Chapter in *Essential Guide of Image Processing*, Elsevier, 2009.

Image Quality Assessment

Kalpana Seshadrinathan, Thrasyvoulos N. Pappas,
Robert J. Safranek,
Junqing Chen, Zhou Wang,
Hamid R. Sheikh and Alan C. Bovik

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. The end user of these images, in an overwhelmingly large number of applications, are human observers. In this chapter, we examine objective criteria for the evaluation of image quality as perceived by an average human observer. 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. This paradigm of image QA (QA) is also known as full reference image QA. The development of objective metrics for evaluating image quality without a reference image is quite different and is outside the scope of this chapter.

Image QA plays a fundamental role in the design and evaluation of imaging and image processing systems. As an example, QA algorithms can be used to systematically evaluate the performance of different image compression algorithms that attempt to minimize the number of bits required to store an image, while maintaining sufficiently high image quality. Similarly, QA algorithms can be used to evaluate image acquisition and display systems. Communication networks have developed tremendously over the past decade and images and video are frequently transported over optic fiber, packet switched networks like the Internet, wireless systems etc. Bandwidth efficiency of applications such as video conferencing and Video on Demand (VoD) can be improved using QA systems to evaluate the effects of channel errors on the transported images and video. Further, QA algorithms can be used in "'perceptually optimal"' design of various components of an image communication system. Finally, QA and the psychophysics of human vision are closely related disciplines. Research on image and video QA may lend deep insights into the functioning of the human visual system (HVS), which would be of great scientific value.

Subjective evaluations 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 image quality [53, 54]. There has also been standards activity on subjective evaluation of image quality [55]. The study of the topic of subjective evaluation of image quality is beyond the scope of this chapter.

The goal of an objective 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. The output of the QA algorithm is often 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.

Perhaps, the earliest image quality metrics are the Mean squared error (MSE) and Peak Signal to Noise Ratio (PSNR) between the reference and distorted images. These metrics are still widely used for performance evaluation, despite their well-known limitations, due to their simplicity. Let $f(\mathbf{n})$ and $g(\mathbf{n})$ represent the value (intensity) 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 MSE between $f(\mathbf{n})$ and $g(\mathbf{n})$ is defined as:

$$\text{MSE}[f(\mathbf{n}), g(\mathbf{n})] = \frac{1}{N} \sum_{\mathbf{n}} [f(\mathbf{n}) - g(\mathbf{n})]^2 \tag{1}$$

where N is the total number of pixel locations in $f(\mathbf{n})$ or $g(\mathbf{n})$. The PSNR between these image patches is defined as:

$$\text{PSNR}[f(\mathbf{n}), g(\mathbf{n})] = 10 \log_{10} \frac{E^2}{\text{MSE}[f(\mathbf{n}), g(\mathbf{n})]^2} \tag{2}$$

where E is the maximum value that a pixel can take. For example, for 8 bit grayscale images, $E = 255$.

In Figure 5, we show two distorted images generated from the same original image. The first distorted image, Figure 5(b), was obtained by adding a constant number to all signal samples. The second distorted image, Figure 5(c), was generated using the same method except that the signs of the constant are randomly chosen to be positive or negative. It can be easily shown that the MSE/PSNR between the original image and both of the distorted images are exactly the same. However, the visual quality of the two distorted images is drastically different. Another example is shown in Figure 6, where Figure 6(b) was generated by adding independent white Gaussian noise to the original texture image in Figure 6(a). In Figure 6(c), the signal sample values remained the same as in Figure 6(a), but the spatial ordering of the samples has been changed (through a sorting procedure). Figure 6(d) was obtained from Figure 6(b), by following the same reordering procedure used to create Figure 6(c). Again, the MSE/PSNR between Figures 6(a) and 6(b) and Figures 6(c) and 6(d) are exactly the same. However, Figure 6(d) appears to be significantly noisier than Figure 6(b).

The above examples clearly illustrate the failure of PSNR as an adequate measure of visual quality. In this chapter, we will discuss three classes of image QA algorithms that correlate with visual perception significantly better - human vision based metrics, Structural SIMilarity (SSIM) metrics and information theoretic metrics. Each of these techniques approaches the image QA problem from a different perspective and using different first principles. As we proceed along this chapter, in addition to discussing these QA techniques, we will also attempt to shed light on the similarities, dissimilarities and interplay between these seemingly diverse techniques.

2 Human Vision Modeling Based Metrics

Human vision modeling based metrics utilize mathematical models of certain stages of processing that occur in the visual systems of humans to construct a quality metric. Most HVS based methods take an engineering approach to solving the quality assessment problem by measuring the threshold of visibility of signals and noise in the signals. These thresholds are then utilized to normalize the error between the reference and distorted images to obtain a perceptually meaningful error metric. To measure visibility thresholds, different aspects of visual processing need to be taken into consideration such as response to average brightness, contrast, spatial frequencies and orientations etc. Other HVS based methods attempt to directly model the different stages of processing that occurs in the HVS that results in the observed visibility thresholds. In Section 2.1, we will discuss the individual building blocks that comprise a HVS based QA system. The function of these blocks is to model concepts from the psychophysics of human perception that apply to image quality metrics. In Section 2.2, we will discuss the details of several well known HVS based QA systems. Each of these QA systems is comprised of some or all of the building blocks discussed in Section 2.1, but uses different mathematical models for each block.

2.1 Building Blocks

2.1.1 Pre-processing

Most QA algorithms include a pre-processing stage that typically comprises of calibration and registration. The array of numbers that represents an image are often mapped to units of visual frequencies or cycles per degree of visual angle and the calibration stage receives input parameters such as viewing distance and physical pixel spacings (screen resolution) to perform this mapping. Other calibration parameters may include fixation depth and eccentricity of the images in the observer's visual field [34, 35]. Display calibration or an accurate model of the display device is an essential part of any image quality metric [52], as the human visual system can only see what the display can reproduce. Many quality metrics require that the input image values be converted to physical luminances[1] before they enter the HVS model. In some cases, when the perceptual model is obtained empirically, the effects of the display are incorporated in the model [37]. The obvious disadvantage of this approach is that when the display changes, a new set of model parameters must be obtained [40]. The study of display models is beyond the scope of this chapter.

Registration, i.e., establishing point-by-point correspondence between two images, is also necessary in most image QA systems. Often times, the performance of a QA model can be extremely sensitive to registration errors since many QA systems operate pixel by pixel (e.g., PSNR) or on local neighborhoods of pixels. Errors in registration would result in a shift in the pixel or coefficient values being compared and degrade the performance of the system.

2.1.2 Frequency Analysis

The frequency analysis stage decomposes the reference and test images into different channels (usually called subbands) with different spatial frequencies and orientations using a set of linear filters. In many QA models, this stage is intended to mimic similar processing

[1] In video practice, the term luminance is sometimes, incorrectly, used to denote a nonlinear transformation of luminance [72, p. 24].

that occurs in the HVS: neurons in the visual cortex respond selectively to stimuli with particular spatial frequencies and orientations. Other QA models that target specific image coders utilize the same decomposition as the compression system and model the thresholds of visibility for each of the channels. Some examples of such decompositions are shown in Figure 4. The range of each axis is from -$u_s/2$ to $u_s/2$ cycles per degree, where u_s is the sampling frequency. Figure 4(a), (b) and (c) show transforms that are polar separable and belong to the former category of decompositions (mimicking processing in the visual cortex). Figure 4(d), (e) and (f) are used in QA models in the latter category and depict transforms that are often used in compression systems.

In the remainder of this chapter, 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.

2.1.3 Contrast Sensitivity

The human visual system's contrast sensitivity function (CSF, 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 [56, Ch. 12].

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 CSF and is illustrated in Figure 2. Note that these experiments used sine-wave gratings at a single orientation. To fully characterize the CSF, 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 QA, it is close enough to isotropic that this 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 often specify a minimum viewing distance and evaluate the 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.

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. 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 $\mathbf{n}$.

2.1.4 Luminance Masking

It is well known that the perception of lightness is a nonlinear function of luminance. Some authors call this "light adaptation." Others prefer the term "luminance masking", which groups it together with the other types of masking we will see below [38]. It is called masking because the luminance of the original image signal masks the variations in the distorted signal.

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

$$\frac{\delta I}{I} = k \tag{3}$$

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

2.1.5 Contrast Masking

We have dealt with stimuli that are either constant or contain a single frequency in describing the luminance masking and contrast sensitivity properties of the visual system respectively. In general, this is not characteristic of natural scenes. They have a wide range of frequency content over many different scales. Consider 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 we mentioned earlier, the visual cortex in 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 [61–63] or even neighboring subbands.

2.1.6 Error Pooling

The final step of an image quality metric is to combine the errors (at the output of the models for various psychophysical phenomena) 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.

An example of error pooling is the following Minkowski metric:

$$E(\mathbf{n}) = \frac{1}{M}\left\{\sum_{\mathbf{k}}\left|\frac{b(\mathbf{k},\mathbf{n}) - \hat{b}(\mathbf{k},\mathbf{n})}{t(\mathbf{k},\mathbf{n})}\right|^{Q}\right\}^{1/Q} \tag{4}$$

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 for each spatial location. The value of Q varies from 2 (energy summation) to infinity (maximum error).

2.2 HVS Based Models

In this section, we will discuss some well known HVS modeling based QA systems. We will first discuss four general purpose quality assessment models: the Visible Differences Predictor (VDP), the Sarnoff JND vision model, Teo and Heeger model and Visual Signal to Noise Ratio (VSNR).

We will then discuss quality models that are designed specifically for different compression systems: Perceptual Image Coder (PIC) and Watson's DCT and Wavelet based metrics. 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. The block diagram of a generic perceptually based coder is shown in Figure 1. 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 coder-specific image quality metrics are quite effective in predicting the performance of the coder they are designed for, they may not be as effective in predicting performance across different coders [33, 80].

2.2.1 Visible Differences Predictor

The Visible Differences Predictors (VDP) is a model developed by Daly for the evaluation of high quality imaging systems [34]. It is one of the most general and elaborate image quality metrics in the literature. It accounts for variations in sensitivity due to light level, spatial frequency (CSF), and signal content (contrast masking).

To model luminance masking or amplitude non-linearities in the HVS, Daly includes a simple point-by-point amplitude nonlinearity where the adaptation level for each image pixel is solely determined from that pixel (as opposed to using the average luminance in a neighborhood of the pixel). To account for the contrast sensitivity of the HVS, the VDP filters the image by the CSF before the frequency decomposition. Once this normalization is accomplished to account for the varying sensitivities of the HVS to different spatial

frequencies, the thresholds derived in the contrast masking stage become the same for all frequencies.

A variation of the Cortex transform shown in Figure 4(b) is used in the VDP for the frequency decomposition. Daly proposes two alternatives to convert the output of the linear filter bank to units of contrast: 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. The conversion to contrast is performed since to a first approximation, the HVS produces a neural image of local contrast [32]. The masking stage in the VDP utilizes a "'threshold elevation"' approach, where a masking function is computed that measures the contrast threshold of a signal as a function of the background (masker) contrast. This function is computed for the case when the masker and signal are single, isolated frequencies. To obtain a masking model for natural images, the VDP considers the results of experiments that have measured the masking thresholds for both single frequencies as well as additive noise. The VDP also allows for *mutual* masking which uses both the original and the distorted image to determine the degree of masking. The masking function used in the VDP is illustrated in Figure 3. Although the threshold elevation paradigm works quite well in determining the discriminability between the reference and distorted images, it fails to generalize to the case of supra-threshold distortions.

In the error pooling stage, a psychometric function is used to compute the probability of discrimination at each pixel of the reference and test images to obtain a spatial map. Further details of this algorithm can be found in [34], along with an interesting discussion of different approaches used in the literature to model various stages of processing in the HVS, their merits and drawbacks.

2.2.2 Sarnoff JND Vision Model

The Sarnoff JND vision model received a technical Emmy award in 2000 and is one of the best known QA systems based on human vision models. This model was developed by Lubin and co-workers and details of this algorithm can be found in [35].

Pre-processing steps in this model include calibration for distance of the observer from the images. In addition, this model also accounts for fixation depth and eccentricity of the observer's visual field. The human eye does not sample an image uniformly since the density of retinal cells drops off with eccentricity, resulting in a decreased spatial resolution as we move away from the point of fixation of the observer. To account for this effect, the Lubin model re-samples the image to generate a modeled retinal image. The Laplacian pyramid of Burt and Adelson [74] is used to decompose the image into seven radial frequency bands. At this stage, the pyramid responses are converted to units of 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. Each pyramid level is then convolved with eight spatially oriented filters of Freeman and Adelson [75], that constitute Hilbert transform pairs for four different orientations. The frequency decomposition so obtained in illustrated in Figure 4(c). The two Hilbert transform pair outputs are squared and summed to obtain a local energy measure at each pixel location, pyramid level and orientation. To account for the contrast sensitivity of human vision, these local energy measures are normalized by the base sensitivities for that position and pyramid level, where the base sensitivities are obtained from the CSF.

The Sarnoff model does not use the threshold elevation approach used by the VDP

to model masking, instead adopting to use a transducer or a contrast gain control model. Gain control models a mechanism that allows a neuron in the HVS to adjust its response to the ambient contrast of the stimulus. Such a model generalizes better to the case of supra-threshold distortions since it models an underlying mechanism in the visual system, as opposed to measuring visibility thresholds. The transducer model used in [35] takes the form of a sigmoid nonlinearity. 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.

Finally, a distance measure is calculated using a Minkowski error between the responses of the test and distorted images at the output of the vision model. A psychometric function is used to convert the distance measure to a probability value and the Sarnoff JND vision model outputs a spatial map that represents the probability that an observer will be able to discriminate between the two input images (reference and distorted) based on the information in that spatial location.

2.2.3 Teo and Heeger Model

The Teo and Heeger metric uses the steerable pyramid transform [76] which decomposes the image into several spatial frequency and orientation bands [36]. A more detailed discussion of this model, with a different transform, can be found in [77]. However, unlike the other two models we saw above, it does not attempt to separate the contrast sensitivity and contrast 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 model has the following form:

$$R(\mathbf{k}, \mathbf{n}, i) = R(\rho, \theta, \mathbf{n}, i) = \kappa_i \, \frac{[b(\rho, \theta, \mathbf{n})]^2}{\sum_\phi [b(\rho, \phi, \mathbf{n})]^2 + {\sigma_i}^2} \tag{5}$$

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. This model is also a contrast gain control model (similar the the Sarnoff JND vision model) that uses a divisive normalization model to explain masking effects. There is increasing evidence for divisive normalization mechanisms in the HVS and this model can account for various aspects of contrast masking in human vision [15, 28–31, 77]. Finally, the quality of the image is computed at each pixel as the Minkowski error between the contrast masked responses to the two input images.

2.2.4 Safranek-Johnston Perceptual Image Coder (PIC)

The Safranek-Johnston PIC image coder was one of the first image coders to incorporate an elaborate perceptual model [37]. It is calibrated for a given CRT display and viewing conditions (six times image height). The PIC coder has the basic structure shown in Figure 1. It uses a separable generalized quadrature mirror filter (GQMF) bank for subband analysis/synthesis shown in Figure 4(d). 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 $t_b(\mathbf{k})$ 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 denoted as $\tau_l(\mathbf{k}, \mathbf{n})$. 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. Safranek and Johnston [37] define as texture any deviation from a flat field within a subband and use the following texture masking adjustment:

$$\tau_t(\mathbf{k}, \mathbf{n}) = \max \left\{ 1, \left[\sum_{\mathbf{k}} w_{MTF}(\mathbf{k}) e_t(\mathbf{k}, \mathbf{n}) \right]^{w_t} \right\} \tag{6}$$

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:

$$e_t(\mathbf{k}, \mathbf{n}) = \begin{cases} \text{local variance}_{\mathbf{m} \in \mathrm{N}(\mathbf{n})}(b(\mathbf{0}, \mathbf{m})), & \text{if } \mathbf{k} = \mathbf{0} \\ b(\mathbf{k}, \mathbf{n})^2, & \text{otherwise} \end{cases} \tag{7}$$

where $\mathrm{N}(\mathbf{n})$ is the neighborhood of the point $\mathbf{n}$ over which the variance is calculated. In the Safranek-Johnston model, the overall sensitivity threshold is the product of three terms

$$t(\mathbf{k}, \mathbf{n}) = \tau_t(\mathbf{k}, \mathbf{n})\ \tau_l(\mathbf{k}, \mathbf{n})\ t_b(\mathbf{k}) \tag{8}$$

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.

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 (4), 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. 4(a)–4(g) demonstrates the performance of the PIC metric. Figure 4(a) shows an original 512×512 image. The gray-scale resolution is 8 bits/pixel. Figure 4(b) shows the image coded with the SPIHT coder [81] at 0.52 bits/pixel; the PSNR is 33.3 DB. Figure 4(b) shows the same image coded with the PIC coder [37] 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. 4(e) and 4(f) 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. Figure 4(d) shows the image coded with the standard JPEG algorithm at 0.52 bits/pixel and Figure 4(g) 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 weighting, which explains why the SPIHT image is superior according to PSNR and inferior according to perceptual PSNR. The above examples illustrate the power of image quality metrics.

2.2.5 Watson's DCTune

Many current compression standards are based on a discrete cosine transform (DCT) decomposition. Watson [3, 38] presented a model known as DCTune 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 [39, 41, 78]. The JPEG compression standard is discussed in Chapter 17.

Because of the popularity of DCT-based coders and computational efficiency of the DCT, we will give a more detailed overview of DCTune and how it can be used to obtain a metric of image quality.

The original reference and degraded images are partitioned into 8×8 pixel blocks and transformed to the frequency domain using the forward DCT. The DCT decomposition is similar to the subband decomposition and is shown in Figure 4(f). Perceptual thresholds are computed from the DCT coefficients of each block of data of the original 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, a threshold $t(\mathbf{k}, \mathbf{n})$ is computed using models for contrast sensitivity, luminance masking, and contrast masking.

The baseline contrast sensitivity thresholds $t_b(\mathbf{k})$ are determined by the Peterson, Ahumada, Watson method [82]. 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 then 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.

The Watson model of contrast masking assumes that the visibility reduction is confined to each coefficient in each block. The overall sensitivity threshold is determined as a function of a contrast masking adjustment and the luminance-masked threshold $t_l(\mathbf{k}, \mathbf{n})$:

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

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.

A distortion visibility threshold $d(\mathbf{k}, \mathbf{n})$ is computed at each location as the error at each location (the difference between the DCT coefficients in the original and distorted images) weighted by the sensitivity threshold.

$$d(\mathbf{k}, \mathbf{n}) = \frac{b(\mathbf{k}, \mathbf{n}) - \hat{b}(\mathbf{k}, \mathbf{n})}{t(\mathbf{k}, \mathbf{n})} \tag{12}$$

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.

To combine the distortion visibilities into a single value denoting the quality of the image, error pooling is first done spatially. Then the pools of spatial errors are pooled across frequency. Both pooling processes utilize the same probability summation framework.

$$p(\mathbf{k}) = \left\{ \sum_{\mathbf{n}} |d(\mathbf{k}, \mathbf{n})|^{Q_s} \right\}^{\frac{1}{Q_s}} \tag{13}$$

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, that are then pooled across frequency using a similar procedure.

$$P = \left\{ \sum_{\mathbf{k}} p(\mathbf{k})^{Q_f} \right\}^{\frac{1}{Q_f}} \tag{14}$$

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.

DCTune 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 [33, 80], it was found that the metric predictions (they used $Q_f = Q_s = 2$) are not always consistent with subjective evaluations when comparing different coders. It was found that this metric is strongly biased towards the JPEG algorithm. This is not surprising since both the metric and JPEG are based on the DCT.

2.2.6 Visual Signal to Noise Ratio

A general purpose quality metric known as the Visual Signal to Noise Ratio (VSNR) was developed by Chandler and Hemami [27]. VSNR differs from other HVS based techniques that we discuss in this section in three main ways. Firstly, the computational models used in VSNR are derived based on psychophysical experiments conducted to quantify the visual detectability of distortions in *natural images*, as opposed to sine wave gratings or Gabor patches used in most other models. Second, VSNR attempts to quantify the perceived contrast of *supra-threshold* distortions and the model is not restricted to the regime of threshold of visibility (such as the Daly model). Third, VSNR attempts to capture a *mid-level* property of the HVS known as global precedence, while most other models discussed here only consider low level processes in the visual system.

In the pre-processing stage, VSNR accounts for viewing conditions (display resolution and viewing distance) and display characteristics. The original image, $f(\mathbf{n})$, and the pixel-wise errors between the original and distorted images, $f(\mathbf{n}) - g(\mathbf{n})$, are decomposed using an M-level discrete wavelet transform (DWT) using the 9/7 bi-orthogonal filters. VSNR defines a model to compute the average contrast signal to noise ratios (CSNR) at the threshold of detection for wavelet distortions in natural images for each sub-band of the wavelet decomposition. To determine whether the distortions are visible within each octave band of frequencies, the actual contrast of the distortions are compared with the corresponding contrast detection threshold. If the contrast of the distortions is lower than the corresponding detection threshold for all frequencies, the distorted image is declared to be of perfect quality.

In Section 2.1.3, we mentioned the CSF of human vision and several models discussed here attempt to model this aspect of human perception. Although the CSF is critical in determining whether the distortions are visible in the test image, the utility of the CSF in measuring the visibility of supra-threshold distortions has been debated. The perceived contrast of supra-threshold targets has been shown to depend much less on spatial frequency than what is predicted by the CSF, a property also known as contrast constancy. The VSNR assumes contrast constancy and if the distortion is supra-threshold, the RMS contrast of the error signal is used as a measure of the perceived contrast of the distortion, denoted by d_{pc}.

Finally, the VSNR models of the global precedence property of human vision - the visual system has a preference for integrating edges in a coarse to fine scale fashion. VSNR models the global precedence preserving CSNR for each octave band of spatial frequencies. This model satisfies the following property - for supra-threshold distortions, the CSNR corresponding to coarse spatial frequencies is greater than the CSNR corresponding to finer scales. Further, as the distortions become increasingly supra-threshold, coarser scales have increasingly greater CSNR than finer scales in order to preserve visual integration of edges in a coarse to fine scale fashion. For a given distortion contrast, the contrast of the distortions within each sub-band is compared with the corresponding global precedence preserving contrast specified by the model to compute a measure d_{gp} of the extent to which global precedence has been disrupted. The final quality metric is a linear combination of d_{pc} and d_{gp}.

3 Structural Approaches

In this section, we will discuss structural approaches to image QA. We will discuss the structural similarity philosophy in Section 3.1. We will show some illustrations of the performance of this metric in Section 3.2. Finally, we will discuss the relation between SSIM and HVS based metrics in Section 3.3.

3.1 The Structural Similarity Index

The most fundamental principle underlying structural approaches to image QA is that the HVS is highly adapted to extract structural information from the visual scene, and therefore a measurement of *structural* similarity (or distortion) should provide a good approximation to perceptual image quality. Depending on how structural information and structural distortion are defined, there may be different ways to develop image QA algorithms. The SSIM index is a specific implementation from the perspective of image formation. The luminance of the surface of an object being observed is the product of the illumination and the reflectance, but the structures of the objects in the scene are independent of the illumination. Consequently, we wish to separate the influence of illumination from the remaining information that represents object structures. Intuitively, the major impact of illumination change in the image is the variation of the average local luminance and contrast, and such variation should not have a strong effect on perceived image quality.

Consider two image patches $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ obtained from the reference and test images. Mathematically, $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ denote two vectors of dimension N, where $\tilde{\mathbf{f}}$ is composed of N elements of $f(\mathbf{n})$ spanned by a window B and similarly for $\tilde{\mathbf{g}}$. To index each element of $\tilde{\mathbf{f}}$, we use the notation $\tilde{\mathbf{f}} = [\tilde{f}_1, \tilde{f}_2, \ldots, \tilde{f}_N]^T$.

First, the luminance of each signal is estimated as the mean intensity:

$$\mu_{\tilde{\mathbf{f}}} = \frac{1}{N}\sum_{i=1}^{N} \tilde{f}_i \tag{15}$$

A luminance comparison function $l(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ is then defined as a function of $\mu_{\tilde{\mathbf{f}}}$ and $\mu_{\tilde{\mathbf{g}}}$:

$$l[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}] = \frac{2\mu_{\tilde{\mathbf{f}}}\mu_{\tilde{\mathbf{g}}} + C_1}{\mu_{\tilde{\mathbf{f}}}^2 + \mu_{\tilde{\mathbf{g}}}^2 + C_1} \tag{16}$$

where the constant C_1 is included to avoid instability when $\mu_{\tilde{\mathbf{f}}}^2 + \mu_{\tilde{\mathbf{g}}}^2$ is very close to zero. One good choice is $C_1 = (K_1 E)^2$, where E is the dynamic range of the pixel values (255 for 8-bit grayscale images), and $K_1 << 1$ is a small constant. Similar considerations also apply to contrast comparison and structure comparison terms described below.

The contrast of each image patch is defined as an unbiased estimate of the standard deviation of the patch:

$$\sigma_{\tilde{\mathbf{f}}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\tilde{f}_i - \mu_{\tilde{\mathbf{f}}}\right)^2 \tag{17}$$

The contrast comparison $c(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ takes a similar form as the luminance comparison function and is defined as a function of $\sigma_{\tilde{\mathbf{f}}}$ and $\sigma_{\tilde{\mathbf{g}}}$:

$$c[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}] = \frac{2\sigma_{\tilde{\mathbf{f}}}\sigma_{\tilde{\mathbf{g}}} + C_2}{\sigma_{\tilde{\mathbf{f}}}^2 + \sigma_{\tilde{\mathbf{g}}}^2 + C_2} \tag{18}$$

where C_2 is a non-negative constant. $C_2 = (K_2E)^2$, where K_2 satisfies $K_2 << 1$.

Third, the signal is normalized (divided) by its own standard deviation, so that the two signals being compared have unit standard deviation. The structure comparison $s(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ is conducted on these normalized signals. The SSIM framework uses a geometric interpretation and the structures of the two images are associated with the direction of the two unit vectors $\tilde{\mathbf{f}} - \mu_{\tilde{\mathbf{f}}}/\sigma_{\tilde{\mathbf{f}}}$ and $\tilde{\mathbf{g}} - \mu_{\tilde{\mathbf{g}}}/\sigma_{\tilde{\mathbf{g}}}$. The angle between the two vectors provides a simple and effective measure to quantify structural similarity. In particular, the correlation coefficient between $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ corresponds to the cosine of the angle between them and is used as the structure comparison function:

$$s[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}] = \frac{\sigma_{\tilde{\mathbf{f}}\tilde{\mathbf{g}}} + C_3}{\sigma_{\tilde{\mathbf{f}}}\sigma_{\tilde{\mathbf{g}}} + C_3} \tag{19}$$

where the sample covariance between $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ is estimated as:

$$\sigma_{\tilde{\mathbf{f}}\tilde{\mathbf{g}}} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\tilde{f}_i - \mu_{\tilde{\mathbf{f}}}\right) (\tilde{g}_i - \mu_{\tilde{\mathbf{g}}}) \tag{20}$$

Finally, the SSIM index between image patches $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ is defined as

$$\text{SSIM}[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}] = l[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}]^{\alpha} \,.\, c[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}]^{\beta} \,.\, s[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}]^{\gamma} \tag{21}$$

where α, β and γ are parameters used to adjust the relative importance of the three components.

The SSIM index and the three comparison functions - luminance, contrast and structure - satisfy the following desirable properties.

- *Symmetry*: $\text{SSIM}(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) = \text{SSIM}(\tilde{\mathbf{g}}, \tilde{\mathbf{f}})$. When quantifying the similarity between two signals, exchanging the order of the input signals should not affect the resulting measurement.

- *Boundedness*: $\text{SSIM}(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) \leq 1$. An upper bound can serve as an indication of how close the two signals are to being perfectly identical.

- *Unique maximum*: $\text{SSIM}(\tilde{\mathbf{f}}, \tilde{\mathbf{g}}) = 1$ if and only if $\tilde{\mathbf{f}} = \tilde{\mathbf{g}}$. The perfect score is achieved only when the signals being compared are identical. In other words, the similarity measure should quantify any variations that may exist between the input signals.

The structure term of the SSIM index is independent of the luminance and contrast of the local patches, which is physically sensible because the change of luminance and/or contrast has little impact on the structures of the objects in the scene. Although the SSIM index is defined by three terms, the structure term in the SSIM index is generally regarded as the most important, since variations in luminance and contrast of an image do not affect visual quality as much as structural distortions [25].

3.2 Image Quality Assessment Using SSIM

The SSIM index measure the structural similarity between two images. If one of the images is regarded as of perfect quality, then the SSIM index can be viewed as an indication of the quality of the other image signal being compared. When applying the SSIM index approach

to large-size images, it is useful to compute it locally rather than globally. The reason is manifold. First, statistical features of images are usually spatially non-stationary. Second, image distortions, which may or may not depend on the local image statistics, may also vary across space. Third, due to the non-uniform retinal sampling feature of the HVS, at typical viewing distances, only a local area in the image can be perceived with high resolution by the human observer at one time instance. Finally, localized quality measurement can provide a spatially varying quality map of the image, which delivers more information about the quality degradation of the image. Such a quality map can be used in different ways. It can be employed to indicate the quality variations across the image. It can also be used to control image quality for space-variant image processing systems, e.g., region-of-interest image coding and foveated image processing.

In early instantiations of the SSIM index approach [25], the local statistics $\mu_{\tilde{\mathbf{f}}}$, $\sigma_{\tilde{\mathbf{f}}}$ and $\sigma_{\tilde{\mathbf{f}}\tilde{\mathbf{g}}}$ defined in Eqs. (15), (17) and (20) were computed within a local 8×8 square window. The window moves pixel-by-pixel from the top-left corner to the bottom-right corner of the image. At each step, the local statistics and SSIM index are calculated within the local window. One problem with this method is that the resulting SSIM index map often exhibits undesirable "blocking" artifacts as exemplified by Fig. 7(c). Such "artifacts" are not desirable because it is created from the choice of the quality measurement method (local square window) and not from image distortions. In [26], a circular-symmetric Gaussian weighting function $\mathbf{w} = \{w_i, i = 1, 2, \dots N\}$ with unit sum ($\sum_{i=1}^{N} w_i = 1$) is adopted. The estimates of $\mu_{\tilde{\mathbf{f}}}$, $\sigma_{\tilde{\mathbf{f}}}$ and $\sigma_{\tilde{\mathbf{f}}\tilde{\mathbf{g}}}$ are then modified accordingly:

$$\mu_{\tilde{\mathbf{f}}} = \sum_{i=1}^{N} w_i \tilde{f}_i \tag{22}$$

$$\sigma_{\tilde{\mathbf{f}}}^2 = \sum_{i=1}^{N} w_i \left(\tilde{f}_i - \mu_{\tilde{\mathbf{f}}}\right)^2 \tag{23}$$

$$\sigma_{\tilde{\mathbf{f}}\tilde{\mathbf{g}}} = \sum_{i=1}^{N} w_i \left(\tilde{f}_i - \mu_{\tilde{\mathbf{f}}}\right)\left(\tilde{g}_i - \mu_{\tilde{\mathbf{g}}}\right) \tag{24}$$

With such a smoothed windowing approach, the quality maps exhibit a locally isotropic property as demonstrated in Fig. 7(d).

Fig. 8 shows the SSIM index maps of a set of sample images with different types of distortions. The absolute error map for each distorted image is also included for comparison. The SSIM index and absolute error maps have been adjusted, so that brighter always indicates better quality in terms of the given quality/distortion measure. It can be seen that the distorted images exhibit variable quality across space. For example, in image 8(b), the noise over the face region appears to be much more significant than that in the texture regions. However, the absolute error map 8(d) is completely independent of the underlying image structures. By contrast, the SSIM index map 8(c) gives perceptually consistent prediction. In image 8(f), the bit allocation scheme of low bit-rate JPEG2000 compression leads to smooth representations of detailed image structures. For example, the texture information of the roof of the building as well as the trees is lost. This is very well indicated by the SSIM index map 8(g), but cannot be predicted from the absolute error map 8(h). Some different types of distortions are caused by low bit-rate JPEG compression. In image 8(j), the major distortions we observe are the blocking effect in the sky and the artifacts around the outer boundaries of the building. Again, the absolute error map 8(l)

fails to provide useful prediction, and the SSIM index map 8(k) successfully predicts image quality variations across space. From these sample images, we see that an image quality measure as simple as the SSIM index can adapt to various kinds of image distortions and provide perceptually consistent quality predictions.

The final step of an image quality measurement system is to combine the quality map into one single quality score for the whole image. A convenient way is to use a weighted summation. If $f(\mathbf{n}$ and $g(\mathbf{n})$ are the two images being compared, and $\mathrm{SSIM}[f(\mathbf{n}), g(\mathbf{n})]$ be the local SSIM index evaluated at location $\mathbf{n}$. Then, the Mean SSIM (MSSIM) index between $f(\mathbf{n})$ and $g(\mathbf{n})$ is defined as:

$$\mathrm{MSSIM}[f(\mathbf{n}), g(\mathbf{n})] = \frac{\sum_{\mathbf{n}} W(\mathbf{n})\, \mathrm{SSIM}[f(\mathbf{n}), g(\mathbf{n})]}{\sum_{\mathbf{n}} W(\mathbf{n})} \tag{25}$$

where $W(\mathbf{n})$ is the weight given to the pixel location $\mathbf{n}$. If all the samples in the quality map are equally weighted, this results in the measure employed in [26]. There are two cases in which non-uniform weighting is desirable. First, depending on the application, some prior knowledge about the importance of different regions in the image is available, and such an importance map can be converted into a weighting function. For example, object-based region-of-interest image processing systems often segment the objects in the scene and give different objects different importance. In a foveated image processing system (see Chapter 4.8), the weighting function can be determined according to the foveation feature of the HVS, i.e., the visual resolution decreases gradually as a function of the distance from the fixation point. Note that the weighting function here is determined only by the spatial location and is independent of the local image content. In the second case, the image content also plays a role. It has been observed that different image textures attract human fixations with varying degrees, and therefore different weights can be assigned. In [13], a variance-weighted weighting function was used. It was observed that this weighting function is useful to balance the extreme case where severe high-variance distortions concentrate at some small areas in the image.

3.3 Relation to HVS Based Models

In this section, we will relate the structure term in the SSIM index to MSE and human vision based metrics [10, 14].

The MSE between image patches $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ is

$$\mathrm{MSE}[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}] = \frac{1}{N} \sum_{i=1}^{N} (\tilde{f}_i - \tilde{g}_i)^2$$

We define normalized random variables

$$\tilde{\mathbf{f}} = \frac{\tilde{\mathbf{f}} - \mu_{\tilde{\mathbf{f}}}}{\sigma_{\tilde{\mathbf{f}}}} \tag{26}$$

$$\tilde{\mathbf{g}} = \frac{\tilde{\mathbf{g}} - \mu_{\tilde{\mathbf{g}}}}{\sigma_{\tilde{\mathbf{g}}}} \tag{27}$$

Observe that:

$$\mathrm{MSE}\left(\frac{\tilde{\mathbf{f}} - \mu_{\tilde{\mathbf{f}}}}{\sigma_{\tilde{\mathbf{f}}}}, \frac{\tilde{\mathbf{g}} - \mu_{\tilde{\mathbf{g}}}}{\sigma_{\tilde{\mathbf{g}}}} \right) = 2 \left(1 - \frac{\sigma_{\tilde{\mathbf{f}}\tilde{\mathbf{g}}}}{\sigma_{\tilde{\mathbf{f}}} \sigma_{\mathbf{g(i)}}} \right) \tag{28}$$

Thus, the structure term in the SSIM index essentially computes an MSE between image patches, after normalizing them for their mean and standard deviations. This is not surprising in view of the fact that the structure term of the SSIM index is defined to be independent of the mean and standard deviation of the image intensity values. This relationship between the structure term of SSIM and MSE illustrates that simple modifications to the MSE computation can help overcome some of its drawbacks such as failure to detect brightness and contrast changes (illustrated in Fig. 5(b)).

We will now discuss the relation of SSIM to HVS based metrics. We first look at the structure term of the SSIM index, which is arguably the most important term in the SSIM index [10, 14]. It is evident that the definition of the normalized variables in (26),(27) is very similar to divisive normalization models of contrast gain control in HVS based metrics. In fact, a SSIM contrast masking model can be defined by:

$$R[\tilde{f}_i] = \frac{\tilde{f}_i - \mu_{\tilde{\mathbf{f}}}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\tilde{f}_i - \mu_{\tilde{\mathbf{f}}}\right]^2}} \tag{29}$$

We discussed in Section 2.1.6 that most HVS based QA systems compute a Minkowski error between the outputs of the contrast gain control model (as well as models of other aspects of the HVS incorporated in QA) to the reference and test image patches as an index of quality, often with a Minkowski exponent of 2 [35, 36, 38]. Similarly, observe that the structure term of SSIM in (28) is a monotonic function of the square of the Minkowski error between the outputs of the SSIM contrast gain control model in (29) with exponent 2.

It is important to note that the SSIM indices perform the gain control normalization in the *image pixel* domain. In (29), the contrast gain control model divisively inhibits each pixel by pooling the responses of a local spatial neighborhood of pixels. However, contrast masking in the HVS is a phenomenon that occurs in a frequency-orientation decomposed domain. For example, the masking effect is maximum when the orientation of the masker and the target are parallel and decreases when their orientations are perpendicular [24]. The contrast masking models that we saw in Section 2.2 can capture these effects since they normalize each filter output by pooling the responses at the same location of *other filters* tuned to different orientations. However, the SSIM metric will not be able to account for such effects. Improved versions of the SSIM index that use a frequency decomposition have been proposed [16, 17] and our discussion of the relation between the SSIM index and contrast masking models helps us understand the reasons for the improved performance of these metrics. Interestingly, the square of the response of the SSIM contrast gain control model defined by (29) is equal to the response of the Teo and Heeger gain control model defined by (5) with $\kappa_i = N$ and $\sigma_i^2 = 0$ for all i. Thus, if the vectors $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ at each spatial location $\mathbf{n}$ are defined using the responses of a filter bank at that location (as is done in the Complex Wavelet SSIM model [17]), the SSIM contrast masking model is just the square root of the Teo and Heeger contrast masking model.

Also, (16) is connected with Weber's law which was discussed in Section 2.1.4 [26]. According to Weber's law, the HVS is sensitive to the relative rather than the absolute luminance change. Letting R represent the ratio of the luminance of the distorted signal relative to the reference signal, then we can write $\mu_{\tilde{\mathbf{g}}} = R\mu_{\tilde{\mathbf{f}}}$. Substituting this into (16) gives

$$l[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}] = \frac{2R}{1 + R^2 + \frac{C_1}{\mu_{\tilde{\mathbf{f}}}^2}} \tag{30}$$

If we assume C_1 is small enough (relative to $\mu_{\tilde{\mathbf{f}}}^2$) to be ignored, then $l[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}]$ is a function only of R. In this sense, it is qualitatively consistent with Weber's law.

This discussion shows that although the SSIM index was derived from very different first principles, at least part of the reasons for its success can be attributed to similarities it shares with models of the HVS.

4 Information Theoretic Approaches

In this section, we discuss information theoretic approaches to image quality assessment. We will discuss the information theoretic metrics in Section 4.1. We will show some illustrations of the performance of this metric in Section 4.2. Finally, we will discuss the relation between structural similarity and information theoretic metrics in Section 4.3.

4.1 Information Theoretic Metrics

In the information-theoretic approach to quality assessment, the quality assessment problem is viewed as an information-fidelity problem rather than a signal-fidelity problem. An image source communicates to a receiver through a channel that limits the amount of information that could flow through it, thereby introducing distortions. The output of the image source is the reference image, the output of the channel is the test image, and the visual quality of the test image is computed as the amount of information shared between the test and the reference signals, i.e., the mutual information between them. Thus, information fidelity methods exploit the relationship between statistical image information and visual quality.

Statistical models for signal sources and transmission channels are at the core of information theoretic analysis techniques. A fundamental component of information fidelity based QA methods is a model for image sources. Images and videos whose quality needs to be assessed are usually optical images of the three dimensional visual environment, or *natural scenes*. Natural scenes form a very tiny subspace in the space of all possible image signals and researchers have developed sophisticated models that capture key statistical features of natural images. A review of these models has been presented in Chapter 4.7.

In this chapter we present two full-reference quality assessment methods based on the information-fidelity paradigm. Both methods share a common mathematical framework. The first method, the information fidelity criterion (IFC) [23], uses a distortion channel model as depicted in Figure 9. The IFC quantifies the information shared between the test image and the distorted image. The other method that we shall present in this chapter is the visual information fidelity (VIF) measure [22], which uses an additional HVS channel model, and utilizes two aspects of image information for quantifying perceptual quality: the information shared between the test and the reference image, and the information content of the reference image itself. This is depicted pictorially in Figure 10.

Images and videos of the visual environment captured using high quality capture devices operating in the visual spectrum are broadly classified as natural scenes. This differentiates them from text, computer generated graphics scenes, cartoons and animations, paintings and drawings, random noise, or images and videos captured from non-visual stimuli such as Radar and Sonar, X-Rays, ultra-sounds etc. The model for natural images that is used in the information theoretic metrics is the Gaussian Scale Mixture (GSM) model in the wavelet domain.

A GSM is a random field (RF) that can be expressed as a product of two independent RFs [11]. That is, a GSM $\mathcal{C} = \{\overrightarrow{C}_{\mathbf{n}} : \mathbf{n} \in \mathscr{N}\}$, where $\mathscr{N}$ denotes the set of spatial indices for the RF, can be expressed as:

$$\mathcal{C} = \mathcal{S} \cdot \mathcal{U} = \{S_{\mathbf{n}} \cdot \overrightarrow{U}_{\mathbf{n}} : \mathbf{n} \in \mathscr{N}\} \tag{31}$$

where $\mathcal{S} = \{S_{\mathbf{n}} : \mathbf{n} \in \mathscr{N}\}$ is an RF of positive scalars also known as the mixing density and $\mathcal{U} = \{\overrightarrow{U}_{\mathbf{n}} : \mathbf{n} \in \mathscr{N}\}$ is a Gaussian vector RF with mean zero and covariance matrix

$\mathbf{C}_U$. $\overrightarrow{C}_{\mathbf{n}}$ and $\overrightarrow{U}_{\mathbf{n}}$ are M dimensional vectors, and we assume that for the RF $\mathcal{U}$, $\overrightarrow{U}_{\mathbf{n}}$ is independent of $\overrightarrow{U}_{\mathbf{m}}$, $\forall \mathbf{n} \neq \mathbf{m}$. We model each subband of a scale-space-orientation wavelet decomposition (such as the steerable pyramid [12]) of an image as a GSM. We partition the subband coefficients into non-overlapping blocks of M coefficients each, and model block $\mathbf{n}$ as the vector $\overrightarrow{C}_{\mathbf{n}}$. Thus image blocks are assumed to be uncorrelated with each other, and any linear correlations between wavelet coefficients is modeled only through the covariance matrix $\mathbf{C}_U$.

One could easily make the following observations regarding the above model: $\mathcal{C}$ is normally distributed given $\mathcal{S}$ (with mean zero, and covariance of $\overrightarrow{C}_{\mathbf{n}}$ being $S_{\mathbf{n}}^2\mathbf{C}_U$), that given $S_{\mathbf{n}}$, $C_{\mathbf{n}}$ are independent of $S_{\mathbf{m}}$ for all $\mathbf{n} \neq \mathbf{m}$, and that given $\mathcal{S}$, $\overrightarrow{C}_{\mathbf{n}}$ are conditionally independent of $\overrightarrow{C}_{\mathbf{m}}$, $\forall \mathbf{n} \neq \mathbf{m}$ [11]. These properties of the GSM model make analytical treatment of information fidelity possible.

The information theoretic metrics assume that the distorted image is obtained by applying a distortion operator on the reference image. The distortion model used in the information theoretic metrics is a signal attenuation and additive noise model in the wavelet domain:

$$\mathcal{D} = \mathcal{GC} + \mathcal{V} = \{g_{\mathbf{n}}\overrightarrow{C}_{\mathbf{n}} + \overrightarrow{V}_{\mathbf{n}} : \mathbf{n} \in \mathcal{N}\} \tag{32}$$

where $\mathcal{C}$ denotes the RF from a subband in the reference signal, $\mathcal{D} = \{\overrightarrow{D}_{\mathbf{n}} : \mathbf{n} \in \mathcal{N}\}$ denotes the RF from the corresponding subband from the test (distorted) signal, $\mathcal{G} = \{g_{\mathbf{n}} : \mathbf{n} \in \mathcal{N}\}$ is a deterministic scalar gain field and $\mathcal{V} = \{\overrightarrow{V}_{\mathbf{n}} : \mathbf{n} \in \mathcal{N}\}$ is a stationary additive zero-mean Gaussian noise RF with covariance matrix $\mathbf{C}_V = \sigma_V^2\mathbf{I}$. The RF $\mathcal{V}$ is white, and is independent of $\mathcal{S}$ and $\mathcal{U}$. We constrain the field $\mathcal{G}$ to be slowly-varying.

This model captures important, and complementary, distortion types: blur, additive noise, and global or local contrast changes. The attenuation factors $g_{\mathbf{n}}$ would capture the loss of signal energy in a subband due to blur distortion, and the process $\mathcal{V}$ would capture the additive noise components separately.

We will now discuss the information fidelity criterion (IFC) and the Visual Information Fidelity (VIF) criteria in the following sections.

4.1.1 The Information Fidelity Criterion

The IFC quantifies the information shared between a test and the reference image. The reference image is assumed to pass through a channel yielding the test image, and the mutual information between the reference and the test images is used for predicting visual quality.

Let $\overrightarrow{C}^N = \{\overrightarrow{C}_1, \overrightarrow{C}_2, \ldots, \overrightarrow{C}_N\}$ denote N elements from $\mathcal{C}$. Let S^N and $\overrightarrow{D}^N$ be correspondingly defined. The IFC uses the mutual information between the reference and test images conditioned on the mixing density in the GSM model, i.e. $I(\overrightarrow{C}^N; \overrightarrow{E}^N | \overrightarrow{S}^N = s^N)$, as an indicator of visual quality. With the stated assumptions on $\mathcal{C}$ and the distortion model, it can easily be shown that [23]:

$$I(\overrightarrow{C}^N; \overrightarrow{D}^N | s^N) = \frac{1}{2}\sum_{\mathbf{n}=1}^{N}\sum_{k=1}^{M}\log_2\left(1 + \frac{g_{\mathbf{n}}^2 s_{\mathbf{n}}^2 \lambda_k}{\sigma_V^2}\right) \tag{33}$$

where λ_k are the eigenvalues of $\mathbf{C}_U$.

Note that in the above treatment it is assumed that the model parameters s^N, $\mathcal{G}$ and σ_V^2 are known. Details of practical estimation of these parameters are given in Section 4.1.3. In

the development of the IFC, we have so far only dealt with one subband. One could easily incorporate multiple subbands by assuming that each subband is completely independent of others in terms of the RFs as well as the distortion model parameters. Thus the IFC is given by:

$$\text{IFC} \;=\; \sum_{j \in \text{subbands}} I(\overrightarrow{C}^{N,j}; \overrightarrow{D}^{N,j} | s^{N,j}) \tag{34}$$

where the summation is carried over the subbands of interest, and $\overrightarrow{C}^{N,j}$ represent N_j elements of the RF $\mathcal{C}_j$ that describes the coefficients from subband j, and so on.

4.1.2 The Visual Information Fidelity Criterion

In addition to the distortion channel, VIF assumes that both the reference and distorted images pass through the HVS, which acts as a 'distortion channel' that imposes limits on how much information could flow through it. The purpose of HVS model in the information fidelity setup is to quantify the uncertainty that the HVS adds to the signal that flows through it. As a matter of analytical and computational simplicity, we lump all sources of HVS uncertainty into one additive noise component that serves as a *distortion baseline* in comparison to which the distortion added by the distortion channel could be evaluated. We call this lumped HVS distortion *visual noise*, and model it as a stationary, zero mean, additive white Gaussian noise model in the wavelet domain. Thus, we model the HVS noise in the wavelet domain as stationary RFs $\mathcal{H} = \{\overrightarrow{H}_{\mathbf{n}} : \mathbf{n} \in \mathscr{N}\}$ and $\mathcal{H}' = \{\overrightarrow{H'}_{\mathbf{n}} : \mathbf{n} \in \mathscr{N}\}$, where $\overrightarrow{H}_i$ and $\overrightarrow{H'}_i$ are zero-mean uncorrelated multivariate Gaussian with the same dimensionality as $\overrightarrow{C}_{\mathbf{n}}$:

$$\mathcal{E} \;=\; \mathcal{C} + \mathcal{H} \;\text{ (reference image)} \tag{35}$$

$$\mathcal{F} \;=\; \mathcal{D} + \mathcal{H}' \;\text{ (test image)} \tag{36}$$

where $\mathcal{E}$ and $\mathcal{F}$ denote the visual signal at the output of the HVS model from the reference and the test images in one subband respectively (Figure 10). The RFs $\mathcal{H}$ and $\mathcal{H}'$ are assumed to be independent of $\mathcal{U}$, $\mathcal{S}$, and $\mathcal{V}$. We model the covariance of $\mathcal{H}$ and $\mathcal{H}'$ as:

$$\mathbf{C}_H = \mathbf{C}_{H'} = \sigma_H^2 \mathbf{I} \tag{37}$$

where σ_H^2 is an HVS model parameter (variance of the visual noise).

It can be shown [22] that :

$$I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N) \;=\; \frac{1}{2} \sum_{\mathbf{n}=1}^{N} \sum_{k=1}^{M} \log_2 \left(1 + \frac{s_{\mathbf{n}}^2 \lambda_k}{\sigma_H^2}\right) \tag{38}$$

$$I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N) \;=\; \frac{1}{2} \sum_{\mathbf{n}=1}^{N} \sum_{k=1}^{M} \log_2 \left(1 + \frac{g_{\mathbf{n}}^2 s_{\mathbf{n}}^2 \lambda_k}{\sigma_V^2 + \sigma_H^2}\right) \tag{39}$$

where λ_k are the eigenvalues of $\mathbf{C}_U$.

$I(\overrightarrow{C}^N; \overrightarrow{E}^N | s^N)$ and $I(\overrightarrow{C}^N; \overrightarrow{F}^N | s^N)$ represent the information that could ideally be extracted by the brain from a particular subband of the reference and the test images respectively. A simple *ratio* of the two information measures relates quite well with visual

quality [22]. It is easy to motivate the suitability of this relationship between image information and visual quality. When a human observer sees a distorted image, he has an idea of the amount information that he expects to receive in the image (modeled through the known $\mathcal{S}$ field), and it is natural to expect the fraction of the expected information that is actually received from the distorted image to relate well with visual quality.

As with the IFC, the VIF could easily be extended to incorporate multiple subbands by assuming that each subband is completely independent of others in terms of the RFs as well as the distortion model parameters. Thus, the VIF is given by:

$$\text{VIF} \quad = \quad \frac{\sum_{j \in \text{subbands}} I(\overrightarrow{C}^{N,j}; \overrightarrow{F}^{N,j} | s^{N,j})}{\sum_{j \in \text{subbands}} I(\overrightarrow{C}^{N,j}; \overrightarrow{E}^{N,j} | s^{N,j})} \tag{40}$$

where we sum over the subbands of interest, and $\overrightarrow{C}^{N,j}$ represent N elements of the RF $\mathcal{C}_j$ that describes the coefficients from subband j, and so on.

The VIF given in (40) is computed for a collection of wavelet coefficients that could either represent an entire subband of an image, or a spatially localized set of subband coefficients. In the former case, the VIF is a single number that quantifies the information fidelity for the entire image, whereas in the latter case, a sliding-window approach could be used to compute a *quality map* that could visually illustrate how the visual quality of the test image varies over space.

4.1.3 Implementation Details

The source model parameters that need to be estimated from the data consist of the field $\mathcal{S}$. For the vector GSM model, the maximum-likelihood estimate of $s_{\mathbf{n}}^2$ can be found as follows [18]:

$$\widehat{s}_{\mathbf{n}}^2 = \frac{\overrightarrow{C}_{\mathbf{n}}^T \mathbf{C}_U^{-1} \overrightarrow{C}_{\mathbf{n}}}{M} \tag{41}$$

Estimation of the covariance matrix $\mathbf{C}_U$ is also straightforward from the reference image wavelet coefficients [18]:

$$\widehat{\mathbf{C}}_U = \frac{1}{N} \sum_{\mathbf{n}=1}^{N} \overrightarrow{C}_{\mathbf{n}} \overrightarrow{C}_{\mathbf{n}}^T \tag{42}$$

In (41) and (42), $\frac{1}{N} \sum_{\mathbf{n}=1}^{N} s_{\mathbf{n}}^2$ is assumed to be unity without loss of generality [18].

The parameters of the distortion channel are estimated locally. A spatially localized block-window centered at coefficient $\mathbf{n}$ could be used to estimate $g_{\mathbf{n}}$ and σ_V^2 at $\mathbf{n}$. The value of the field $\mathcal{G}$ over the block centered at coefficient $\mathbf{n}$, which we denote as $g_{\mathbf{n}}$, and the variance of the RF $\mathcal{V}$, which we denote as $\sigma_{V,\mathbf{n}}^2$, are fairly easy to estimate (by linear regression) since both the input (the reference signal) as well as the output (the test signal) of the system (32) are available:

$$\widehat{g}_{\mathbf{n}} \quad = \quad \widehat{\text{Cov}}(C, D) \widehat{\text{Cov}}(C, C)^{-1} \tag{43}$$

$$\widehat{\sigma}_{V,\mathbf{n}}^2 \quad = \quad \widehat{\text{Cov}}(D, D) - \widehat{g}_{\mathbf{n}} \widehat{\text{Cov}}(C, D) \tag{44}$$

where the covariances are approximated by sample estimates using sample points from the corresponding blocks centered at coefficient $\mathbf{n}$ in the reference and the test signals.

For VIF, the HVS model is parameterized by only one parameter: the variance of visual noise σ_H^2. It is easy to hand-optimize the value of the parameter σ_H^2 by running the algorithm over a range of values and observing its performance.

4.2 Image Quality Assessment Using Information Theoretic Metrics

Firstly, note that the IFC is bounded below by zero (since mutual information is a non-negative quantity) and bounded above by ∞, which occurs when the reference and test images are identical. One advantage of the IFC is that like the MSE, it does not depend upon model parameters such as those associated with display device physics, data from visual psychology experiments, viewing configuration information, or stabilizing constants.

Note that VIF is basically IFC normalized by the reference image information. The VIF has a number of interesting features. Firstly, note that VIF is bounded below by zero, which indicates that all information about the reference image has been lost in the distortion channel. Secondly, if the test image is an exact copy of the reference image, then VIF is *exactly* unity (this property is satisfied by the SSIM index also). For many distortion types, VIF would lie in the interval $[0, 1]$. Thirdly, a linear contrast enhancement of the reference image that does not add noise would result in a VIF value *larger* than unity, signifying that the contrast enhanced image has a *superior* visual quality than the reference image! It is common observation that contrast enhancement of images increases their perceptual quality unless quantization, clipping, or display non-linearities add additional distortion. This improvement in visual quality is captured by VIF.

We now illustrate the performance of VIF by example. Figure 11 shows a reference image and three of its distorted versions that come from three different types of distortion, all of which have been adjusted to have about the same MSE with the reference image. The distortion types illustrated in Figure 11 are contrast stretch, Gaussian blur and JPEG compression. In comparison with the reference image, the contrast enhanced image has a better visual quality despite the fact that the 'distortion' (in terms of a perceivable difference with the reference image) is clearly visible. A VIF value larger than unity indicates that the perceptual difference in fact constitutes improvement in visual quality. In contrast, both the blurred image and the JPEG compressed image have clearly visible distortions and poorer visual quality, which is captured by a low VIF measure.

Figure 12 illustrates spatial quality maps generated by VIF. Figure 12(a) shows a reference image and Figure 12(b) the corresponding JPEG2000 compressed image in which the distortions are clearly visible. Figure 12(c) shows the reference image information map. The information map shows the spread of statistical information in the reference image. The statistical information content of the image is low in flat image regions, whereas in textured regions and regions containing strong edges, it is high. The quality map in Figure 12(d) shows the proportion of the image information that has been lost to JPEG2000 compression. Note that due to the nonlinear normalization in the denominator of VIF, the scalar VIF value for a reference/test pair is *not* the mean of the corresponding VIF-map.

4.3 Relation to HVS Based Metrics and Structural Similarity

We will first discuss the relation between IFC and the SSIM index [10, 14]. First of all, the GSM model used in the information theoretic metrics results in the sub-band coefficients being Gaussian distributed, when conditioned on the mixing density. The linear distortion channel model results in the reference and test image being jointly Gaussian. The definition of the correlation coefficient in the SSIM index in (19) is obtained from regression analysis and implicitly assumes that the reference and test image vectors are jointly Gaussian [19]. These observations hint at the possibility that the IFC index may be closely related to SSIM. A well known result in information theory states that when two variables are jointly

Gaussian, the mutual information between them is a function of just the correlation coefficient [20, 21]. Thus, recent results show that a scalar version of the IFC metric is a monotonic function of the square of the structure term of the SSIM index when the SSIM index is applied on sub-band filtered coefficients [10, 14]. The reasons for the monotonic relationship between the SSIM index and the IFC index are the explicit assumption of a Gaussian distribution on the reference and test image coefficients in the IFC index (conditioned on the mixing density) and the implicit assumption of a Gaussian distribution in the SSIM index (due to the use of regression analysis). These results indicate that the IFC index is equivalent to multi-scale SSIM indices since they satisfy a monotonic relationship.

Further, the concept of the correlation coefficient in SSIM was generalized to vector valued variables using canonical correlation analysis to establish a monotonic relation between the squares of the canonical correlation coefficients and the vector IFC index [10,14]. It was also established that the VIF index includes a structure comparison term and a contrast comparison term (similar to the SSIM index), as opposed to just the structure term in IFC. One of the properties of the VIF index observed in Section 4.2 was the fact that it can predict improvement in quality due to contrast enhancement. The presence of the contrast comparison term in VIF explains this effect [10, 14].

We showed the relation between SSIM and HVS based metrics in Section 3.3. From our discussion here, the relation between IFC, VIF and HVS based metrics are also immediately apparent. Similarities between the scalar IFC index and HVS based metrics were also observed in [23]. It was shown that the IFC is functionally similar to HVS based FR QA algorithms [23]. The reader is referred to [10, 14] for a more thorough treatment of this subject.

Having discussed the similarities between the structural similarity and the information theoretic frameworks, we will now discuss the differences between them. The structural similarity metrics use a measure of *linear* dependence between the reference and test image pixels, namely the Pearson product moment correlation coefficient. However, the information theoretic metrics use the mutual information, which is a more general measure of correlation that can capture non-linear dependencies between variables. The reason for the monotonic relation between the square of the structure term of the SSIM index applied in the sub-band filtered domain and the IFC index is due to the assumption that the reference and test image coefficients are jointly Gaussian. This indicates that the structure term of SSIM and the IFC are equivalent under the statistical source model used in [23] and more sophisticated statistical models are required in the IFC framework to distinguish it from the SSIM index.

Although the information theoretic metrics use a better notion of correlation than the structural similarity philosophy, the form of the relationship between the reference and test images might affect visual quality. As an example, if one test image is a deterministic linear function of the reference image, while another test image is a deterministic parabolic function of the reference, the mutual information between the reference and the test image is identical in both cases. However, it is unlikely that the visual quality of both images are identical. We believe that further investigation of suitable models for the distortion channel and the relation between such channel models and visual quality are required to answer this question.

5 Performance of Image Quality Metrics

In this section we present results on the validation of some of the image quality metrics presented in this chapter and present comparisons with PSNR. All results use the LIVE image QA database [5] developed by Bovik and co-workers and further details can be found in [4]. The validation is done using subjective quality scores obtained from a group of human observers, and the performance of the QA algorithms is evaluated by comparing the quality predictions of the algorithms against subjective scores.

In the LIVE database, $20 - 28$ human subjects were asked to assign each image with a score indicating their assessment of the quality of that image, defined as the extent to which the artifacts were visible and annoying. Twenty-nine high-resolution 24-bits/pixel RGB color images (typically 768×512) were distorted using five distortion types: JPEG2000, JPEG, white noise in the RGB components, Gaussian blur, and transmission errors in the JPEG2000 bit stream using a fast-fading Rayleigh channel model. A database was derived from the 29 images to yield a total of 779 distorted images, which, together with the undistorted images, were then evaluated by human subjects. The raw scores were processed to yield Difference Mean Opinion Scores (DMOS) scores for validation and testing.

Usually, the predicted quality scores from a QA method are fitted to the subjective quality scores using a monotonic non-linear function to account for any non-linearities in the objective model. Numerical methods are used to do this fitting. For the results presented here, a five-parameter non-linearity (a logistic function with additive linear term) was used and the mapping function used is given by

$$\text{Quality}(x) = \beta_1 \text{logistic}\left(\beta_2, (x - \beta_3)\right) + \beta_4 x + \beta_5 \tag{45}$$

$$\text{logistic}(\tau, x) = \frac{1}{2} - \frac{1}{1 + \exp(\tau x)} \tag{46}$$

Table 1 quantifies the performance of the various methods in terms of well known validation quantities: the linear correlation coefficient (LCC) between objective model prediction and subjective quality and the spearman rank order correlation coefficient (SROCC) between them. Clearly, several of these quality metrics correlate very well with visual perception. The performance of IFC and multi-scale SSIM indices are comparable, which is not surprising in view of the discussion in Section 4.3. Interestingly, the SSIM index correlates very well with visual perception despite its simplicity and ease of computation.

6 Conclusion

Hopefully, the reader has captured an understanding of the basic principles and difficulties underlying the problem of image QA. Even when there is a reference image available, as we have assumed in this chapter, the problem remains difficult owing to the subtleties and remaining mysteries of human visual perception. Hopefully, the reader has also found that recent progress has been significant, and that image QA algorithms exist that correlate quite highly with human judgments. Ultimately, it is hoped that confidence in these algorithms will become high enough that image quality algorithms can be used as surrogates for human subjectivity.

Naturally, significant problems remain. The use of partial image information instead of a reference image - so-called reduced reference image QA - presents interesting opportunities where good performance can be achieved in realistic applications where only partial about the reference image may be available. More difficult yet is the situation where no reference image information is available. This problem, called No-Reference or Blind image QA is very difficult to approach unless there is at least some information regarding the types of distortions that might be encountered [2].

An interesting direction for future work is the further use of image quality assessment algorithms as objective functions for image optimization problems. For example, the SSIM index has been used to optimize several important image processing problems, including image restoration, image quantization, and image de-noising [6–9]. Another interesting line of inquiry is the use of image quality algorithms - or variations of them - for other purposes than image quality assessment - such as speech quality assessment [1].

Lastly, we have not covered methods for assessing the quality of digital videos. There are many sources of distortion that may occur owing to time-dependent processing of videos, and interesting aspects of spatio-temporal visual perception come into play when developing algorithms for video QA. Such algorithms are by necessity more involved in their construction and complex in their execution. The reader is encouraged to read Chapter 14 of the companion volume, The Essential Guide to Video Processing, for a thorough discussion of this topic.

References

[1] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? A new look at signal fidelity measures," *IEEE Signal Process. Mag.*, to appear, 2008.

[2] Z. Wang and A. C. Bovik, *Modern image quality assessment*. Morgan and Claypool Publishing Co., 2006.

[3] A. B. Watson, "DCTune: a technique for visual optimization of dct quantization matrices for individual images," *Society for Information Display Digest of Technical Papers*, vol. 24, pp. 946–949, 1993.

[4] H. R. Sheikh, M. F. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[5] (2003) LIVE image quality assessment database. [Online]. Available: `http://live.ece.utexas.edu/research/quality/subjective.htm`

[6] S. S. Channappayya, A. C. Bovik, and R. W. Heath, Jr., "A linear estimator optimized for the structural similarity index and its application to image denoising," in *IEEE Intl. Conf. Image Process.*, Atlanta, GA, Jan. 2006.

[7] S. S. Channappayya, A. C. Bovik, C. Caramanis, and R. W. Heath, Jr., "Design of linear equalizers optimized for the structural similarity index," *IEEE Trans. Image Process.*, to appear, 2008.

[8] S. S. Channappayya, A. C. Bovik, and R. W. Heath, Jr., "Rate bounds on SSIM index of quantized images," *IEEE Trans. Image Process.*, to appear, 2008.

[9] S. S. Channappayya, A. C. Bovik, R. W. Heath, Jr., and C. Caramanis, "Rate bounds on the SSIM index of quantized image DCT coefficients," in *Data Compression Conf.*, Snowbird, Utah, Mar. 2008.

[10] K. Seshadrinathan and A. C. Bovik, "Unifying analysis of full reference image quality assessment," to appear in IEEE Intl. Conf. on Image Proc., 2008.

[11] M. J. Wainwright, E. P. Simoncelli, and A. S. Willsky, "Random cascades on wavelet trees and their use in analyzing and modeling natural images," *Applied and Computational Harmonic Analysis*, vol. 11, no. 1, pp. 89–123, Jul. 2001.

[12] E. P. Simoncelli and W. T. Freeman, "The steerable pyramid: a flexible architecture for multi-scale derivative computation," in *Proc. Intl. Conf. on Image Proc.*, vol. vol.3, Jan. 1995.

[13] Z. Wang and E. P. Simoncelli, "Stimulus synthesis for efficient evaluation and refinement of perceptual image quality metrics," in *Proc. SPIE*, vol. 5292, no. 1, Jan. 2004, pp. 99–108.

[14] K. Seshadrinathan and A. C. Bovik, "Unified treatment of full reference image quality assessment algorithms," submitted to the IEEE Trans. on Image Proc.

[15] D. J. Heeger, "Normalization of cell responses in cat striate cortex." *Vis Neurosci*, vol. 9, no. 2, pp. 181–197, Aug 1992.

[16] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Thirty-Seventh Asilomar Conf. on Signals, Systems and Computers*, Pacific Grove, CA, 2003.

[17] Z. Wang and E. P. Simoncelli, "Translation insensitive image similarity in complex wavelet domain," in *IEEE Intl. Conf. Acoustics, Speech, and Signal Process.*, Philadelphia, PA, 2005.

[18] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," in *Adv. Neural Inf. Proc. Sys.*, S. A. Solla, T. Leen, and S.-R. Muller, Eds., vol. 12, 1999.

[19] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. John Wiley and Sons, 1984.

[20] I. M. Gelfand and A. M. Yaglom, "Calculation of the amount of information about a random function contained in another such function," *Amer. Math. Soc. Transl.*, vol. 12, no. 2, pp. 199–246, 1959.

[21] S. Kullback, *Information Theory and Statistics*. Dover Publications, 1968.

[22] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[23] H. R. Sheikh, A. C. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, 2005.

[24] J. Ross and H. D. Speed, "Contrast adaptation and contrast masking in human vision," *Proc. Biol. Sci.*, vol. 246, no. 1315, pp. 61–70, Oct. 1991.

[25] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, 2002.

[26] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[27] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, 2007.

[28] A. Watson and J. Solomon, "Model of visual contrast gain control and pattern masking," *J. Opt. Soc. Am. A (Optics, Image Science and Vision)*, vol. 14, no. 9, pp. 2379–2391, Sep. 1997.

[29] O. Schwartz and E. P. Simoncelli, "Natural signal statistics and sensory gain control." *Nat. Neurosci.*, vol. 4, no. 8, pp. 819–825, Aug 2001.

[30] J. Foley, "Human luminance pattern-vision mechanisms: masking experiments require a new model," *J. Opt. Soc. Am. A (Optics and Image Science)*, vol. 11, no. 6, pp. 1710–1719, Jun. 1994.

[31] D. G. Albrecht and W. S. Geisler, "Motion selectivity and the contrast-response function of simple cells in the visual cortex." *Vis. Neurosci.*, vol. 7, no. 6, pp. 531–546, Dec 1991.

[32] R. Shapley and C. Enroth-Cugell, "Visual adaptation and retinal gain controls," *Progress in Retinal Research*, vol. 3, pp. 263–346, 1984.

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

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

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

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

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

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

[44] I. Höntsch and L. J. Karam, "Locally adaptive perceptual image coding," *IEEE Trans. Image Processing*, vol. 9, pp. 1472–1483, Sept. 2000.

[45] I. Höntsch and L. J. Karam, "Adaptive image coding with perceptual distortion control," *IEEE Trans. Image Processing*, vol. 9, pp. 1472–1483, Sept. 2000.

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

[47] 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.

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

[49] 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.

[50] 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.

[51] 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.

[52] T. N. Pappas and D. L. Neuhoff, "Least-squares model-based halftoning," *IEEE Trans. Image Processing*, vol. 8, pp. 1102–1116, Aug. 1999.

[53] 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.

[54] 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.

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

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

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

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

[59] 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.

[60] 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.

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

[62] G. E. Legge and J. M. Foley, "Contrast masking in human vision," *Journal of the Optical Society of America*, vol. 70, no. 12, pp. 1458–1471, 1980.

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

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

[65] A. J. Seyler and Z. L. Budrikas, "Detail perception after scene change in television image presentations," *IEEE Trans. Inform. Theory*, vol. IT-11, no. 1, pp. 31–43, 1965.

[66] Y. Ninomiya, T. Fujio, and F. Namimoto, "Perception of impairment by bit reduction on cut-changes in television pictures. (in japanese)," *Electrical Communication Association Essay Periodical*, vol. J62-B, no. 6, pp. 527–534, 1979.

[67] 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.

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

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

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

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

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

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

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

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

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

[77] 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.

[78] 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), pp. 83–91, Feb. 1995.

[79] 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* (V. Bhaskaran, F. Sijstermans, and S. Panchanathan, eds.), vol. Proc. SPIE, Vol. 2668, (San Jose, CA), pp. 450–461, Jan./Feb. 1996.

[80] J. Chen and T. N. Pappas, "Perceptual coders and perceptual metrics," in *Human Vision and Electronic Imaging VI* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE Vol. 4299, (San Jose, CA), pp. 150–162, Jan. 2001.

[81] 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 Technol.*, vol. 6, pp. 243–250, June 1996.

[82] 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.

[83] B. E. Usevitch, "A tutorial on modern lossy wavelet image compression: Foundations of JPEG 2000," *IEEE Signal Processing Mag.*, vol. 18, pp. 22–35, Sept. 2001.

[84] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Processing Mag.*, vol. 18, pp. 36–58, Sept. 2001.

[85] D. S. Taubman and M. W. Marcellin, "JPEG2000: Standard for interactive imaging," *Proc. IEEE*, vol. 90, pp. 1336–1357, Aug. 2002.

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

[87] D. Taubman, "High performance scalable image compression with ebcot," *IEEE Trans. Image Processing*, vol. 9, pp. 1158–1170, July 2000.

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

[89] A. J. Ahumada, Jr. and H. A. Peterson, "An visual detection model for DCT coefficient quantization," in *AIAA Computing in Aerospace 9: A Collection of Technical Papers*, (San Diego, CA), pp. 314–317, Oct. 1993.

[90] M. P. Eckert and A. P. Bradley, "Perceptual quality metrics applied to still image compression," *Signal Processing*, vol. 70, pp. 177–200, 1998.

[91] S. Daly, "Subroutine for the generation of a two dimensional human visual contrast sensitivity function," Technical Report 233203Y, Eastman Kodak, 1987.

[92] 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.

[93] 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.

[94] 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.

[95] N. Bekkat and A. Saadane, "Coded image quality assessment based on a new contrast masking model," *Journal of Electronic Imaging*, vol. 13, pp. 341–348, Apr. 2004.

[96] S. Winkler and S. Süsstrunk, "Visibility of noise in natural images," in *Human Vision and Electronic Imaging IX*, vol. Proc. SPIE, Vol. 5292, (San Jose, CA), Jan. 2004.

[97] 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.

[98] 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), pp. 197–206, Jan. 1999.

[99] 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.

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

[101] M. G. Ramos and S. S. Hemami, "Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis," *J. Opt. Soc. Am. A*, Oct. 2001.

[102] D. M. Chandler and S. S. Hemami, "Additivity models for suprathreshold distortion in quantized wavelet-coded images," in *Human Vision and Electronic Imaging VII* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE Vol. 4662, (San Jose, CA), pp. 105–118, Jan. 2002.

[103] D. M. Chandler and S. S. Hemami, "Effects of natural images on the detectability of simple and compound wavelet subband quantization distortions," *J. Opt. Soc. Am. A*, vol. 20, July 2003.

[104] D. M. Chandler and S. S. Hemami, "Suprathreshold image compression based on contrast allocation and global precedence," in *Human Vision and Electronic Imaging VIII* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE Vol. 5007, (Santa Clara, CA), Jan. 2003.

[105] D. M. Chandler and S. S. Hemami, "Contrast-based quantization and rate control for wavelet-coded images," in *Proc. Int. Conf. Image Processing (ICIP-02)*, (Rochester, NY), Sept. 2002.

[106] T. N. Pappas, J. P. Allebach, and D. L. Neuhoff, "Model-based digital halftoning," *IEEE Signal Processing Mag.*, vol. 20, pp. 14–27, July 2003.

[107] 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.

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

[109] S. Winkler, "Quality metric design: a closer look," in *Human Vision and Electronic Imaging V* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE Vol. 3959, (San Jose, CA), Jan. 2000.

[110] S. Winkler, "Visual fdelity and perceived quality: towards comprehensive metrics," in *Human Vision and Electronic Imaging VI* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE Vol. 4299, (San Jose, CA), Jan. 2001.

[111] 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), pp. 218–225, Jan. 1999.

[112] 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.

[113] 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.

[114] A. B. Watson, J. Hu, and J. F. M. III, "Digital video quality metric based on human vision," *Journal of Electronic Imaging*, vol. 10, pp. 20–29, Jan. 2001.

[115] 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), pp. 168–174, Jan. 1999.

[116] R. O. Hinds and T. N. Pappas, "Effect of concealment techniques on perceived video quality," in *Human Vision and Electronic Imaging IV* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE Vol. 3644, (San Jose, CA), pp. 207–217, Jan. 1999.

[117] K. Brunnström and B. N. Schenkman, "Quality of video affected by packet loss distortion, compared to the predictions of a spatio-temporal model," in *Human Vision and Electronic Imaging VII* (B. E. Rogowitz and T. N. Pappas, eds.), vol. Proc. SPIE Vol. 4662, (San Jose, CA), Jan. 2002.

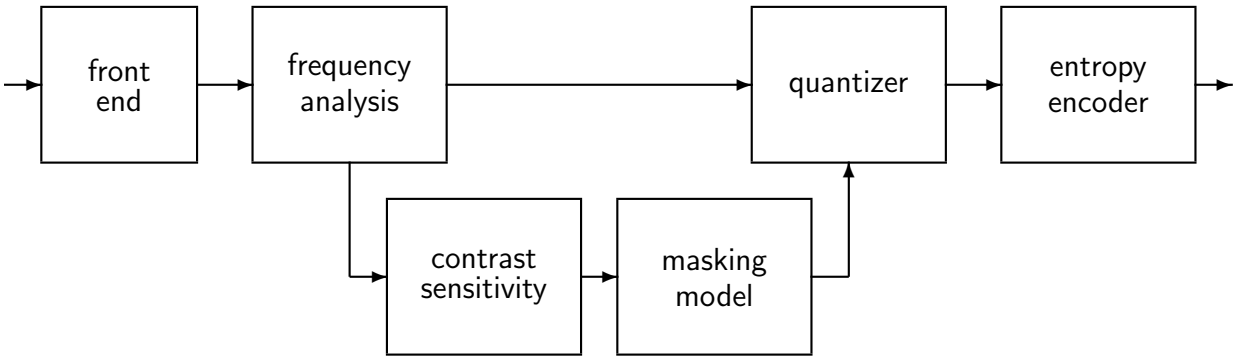


Figure 1: Perceptual coder

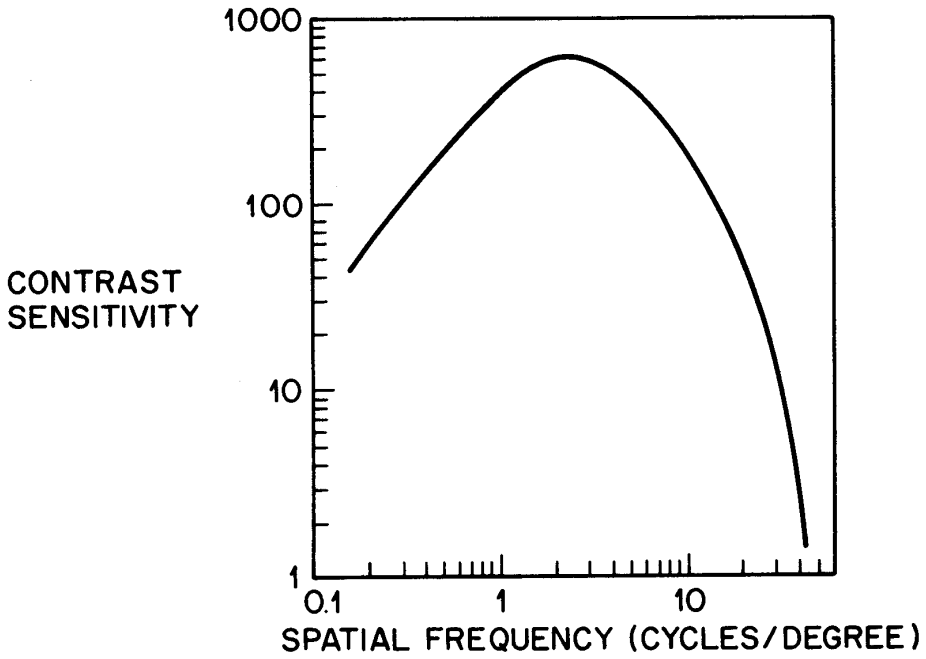


Figure 2: Spatial contrast sensitivity function (Reprinted with permission from reference [63], page 269)

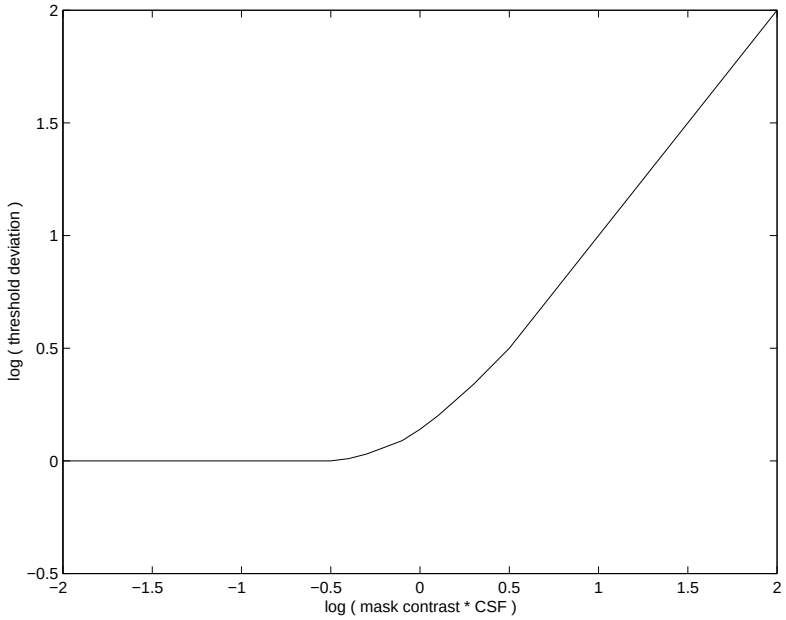
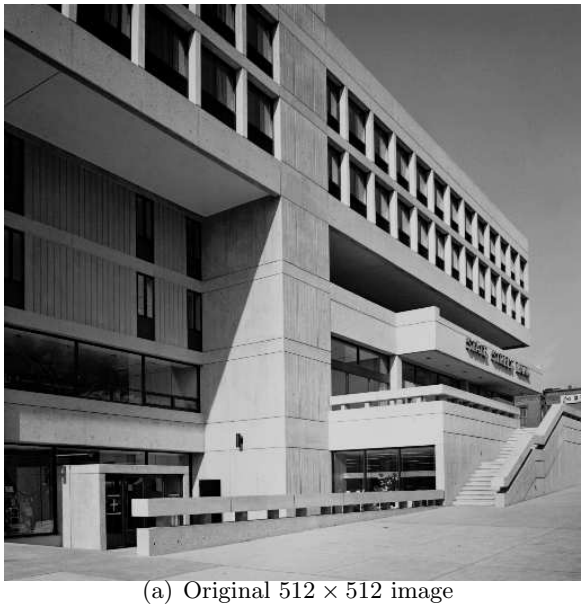
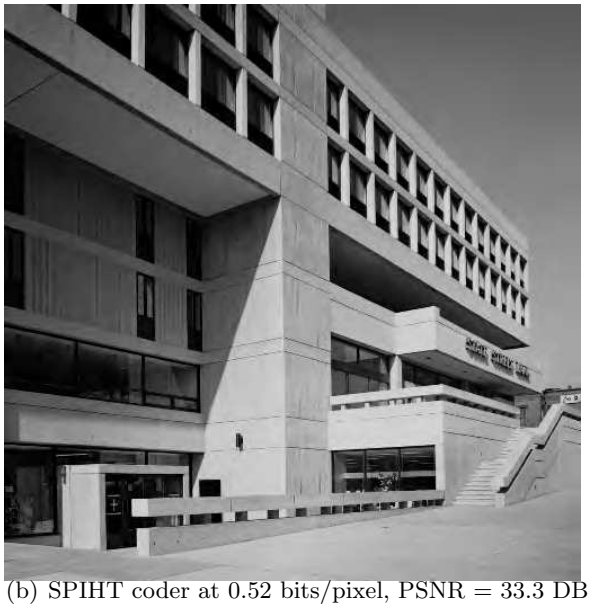


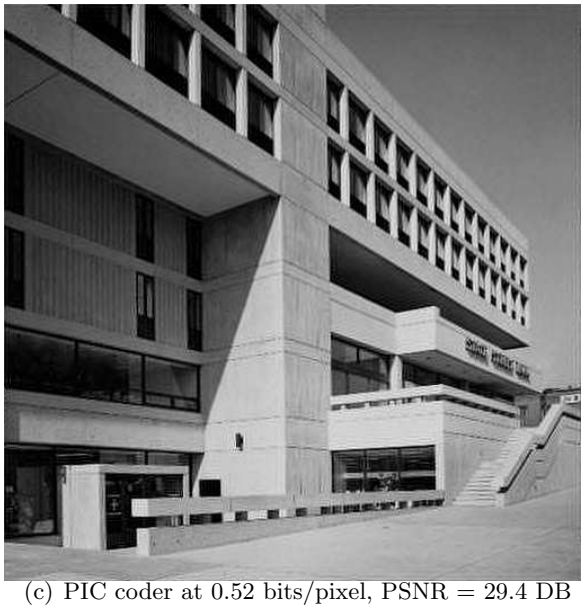
Figure 3: Contrast masking function



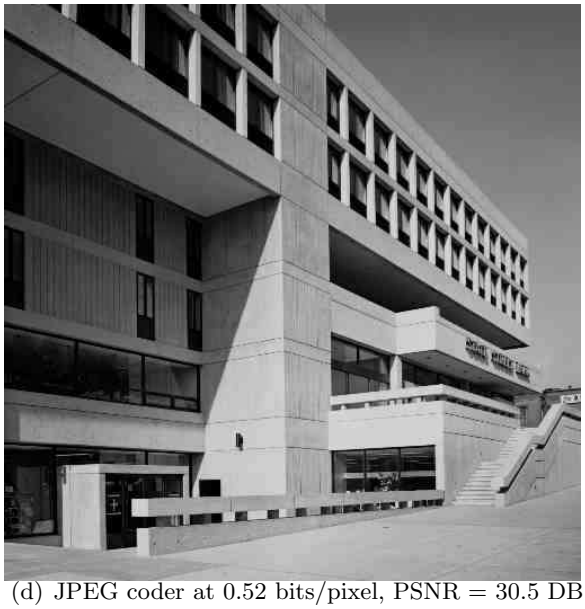
(a) Original 512×512 image



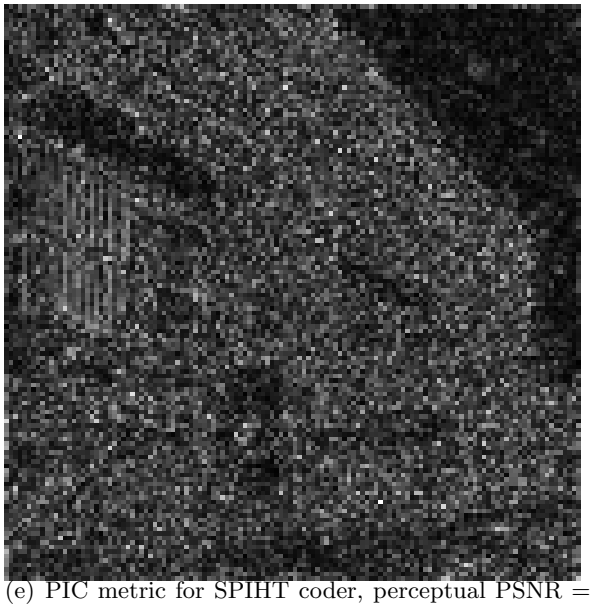
(b) SPIHT coder at 0.52 bits/pixel, PSNR = 33.3 DB



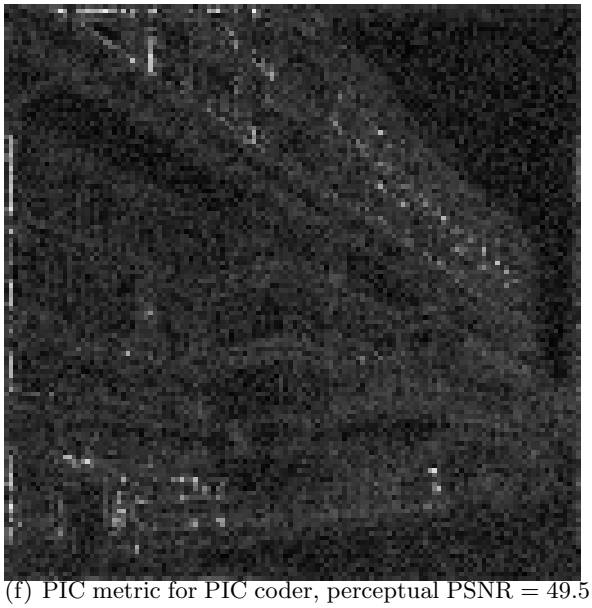
(c) PIC coder at 0.52 bits/pixel, PSNR = 29.4 DB



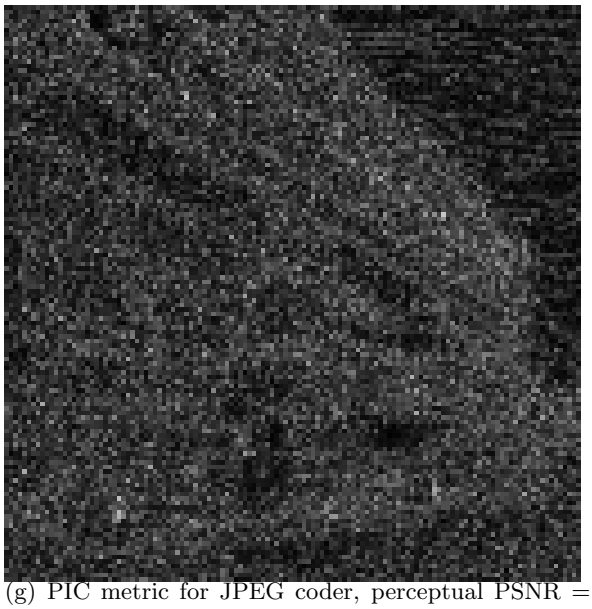
(d) JPEG coder at 0.52 bits/pixel, PSNR = 30.5 DB



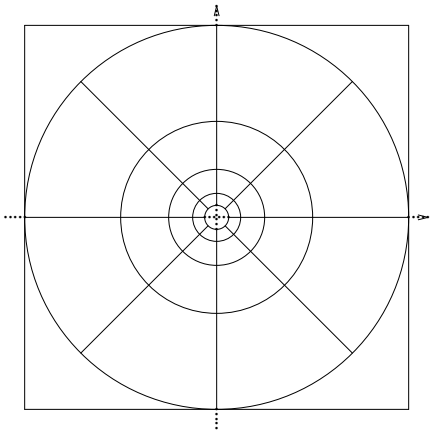
(e) PIC metric for SPIHT coder, perceptual PSNR = 46.8 DB



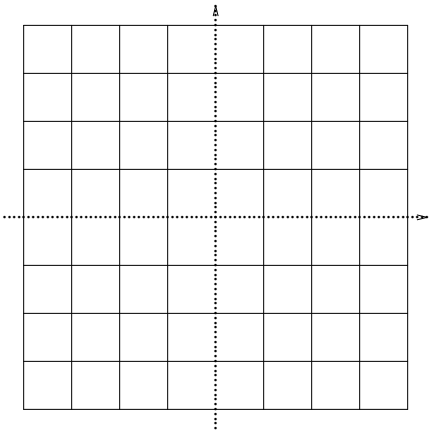
(f) PIC metric for PIC coder, perceptual PSNR = 49.5 DB



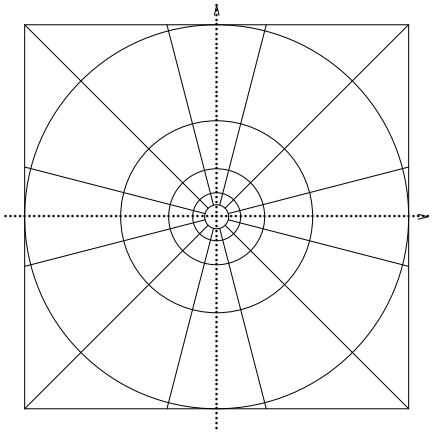
(g) PIC metric for JPEG coder, perceptual PSNR = 47.9 DB



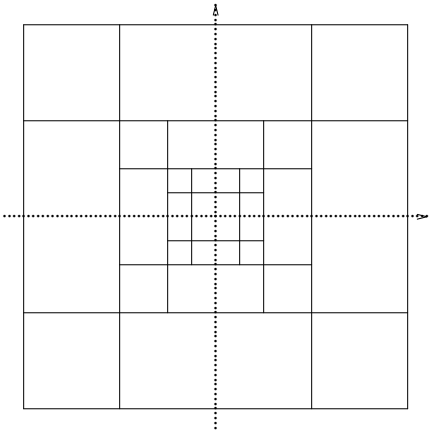
(a) Cortex transform (Watson)



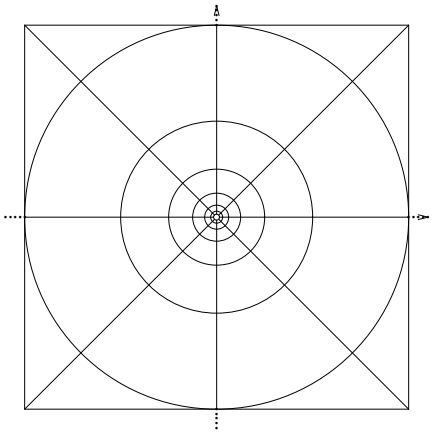
(d) Subband transform



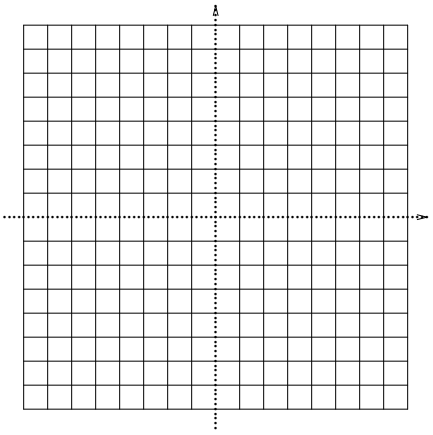
(b) Cortex transform (Daly)



(e) Wavelet transform



(c) Lubin's transform



(f) DCT transform

Figure 4: 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.

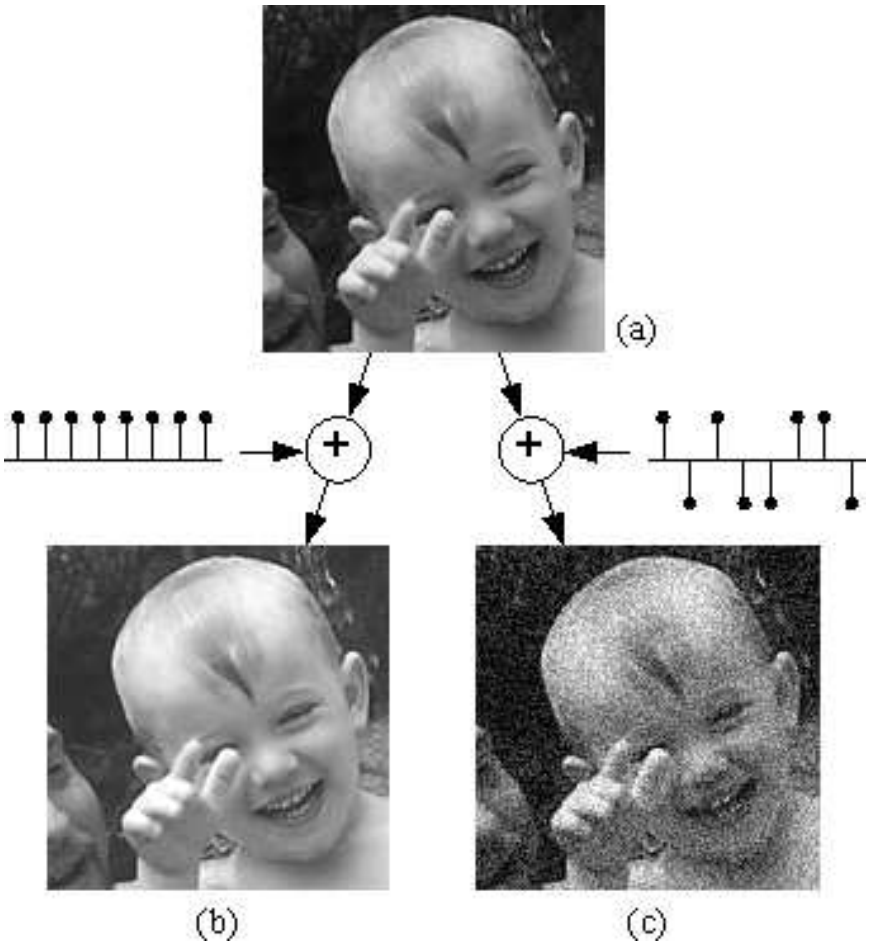


Figure 5: Failure of the Minkowski metric for image quality prediction. (a) original image; (b) distorted image by adding a positive constant; (c) distorted image by adding the same constant, but with random sign. Images (b) and (c) have the same Minkowski metric with respect to image (a), but drastically different visual quality.

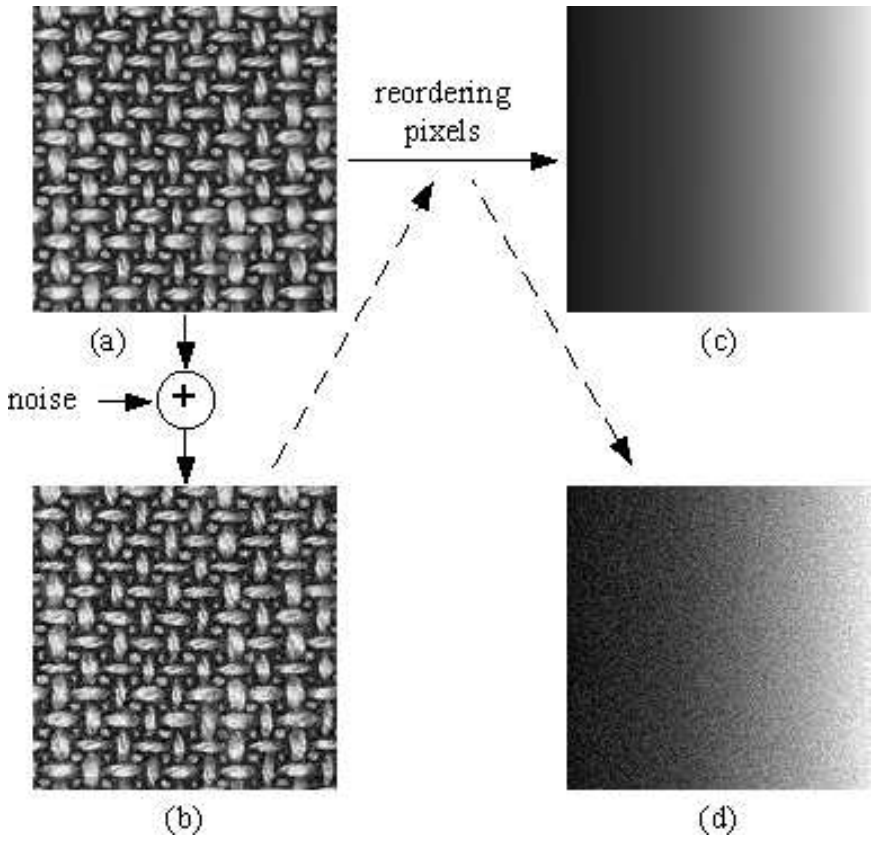


Figure 6: Failure of the Minkowski metric for image quality prediction. (a) original texture image; (b) distorted image by adding independent white Gaussian noise; (c) reordering of the pixels in image (a) (by sorting pixel intensity values); (d) reordering of the pixels in image (b), by following the same reordering used to create image (c). The Minkowski metrics between images (a) and (b) and images (c) and (d) are the same, but image (d) appears much noisier than image (b).

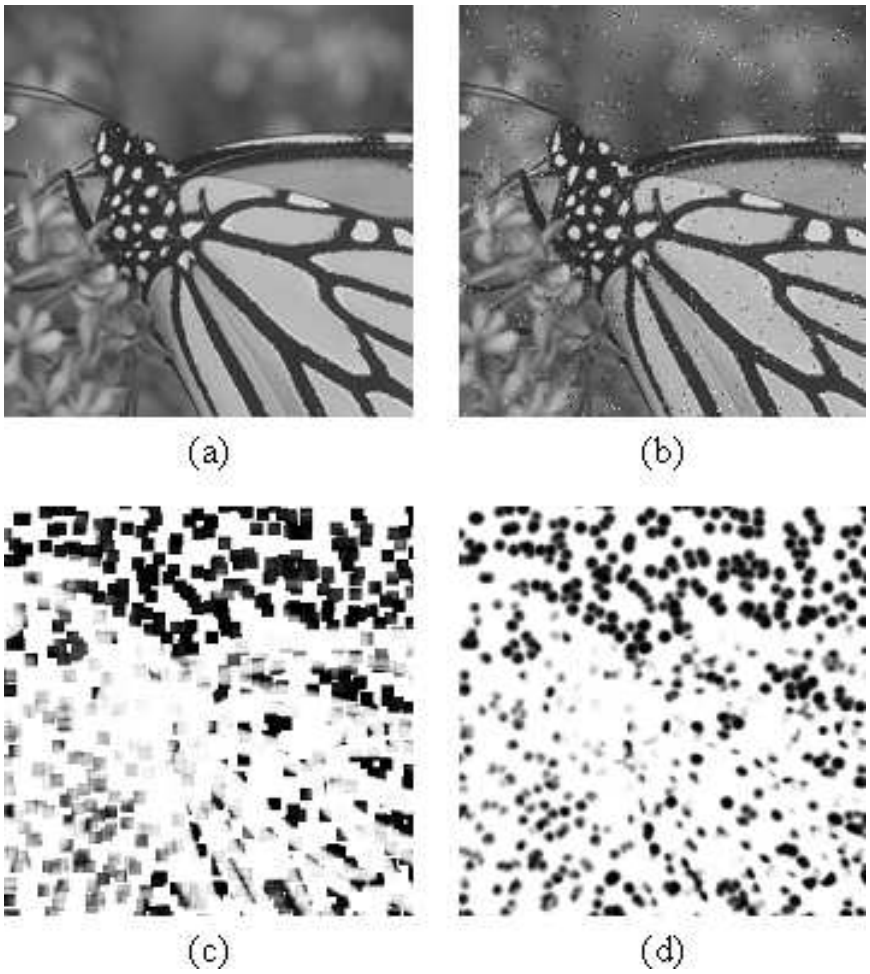


Figure 7: The effect of local window shape on SSIM index map. (a) original image; (b) impulsive noise contaminated image; (c) SSIM index map using square windowing approach; (d) SSIM index map using smoothed windowing approach. In both SSIM index maps, brighter indicates better quality.

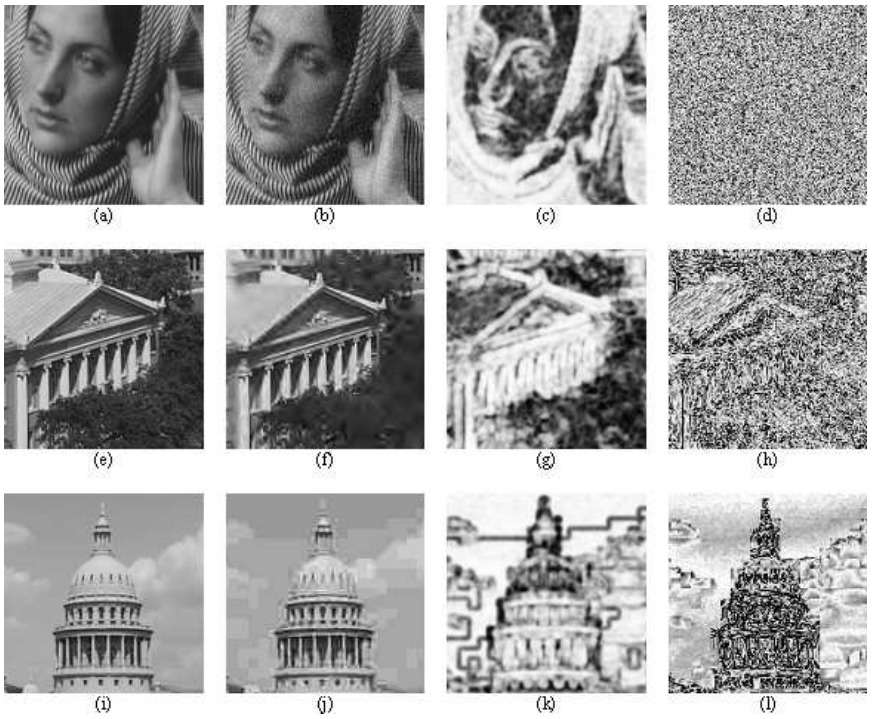


Figure 8: Sample distorted images and their quality/distortion maps (images are cropped to 160160 for visibility); (a), (e), (i): original images; (b): Gaussian noise contaminated image; (f) JPEG2000 compressed image; (j) JPEG compressed images; (c), (g), (k): SSIM index maps of the distorted images, where brightness indicates the magnitude of the local SSIM index (squared for visibility); (d), (h), (l): absolute error maps of the distorted images, where darkness indicates the absolute value of the local pixel difference. Note that in all quality/distortion maps ((c), (d), (g), (h), (k) and (l)), brighter indicates better quality in terms of the underlying quality/distortion measure.

Figure 9: The information-fidelity problem: A channel distorts images and limits the amount of information that could flow from the source to the receiver. Quality should relate to the amount of information about the reference image that could be extracted from the test image.

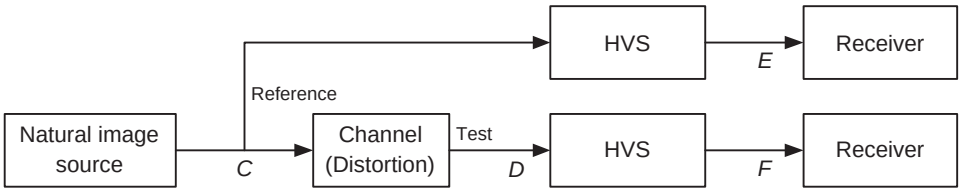
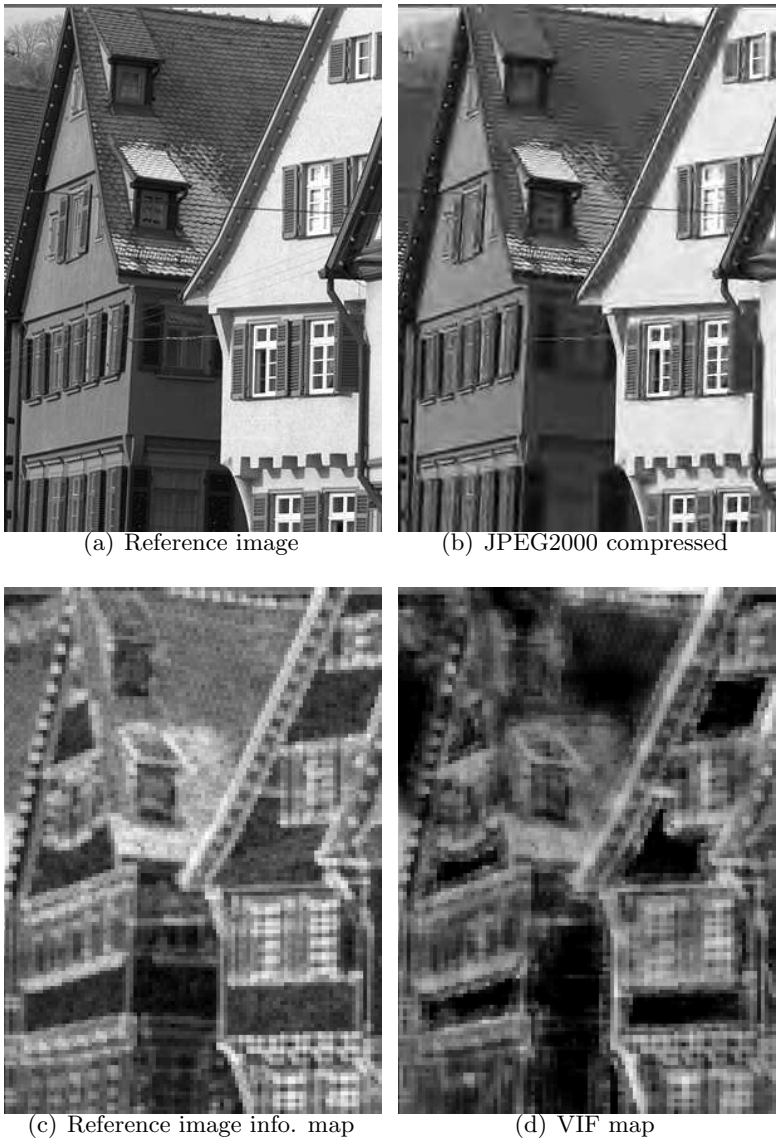


Figure 10: An information-theoretic setup for quantifying visual quality using a distortion channel model as well as an HVS model. The HVS also acts as a channel that limits the flow of information from the source to the receiver. Image quality could also be quantified using a relative comparison of the information in the upper path of the figure and the information in the lower path.

(a) Reference image (b) Contrast enhancement

(c) Blurred (d) JPEG compressed

Figure 11: The VIF has an interesting feature: it can capture the effects of linear contrast enhancements on images, and quantify the improvement in visual quality. A VIF value greater than unity indicates this improvement, while a VIF value less than unity signifies a loss of visual quality. (a) Reference Lena image (VIF = 1.0). (b) Contrast stretched Lena image (VIF = 1.17). (c) Gaussian blur (VIF = 0.05) and (d) JPEG compressed (VIF = 0.05).



(a) Reference image (b) JPEG2000 compressed

(c) Reference image info. map (d) VIF map

Figure 12: Spatial maps showing how VIF captures spatial information loss.

Performance		
Model	LCC	SROCC
PSNR	0.8709	0.8755
Sarnoff JND	0.9266	0.9291
DCTune	0.8046	0.8032
Multi-Scale SSIM	0.9393	0.9527
IFC	0.9441	0.9459
VIF	0.9533	0.9584
VSNR	0.889	0.889

Table 1: Performance of different QA methods.