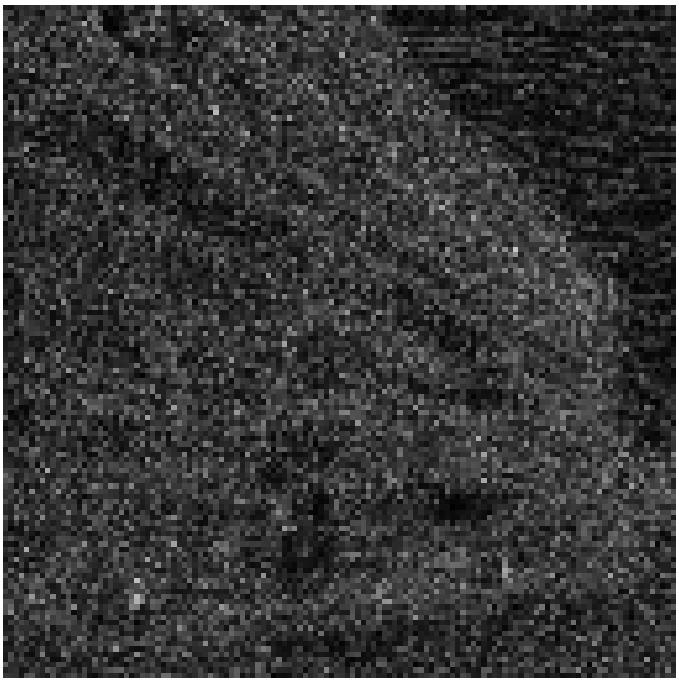
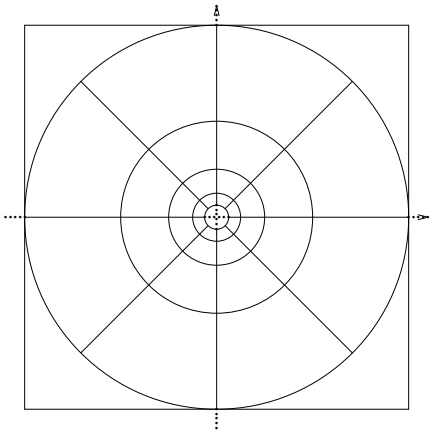
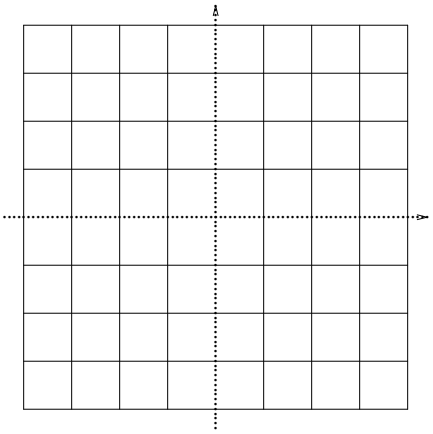


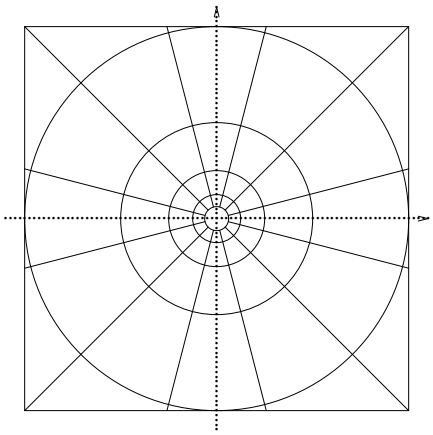
Figure 13: PIC metric for JPEG coder, perceptual PSNR = 47.9 DB



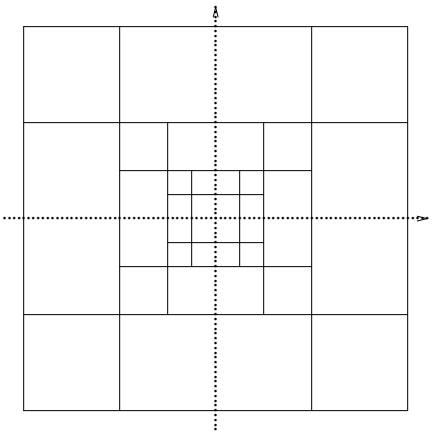
(a) Cortex transform (Watson)



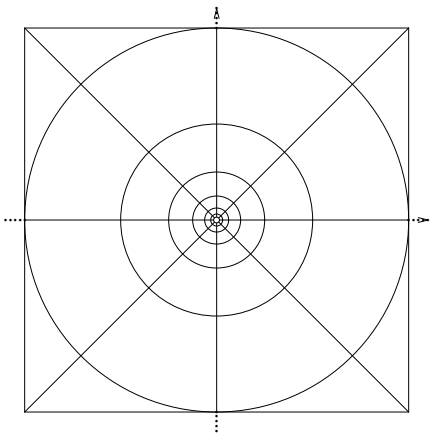
(d) Subband transform



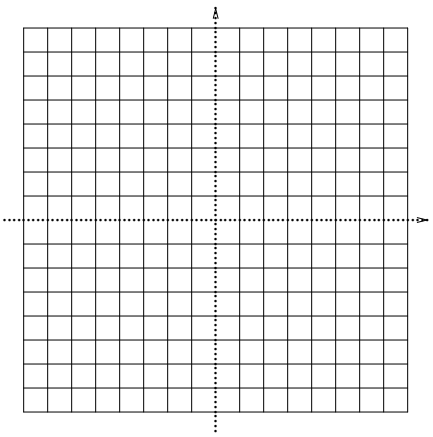
(b) Cortex transform (Daly)



(e) Wavelet transform



(c) Lubin's transform



(f) DCT transform

Figure 14: The decomposition of the frequency plane corresponding to various transforms. The range of each axis is from -$u_s/2$ to $u_s/2$ cycles per degree, where u_s is the sampling frequency.