Chapter 41 in *The Handbook of Video Databases: Design and Applications*, B. Furht and O. Marqure, ed., CRC Press, pp. 1041-1078, September 2003

41

OBJECTIVE VIDEO QUALITY ASSESSMENT

Zhou Wang, Hamid R. Sheikh and Alan C. Bovik
Department of Electrical and Computer Engineering
The University of Texas at Austin
Austin, Texas, USA
zhouwang@ieee.org, hamid.sheikh@ieee.org,
bovik@ece.utexas.edu

1. INTRODUCTION

Digital video data, stored in video databases and distributed through communication networks, is subject to various kinds of distortions during acquisition, compression, processing, transmission, and reproduction. For example, lossy video compression techniques, which are almost always used to reduce the bandwidth needed to store or transmit video data, may degrade the quality during the quantization process. For another instance, the digital video bitstreams delivered over error-prone channels, such as wireless channels, may be received imperfectly due to the impairment occurred during transmission. Package-switched communication networks, such as the Internet, can cause loss or severe delay of received data packages, depending on the network conditions and the quality of services. All these transmission errors may result in distortions in the received video data. It is therefore imperative for a video service system to be able to realize and quantify the video quality degradations that occur in the system, so that it can maintain, control and possibly enhance the quality of the video data. An effective image and video quality metric is crucial for this purpose.

The most reliable way of assessing the quality of an image or video is subjective evaluation, because human beings are the ultimate receivers in most applications. The mean opinion score (MOS), which is a subjective quality measurement obtained from a number of human observers, has been regarded for many years as the most reliable form of quality measurement. However, the MOS method is too inconvenient, slow and expensive for most applications.

The goal of objective image and video quality assessment research is to design quality metrics that can predict perceived image and video quality automatically.

Generally speaking, an objective image and video quality metric can be employed in three ways:

1. It can be used to *monitor* image quality for quality control systems. For example, an image and video acquisition system can use the quality metric to monitor and automatically adjust itself to obtain the best quality image and video data. A network video server can examine the quality of the digital video transmitted on the network and control video streaming.
2. It can be employed to *benchmark* image and video processing systems and algorithms. If multiple video processing systems are available for a specific task, then a quality metric can help in determining which one of them provides the best quality results.
3. It can be embedded into an image and video processing system to *optimize* the algorithms and the parameter settings. For instance, in a visual communication system, a quality metric can help optimal design of the prefiltering and bit assignment algorithms at the encoder and the optimal reconstruction, error concealment and postfiltering algorithms at the decoder.

Objective image and video quality metrics can be classified according to the availability of the original image and video signal, which is considered to be distortion-free or perfect quality, and may be used as a reference to compare a distorted image or video signal against. Most of the proposed objective quality metrics in the literature assume that the undistorted reference signal is fully available. Although "image and video quality" is frequently used for historical reasons, the more precise term for this type of metric would be image and video *similarity* or *fidelity* measurement, or full-reference (FR) image and video quality assessment. It is worth noting that in many practical video service applications, the reference images or video sequences are often not accessible. Therefore, it is highly desirable to develop measurement approaches that can evaluate image and video quality blindly. Blind or no-reference (NR) image and video quality assessment turns out to be a very difficult task, although human observers usually can effectively and reliably assess the quality of distorted image or video without using any reference. There exists a third type of image quality assessment method, in which the original image or video signal is not fully available. Instead, certain features are extracted from the original signal and transmitted to the quality assessment system as side information to help evaluate the quality of the distorted image or video. This is referred to as reduced-reference (RR) image and video quality assessment.

Currently, the most widely used FR objective image and video distortion/quality metrics are mean squared error (MSE) and peak signal-to-noise ratio (PSNR), which are defined as:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{41.1}$$

$$\mathrm{PSNR} = 10\log_{10}\frac{L^2}{\mathrm{MSE}} \tag{41.2}$$

where N is the number of pixels in the image or video signal, and x_i and y_i are the i-th pixels in the original and the distorted signals, respectively. L is the dynamic range of the pixel values. For an 8bits/pixel monotonic signal, L is equal to 255. MSE and PSNR are widely used because they are simple to

calculate, have clear physical meanings, and are mathematically easy to deal with for optimization purposes (MSE is differentiable, for example). However, they have been widely criticized as well for not correlating well with perceived quality measurement [1-8]. In the last three to four decades, a great deal of effort has been made to develop objective image and video quality assessment methods (mostly for FR quality assessment), which incorporate perceptual quality measures by considering human visual system (HVS) characteristics. Some of the developed models are commercially available. However, image and video quality assessment is still far from being a mature research topic. In fact, only limited success has been reported from evaluations of sophisticated HVS-based FR quality assessment models under strict testing conditions and a broad range of distortion and image types [3,9-11].

This chapter will mainly focus on the basic concepts, ideas and approaches for FR image and video quality assessment. It is worth noting that a dominant percentage of proposed FR quality assessment models share a common error sensitivity based philosophy, which is motivated from psychophysical vision science research. Section 2 reviews the background and various implementations of this philosophy and also attempts to point out the limitations of this approach. In Section 3, we introduce a new way to think about the problem of image and video quality assessment and provide some preliminary results of a novel structural distortion based FR quality assessment method. Section 4 introduces the current status of NR/RR quality assessment research. In Section 5, we discuss the issues that are related to the validation of image and video quality metrics, including the recent effort by the video quality experts group (VQEG) in developing, validating and standardizing FR/RR/NR video quality metrics for television and multimedia applications. Finally, Section 6 makes some concluding remarks and provides a vision for future directions of image and video quality assessment.

2. FULL-REFERENCE QUALITY ASSESSMENT USING ERROR SENSITIVITY MEASURES

An image or video signal whose quality is being evaluated can be thought of as a sum of a perfect reference signal and an error signal. We may assume that the loss of quality is directly related to the strength of the error signal. Therefore, a natural way to assess the quality of an image is to quantify the error between the distorted signal and the reference signal, which is fully available in FR quality assessment. The simplest implementation of the concept is the MSE as given in (41.1). However, there are a number of reasons why MSE may not correlate well with the human perception of quality:

1. Digital pixel values on which the MSE is typically computed, may not exactly represent the light stimulus entering the eye.
2. The sensitivity of the HVS to the errors may be different for different types of errors, and may also vary with visual context. This difference may not be captured adequately by the MSE.
3. Two distorted image signals with the same amount of error energy may have very different types of errors.

4. Simple error summation, like the one implemented in the MSE formulation, may be markedly different from the way the HVS and the brain arrives at an assessment of the perceived distortion.

In the last three decades, most of the proposed image and video quality metrics have tried to improve upon the MSE by addressing the above issues. They have followed an error sensitivity based paradigm, which attempts to analyze and quantify the error signal in a way that simulates the characteristics of human visual error perception. Pioneering work in this area was done by Mannos and Sakrison [12], and has been extended by other researchers over the years. We shall briefly describe several of these approaches in this section. But first, a brief introduction to the relevant physiological and psychophysical components of the HVS will aid in the understanding of the algorithms better.

2.1 THE HUMAN VISUAL SYSTEM

Figure 41.1 schematically shows the early stages of the HVS. It is not clearly understood how the human brain extracts higher-level cognitive information from the visual stimulus in the later stages of vision, but the components of the HVS depicted in Figure 41.1 are fairly well understood and accepted by the vision science community. A more detailed description of the HVS may be found in [13-15].

2.1.1 Anatomy of the HVS

The visual stimulus in the form of light coming from objects in the environment is focussed by the optical components of the eye onto the retina, a membrane at the back of the eyes that contains several layers of neurons, including photoreceptor cells. The optics consists of the cornea, the pupil (the aperture that controls the amount of light entering the eye), the lens and the fluids that fill the eye. The optical system focuses the visual stimulus onto the retina, but in doing so blurs the image due to the inherent limitations and imperfections. The blur is low-pass, typically modelled as a linear space-invariant system characterized by a point spread function (PSF). Photoreceptor cells in the retina sample the image that is projected onto it.

There are two types of photoreceptor cells in the retina: the cone cells and the rod cells. The cones are responsible for vision in normal light conditions, while the rods are responsible for vision in very low light conditions, and hence are generally ignored in the modelling. There are three different types of cones, corresponding to three different light wavelengths to which they are most sensitive. The L-cones, M-cones and S-cones (corresponding to the Long, Medium and Short wavelengths at which their respective sensitivities peak) split the image projected onto the retina into three visual streams. These visual streams can be thought of as the Red, Green and Blue color components of the visual stimulus, though the approximation is crude. The signals from the photoreceptors pass through several layers of interconnecting neurons in the retina before being carried off to the brain by the optic nerve.

The photoreceptor cells are non-uniformly distributed over the surface of the retina. The point on the retina that lies on the visual axis is called the fovea (Figure 41.1), and it has the highest density of cone cells. This density falls

off rapidly with distance from the fovea. The distribution of the ganglion cells, the neurons that carry the electrical signal from the eye to the brain through the optic nerve, is also highly non-uniform, and drops off even faster than the density of the cone receptors. The net effect is that the HVS cannot perceive the entire visual stimulus at uniform resolution.

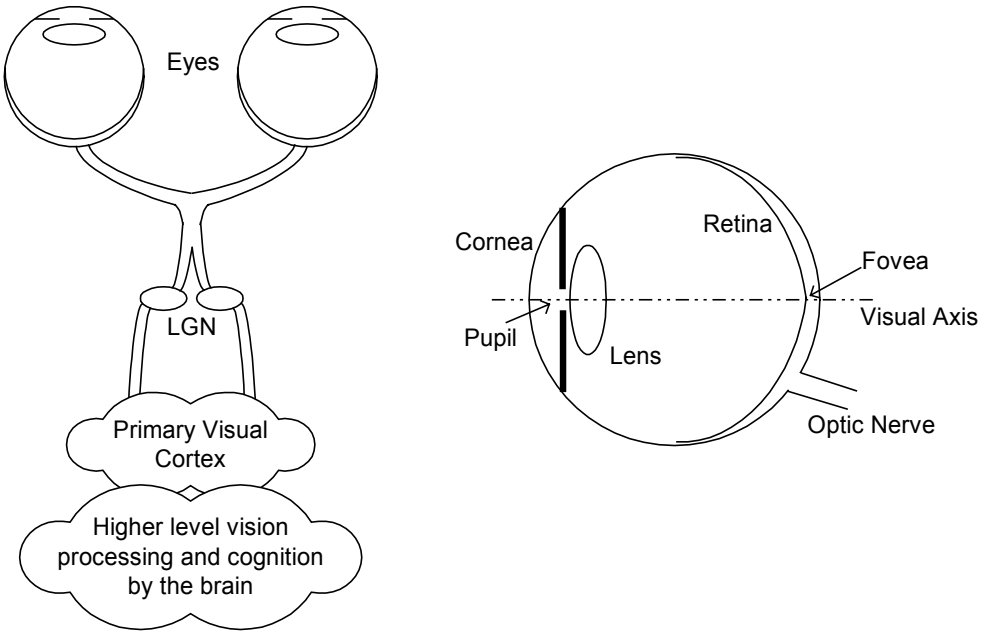


Figure 41.1 Schematic diagram of the human visual system.

The visual streams originating from the eye are reorganized in the optical chiasm and the lateral geniculate nucleus (LGN) in the brain, before being relayed to the primary visual cortex. The neurons in the visual cortex are known to be tuned to various aspects of the incoming streams, such as spatial and temporal frequencies, orientations, and directions of motion. Typically, only the spatial frequency and orientation selectivity is modelled by quality assessment metrics. The neurons in the cortex have receptive fields that are well approximated by two-dimensional Gabor functions. The ensemble of these neurons is effectively modelled as an octave-band Gabor filter bank [14,15], where the spatial frequency spectrum (in polar representation) is sampled at octave intervals in the radial frequency dimension and uniform intervals in the orientation dimension [16]. Another aspect of the neurons in the visual cortex is their saturating response to stimulus contrast, where the output of a neuron saturates as the input contrast increases.

Many aspects of the neurons in the primary visual cortex are not modelled for quality assessment applications. The visual streams generated in the cortex are carried off into other parts of the brain for further processing, such as motion sensing and cognition. The functionality of the higher layers of the HVS is currently an active research topic in vision science.

2.1.2 Psychophysical HVS Features

Foveal and Peripheral Vision

As stated above, the densities of the cone cells and the ganglion cells in the retina are not uniform, peaking at the fovea and decreasing rapidly with distance from the fovea. A natural result is that whenever a human observer fixates at a point in his environment, the region around the fixation point is resolved with the highest spatial resolution, while the resolution decreases with distance from fixation point. The high-resolution vision due to fixation by the observer onto a region is called *foveal* vision, while the progressively lower resolution vision is called *peripheral* vision. Most image quality assessment models work with foveal vision; a few incorporate peripheral vision as well [17-20]. Models may also resample the image with the sampling density of the receptors in the fovea in order to provide a better approximation of the HVS as well as providing more robust calibration of the model [17,18].

Light Adaptation

The HVS operates over a wide range of light intensity values, spanning several orders of magnitude from a moonlit night to a bright sunny day. It copes with such a large range by a phenomenon known as *light adaptation*, which operates by controlling the amount of light entering the eye through the pupil, as well as adaptation mechanisms in the retinal cells that adjust the gain of post-receptor neurons in the retina. The result is that the retina encodes the contrast of the visual stimulus instead of coding absolute light intensities. The phenomenon that maintains the contrast sensitivity of the HVS over a wide range of background light intensity is known as *Weber's Law*.

Contrast Sensitivity Functions

The contrast sensitivity function (CSF) models the variation in the sensitivity of the HVS to different spatial and temporal frequencies that are present in the visual stimulus. This variation may be explained by the characteristics of the receptive fields of the ganglion cells and the cells in the LGN, or as internal noise characteristics of the HVS neurons. Consequently, some models of the HVS choose to implement CSF as a filtering operation, while others implement CSF through weighting factors for subbands after a frequency decomposition. The CSF varies with distance from the fovea as well, but for foveal vision, the spatial CSF is typically modelled as a space-invariant band-pass function (Figure 41.2). While the CSF is slightly band-pass in nature, most quality assessment algorithms implement a low-pass version. This makes the quality assessment metrics more robust to changes in the viewing distance. The contrast sensitivity is also a function of temporal frequency, which is irrelevant for image quality assessment but has been modelled for video quality assessment as simple temporal filters [21-24].

Masking and Facilitation

Masking and facilitation are important aspects of the HVS in modelling the interactions between different image components present at the same spatial location. Masking/facilitation refers to the fact that the presence of one image component (called the *mask*) will decrease/increase the visibility of another image component (called the *test* signal). The mask generally reduces the visibility of the test signal in comparison with the case that the mask is absent. However, the mask may sometimes facilitate detection as well.

Usually, the masking effect is the strongest when the mask and the test signal have similar frequency content and orientations. Most quality assessment methods incorporate one model of masking or the other, while some incorporate facilitation as well [1,18,25].

Pooling

Pooling refers to the task of arriving at a single measurement of quality, or a decision regarding the visibility of the artifacts, from the outputs of the visual streams. It is not quite understood as to how the HVS performs pooling. It is quite obvious that pooling involves cognition, where a perceptible distortion may be more annoying in some areas of the scene (such as human faces) than at others. However, most quality assessment metrics use Minkowski pooling to pool the error signal from the different frequency and orientation selective streams, as well as across spatial coordinates, to arrive at a fidelity measurement.

2.1.3 Summary

Summarizing the above discussion, an elaborate quality assessment algorithm may implement the following HVS features:

1. Eye optics modelled by a low-pass PSF.
2. Color processing.
3. Non-uniform retinal sampling.
4. Light adaptation (luminance masking).
5. Contrast sensitivity functions.
6. Spatial frequency, temporal frequency and orientation selective signal analysis.
7. Masking and facilitation.
8. Contrast response saturation.
9. Pooling.

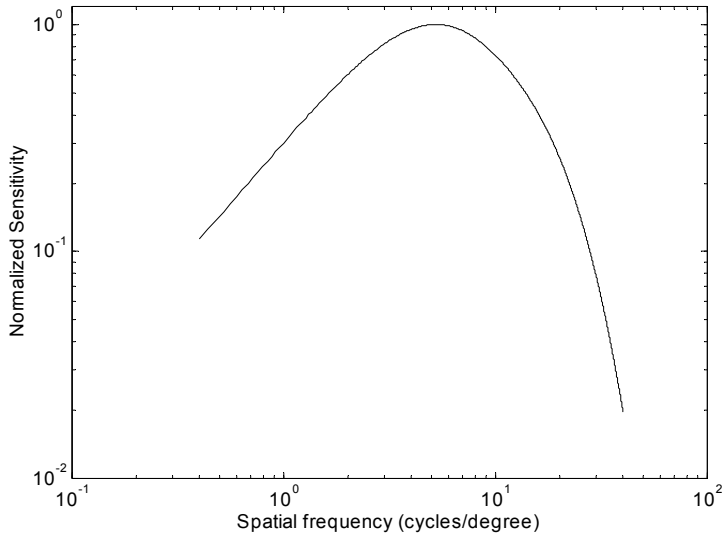


Figure 41.2 Normalized contrast sensitivity function.

2.2 GENERAL FRAMEWORK OF ERROR SENSITIVITY BASED METRICS

Most HVS based quality assessment metrics share an error-sensitivity based paradigm, which aims to quantify the strength of the errors between the reference and the distorted signals in a perceptually meaningful way. Figure 41.3 shows a generic error-sensitivity based quality assessment framework that is based on HVS modelling. Most quality assessment algorithms that model the HVS can be explained with this framework, although they may differ in the specifics.

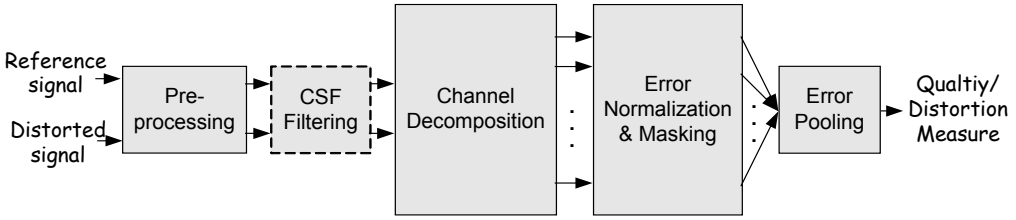


Figure 41.3 Framework of error sensitivity based quality assessment system. Note: the CSF feature can be implemented either as "CSF Filtering" or within "Error Normalization".

Pre-processing

The pre-processing stage may perform the following operations: alignment, transformations of color spaces, calibration for display devices, PSF filtering, and light adaptation. First, the distorted and the reference signals need to be properly aligned. The distorted signal may be misaligned with respect to the reference, globally or locally, for various reasons during compression, processing, and transmission. Point-to-point correspondence between the reference and the distorted signals needs to be established. Second, it is sometimes preferable to transform the signal into a color space that conforms better to the HVS. Third, quality assessment metrics may need to convert the digital pixel values stored in the computer memory into luminance values of pixels on the display device through point-wise non-linear transformations. Fourth, a low-pass filter simulating the PSF of the eye optics may be applied. Finally, the reference and the distorted videos need to be converted into corresponding contrast stimuli to simulate light adaptation. There is no universally accepted definition of contrast for natural scenes. Many models work with band-limited contrast for complex natural scenes [26], which is tied with the channel decomposition. In this case, the contrast calculation is implemented later in the system, during or after the channel decomposition process.

CSF Filtering

CSF may be implemented before the channel decomposition using linear filters that approximate the frequency responses of the CSF. However, some metrics choose to implement CSF as weighting factors for channels after the channel decomposition.

Channel Decomposition

Quality metrics commonly model the frequency selective channels in the HVS within the constraints of application and computation. The channels serve to separate the visual stimulus into different spatial and temporal *subbands*. While some quality assessment algorithms implement sophisticated channel decompositions, simpler transforms such as the wavelet transform, or even

the Discrete Cosine Transform (DCT) have been reported in the literature primarily due to their suitability for certain applications, rather than their accuracy in modelling the cortical neurons.

While the cortical receptive fields are well represented by 2D Gabor functions, the Gabor decomposition is difficult to compute and lacks some of the mathematical conveniences that are desired for good implementation, such as invertibility, reconstruction by addition, etc. Watson constructed the cortex transform [27] to model the frequency and orientation selective channels, which have similar profiles as 2D Gabor functions but are more convenient to implement. Channel decomposition models used by Watson, Daly [28,29], Lubin [17,18] and Teo and Heeger [1,25,30] attempt to model the HVS as closely as possible without incurring prohibitive implementation difficulties. The subband configurations for some of the models described in this chapter is given in Figure 41.4. Channels tuned to various temporal frequencies have also been reported in the literature [5,22,31,32].

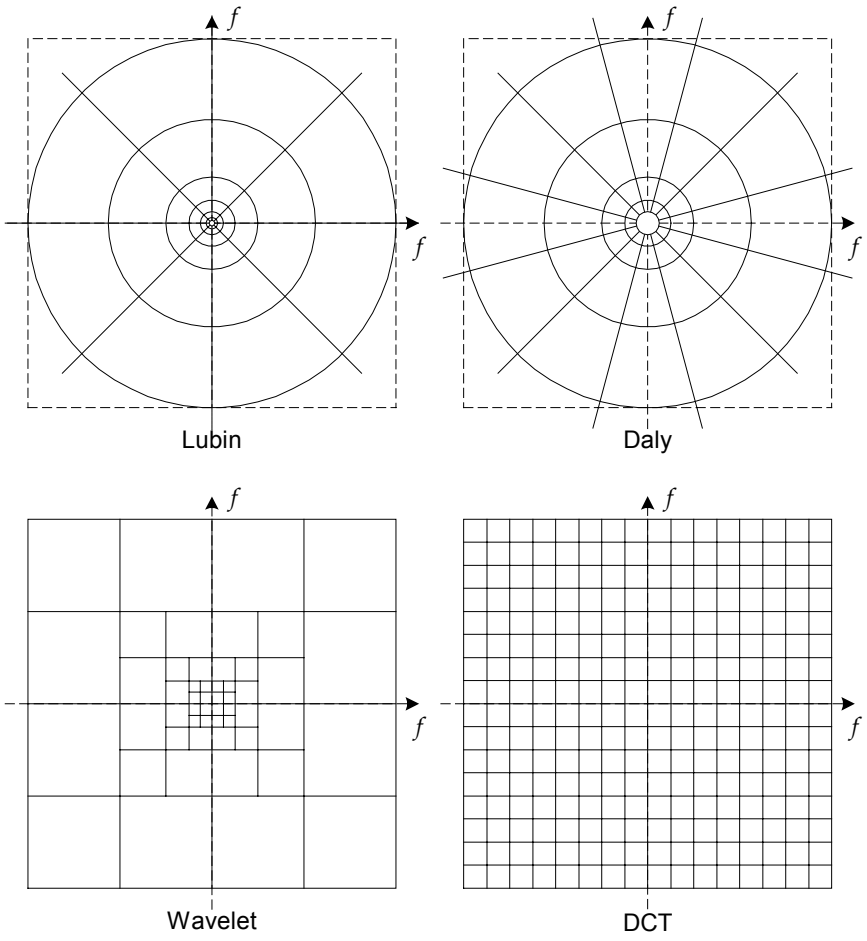
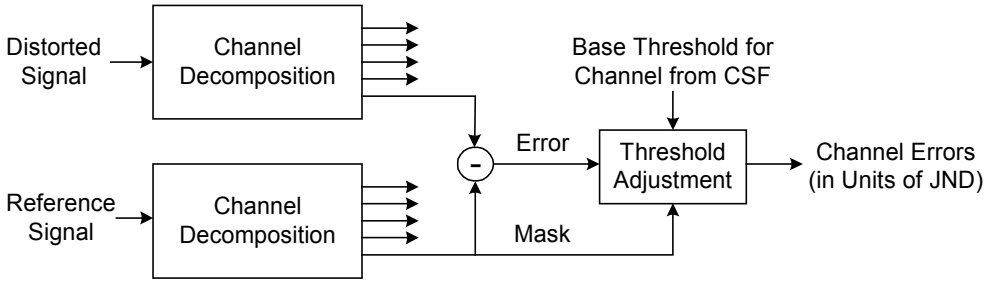


Figure 41.4 Frequency decompositions of various models.

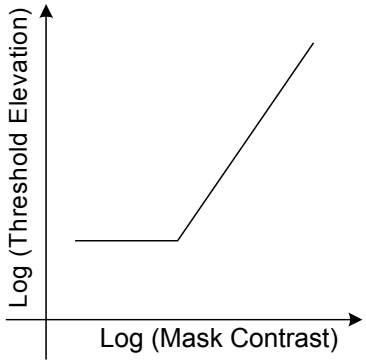
Error Normalization and Masking

Error normalization and masking is typically implemented within each channel. Most models implement masking in the form of a gain-control mechanism that weights the error signal in a channel by a space-varying

visibility threshold for that channel [33]. The visibility threshold adjustment at a point is calculated based on the energy of the reference signal (or both the reference and the distorted signals) in the neighbourhood of that point, as well as the HVS sensitivity for that channel in the absence of masking effects (also known as the *base-sensitivity*). Figure 41.5(a) shows how masking is typically implemented in a channel. For every channel the base error threshold (the minimum visible contrast of the error) is elevated to account for the presence of the masking signal. The threshold elevation is related to the contrast of the reference (or the distorted) signal in that channel through a relationship that is depicted in Figure 41.5(b). The elevated visibility threshold is then used to normalize the error signal. This normalization typically converts the error into units of Just Noticeable Difference (JND), where a JND of 1.0 denotes that the distortion at that point in that channel is just at the threshold of visibility. Some methods implement masking and facilitation as a manifestation of contrast response saturation. Figure 41.6 shows a set of curves each of which may represent the saturation characteristics of neurons in the HVS. Metrics may model masking with one or more of these curves.



(a) Masking implementation



(b) Visibility threshold model

Figure 41.5 (a) Implementation of masking effect for channel based HVS models. (b) Visibility threshold model (simplified): threshold elevation versus mask contrast.

Error Pooling

Error pooling is the process of combining the error signals in different channels into a single distortion/quality interpretation. For most quality assessment methods, pooling takes the form:

$$E = \left(\sum_{l} \sum_{k} \left| e_{l,k} \right|^{\beta} \right)^{1/\beta} \tag{41.3}$$

where $e_{l,k}$ is the normalized and masked error of the k-th coefficient in the l-th channel, and β is a constant typically with a value between 1 and 4. This form of error pooling is commonly called Minkowski error pooling. Minkowski pooling may be performed over space (index k) and then over frequency (index l), or vice-versa, with some non-linearity between them, or possibly with different exponents β. A spatial map indicating the relative importance of different regions may also be used to provide spatially variant weighting to different $e_{l,k}$ [31,34,35].

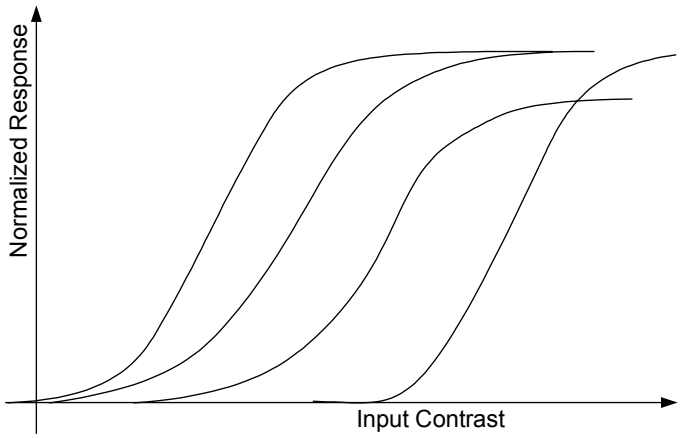


Figure 41.6 Non-linear contrast response saturation effects.

2.3 IMAGE QUALITY ASSESSMENT ALGORITHMS

Most of the efforts in the research community have been focussed on the problem of image quality assessment, and only recently has video quality assessment received more attention. Current video quality assessment metrics use HVS models similar to those used in many image quality assessment metrics, with appropriate extensions to incorporate the temporal aspects of the HVS. In this section, we present some image quality assessment metrics that are based on the HVS error sensitivity paradigm. Later we will present some video quality assessment metrics. A more detailed review of image quality assessment metrics may be found in [2,36].

The visible differences predictor (VDP) by Daly [28,29] aims to compute a probability-of-detection map between the reference and the distorted signal. The value at each point in the map is the probability that a human observer will perceive a difference between the reference and the distorted images at that point. The reference and the distorted images (expressed in luminance values instead of pixels) are passed through a series of processes: point non-linearlity, CSF filtering, channel decomposition, contrast calculation, masking effect modelling, and probability-of-detection calculation. A modified cortex transform [27] is used for channel decomposition, which transforms

the image signal into five spatial levels followed by six orientation levels, leading to a total of 31 independent channels (including the baseband). For each channel, a threshold elevation map is computed from the contrast in that channel. A psychometric function is used to convert error strengths (weighted by the threshold elevations) into a probability-of-detection map for each channel. Pooling is then carried out across the channels to obtain an overall detection map.

Lubin's algorithm [17,18] also attempts to estimate a detection probability of the differences between the original and the distorted versions. A blur is applied to model the PSF of the eye optics. The signals are then re-sampled to reflect the photoreceptor sampling in the retina. A Laplacian pyramid [37] is used to decompose the images into seven resolutions (each resolution is one-half of the immediately higher one), followed by band-limited contrast calculations [26]. A set of orientation filters implemented through steerable filters of Freeman and Adelson [38] is then applied for orientation selectivity in four orientations. The CSF is modelled by normalizing the output of each frequency-selective channel by the base-sensitivity for that channel. Masking is implemented through a sigmoid non-linearity, after which the errors are convolved with disk-shaped kernels at each level before being pooled into a distortion map using Minkowski pooling across frequency. An additional pooling stage may be applied to obtain a single number for the entire image.

Teo and Heeger's metric [1,25,30] uses PSF modelling, luminance masking, channel decomposition, and contrast normalization. The channel decomposition process uses quadrature steerable filters [39] with six orientation levels and four spatial resolutions. A detection mechanism is implemented based on squared error. Masking is modelled through contrast normalization and response saturation. The contrast normalization is different from Daly's or Lubin's method in that they take the outputs of channels at all orientations at a particular resolution to perform the normalization. Thus, this model does not assume that the channels at the same resolution are independent. Only channels at different resolutions are considered to be independent. The output of the channel decomposition after contrast normalization is decomposed four-fold by passing through four non-linearities of shapes as illustrated in Figure 41.6, the parameters for which are optimized to fit the data from psychovisual experiments.

Watson's DCT metric [40] is based on an 8×8 DCT transform commonly used in image and video compression. Unlike the models above, this method partitions the spectrum into 64 uniform subbands (8 in each Cartesian dimension). After the block-based DCT and the associated subband contrasts are computed, a visibility threshold is calculated for each subband coefficient within each block using the base-sensitivity for that subband. The base sensitivities are derived empirically. The thresholds are corrected for luminance and texture masking. The error in each subband is weighted by the corresponding visibility threshold and pooled using Minkowski pooling spatially. Pooling across subbands is then performed using the Minkowski formulation with a different exponent.

Safranek-Johnston's perceptual image coder [41] incorporates a quality metric using a similar strategy as in Watson's DCT metric. The channel decomposition uses a generalized quadrature mirror filter (GQMF) [42] for analysis and synthesis. This transform splits the spectrum into 16 uniform

subbands (four in each Cartesian dimension). Masking and pooling methods are similar to those in Watson's DCT metric.

Bradley [43] reports a wavelet visible difference predictor (WVDP), which is a simplification of Daly's VDP described above. He uses Watson's derivation of 9/7 Wavelet quantization-noise detection thresholds [44] for a 9/7 biorthogonal wavelet [45] and combines it with a threshold elevation and psychometric detection probability scheme similar to Daly's. Another wavelet based metric has been proposed by Lai and Kuo [6]. Their metric is based on the Haar Wavelet and their masking model can account for channel interactions as well as suprathreshold effects.

The quality metrics proposed above are scalar valued metrics. Damera-Venkata *et al.* proposed a metric for quantifying performance of image restoration systems, in which the degradation is modelled as a linear frequency distortion and additive noise injection [46]. Two complementary metrics were developed to separately quantify these distortions. They observed that if the additive noise is uncorrelated with the reference image, then an error measure from an HVS based metric will correlate well with the subjective judgement. Using a spatially adaptive restoration algorithm [47] (which was originally designed for inverse-halftoning), they isolate the effects of noise and linear frequency distortion. The noise is quantified using a multichannel HVS based metric. A distortion measure quantifies the spectral distortion between the reference and the model restored image.

Some researchers have attempted to measure image quality using single-channel models with the masking-effect models specifically targeting certain types of distortions, such as the blocking artifact. Blocking is recognized as one of the most annoying artifacts in block-DCT based image/video compression such as JPEG, especially at high compression ratios. In [48] and [49], Karunasekera and Kingsbury proposed a quality metric for blocking artifacts. Edge detection is performed first on the error image. An activity map is calculated from the reference image in the neighbourhood of the edges, and an activity-masked edge image is computed such that edges that occur in high activity areas are de-emphasized. The activity-masked edge image is adjusted for luminance masking. A non-linear transformation is applied before pooling. The parameters for the model are obtained from experiments that measure the sensitivity of human observers to edge artifacts embedded in narrow-band test patterns.

In [50], Chou and Li defined a peak signal to perceptible noise ratio (PSPNR), which is a single-channel metric. They model luminance masking and activity masking to obtain a JND profile. The PSPNR has the same definition as given in (41.2) except that the MSE expression is adjusted for the JND profile.

Another single-channel metric is the objective picture quality scale (PQS) by Miyahara [51], a number of features that can capture various distortions are combined into one score. The method has also been extended for video quality assessment [52].

2.4 VIDEO QUALITY ASSESSMENT ALGORITHMS

One obvious way to implement video quality metrics is to apply a still image quality assessment metric on a frame-by-frame basis. However, a more

sophisticated approach would model the temporal aspects of the HVS in the design of the metric. A number of algorithms have been proposed to extend the HVS features into the dimensions of time and motion [5,22,24,32,52-54]. A survey of video coding distortions can be found in [55]. A review of HVS modelling for video quality metrics is presented in [56].

In [53], Tan *et al.* implemented a Video Distortion Meter by using an image quality assessment metric followed by a "cognitive emulator" that models temporal effects such as smoothing and temporal masking of the frame quality measure, saturation and asymmetric tracking. Asymmetric tracking models the phenomenon that humans tend to notice a quality transition from good to poor more readily than a quality transition from poor to good.

Van den Branden Lambrecht *et al.* has extended the HVS modelling into the time dimension by modelling the temporal dimension of the CSF, and by generating two visual streams tuned to different temporal aspects of the stimulus from the output of each spatial channel [21,22,31,32]. The two streams model the transient and the sustained temporal mechanisms in the HVS. His proposed moving picture quality metric (MPQM) consists of a channel decomposition into four scales, four orientations and two temporal streams. The resulting channel outputs are subtracted to create the error signal. Masking is implemented by normalization of the channel errors by the stimulus dependent visibility thresholds (similar to those used in still image quality assessment metrics). Motion rendering quality assessment has also been proposed by extending the MPQM by extraction of motion information [32].

In [5], Winkler presented a quality assessment metric for color video. The algorithm uses a color space transformation and applies the quality assessment metric on each transformed color channel. Two temporal streams are generated using IIR filters, with spatial decomposition into five subband levels and four orientations. Channels are weighted by the corresponding CSF, and masking is implemented based on the excitatory-inhibitory masking model proposed by Watson and Solomon [33].

Watson's digital video quality (DVQ) metric operates in the DCT domain and is therefore more attractive from an implementation point of view [24,57] since the DCT is efficient to implement and most video coding standards are based on the DCT. A three-dimensional visibility threshold model for spatiotemporal DCT channels was proposed. The DVQ algorithm first takes the DCT of the reference and the distorted signals, respectively. It then computes local contrast, applies temporal CSF filtering, and converts the results into JND units by normalizing them with the visibility thresholds, following which the error signal is computed. Finally, masking and pooling are applied to the error signal. DVQ implements color transformation before applying the metric to each of the chrominance dimensions.

Another metric that models the temporal aspects of HVS is presented by Tan and Ghanbari [54], which aims to evaluate the quality of MPEG video and combines a typical error-sensitivity based perceptual model with a blockiness measurement model. The perceptual model consists of display gamma correction, point non-linearity, contrast calculation, spatial CSF filtering, temporal filtering, frequency decomposition into two channels (diagonal and horizontal/vertical), contrast response non-linearity, error averaging,

masking, pooling, temporal averaging and motion masking. The blockiness detector is based on harmonic analysis of the block-edge signal, combined with a visual masking model. The final quality score is either the perceptual model score or the blockiness detector score, based on the amount of blockiness artifact detected.

In [58], Yu *et al.* propose a video quality metric based on the extension of the perceptual distortion metric by Winkler [5] to a perceptual blocking distortion metric. The parameters for the models are obtained by minimizing the error in quality predictions for video sequences obtained from VQEG subjective testing database. This is in contrast to most methods that obtain parameters to fit threshold psychovisual experiments with simple patterns. They specifically address the blocking artifact by pooling spatially over those areas where blocking effects are dominant.

There are several implementation issues that need to be considered before developing a practical video quality assessment system. One important factor affecting the feasibility of a quality metric for video is its computational complexity. While complex quality assessment methods may model the HVS more accurately, their computational complexity may be prohibitively large for many platforms, especially for real-time quality assessment of high-resolution video. Memory requirements are another important issue. For example, in order to implement temporal filtering, a large memory space may be needed to store a number of video frames, which is expensive on many platforms. Another problem of HVS based metrics might be their dependence on viewing configurations, which include the resolution of the display devices, the non-linear relationships between the digital pixel values and the output luminance values, and the viewing distance of the observers. Most models either require that viewing configurations be known or simply assume a fixed set of configurations. How these metrics would perform when the configurations are unknown or the assumptions about the configurations do not hold is another issue that needs to be studied.

2.5 LIMITATIONS

The underlying principle of visual error sensitivity based algorithms is to predict perceptual quality by quantifying perceptible errors. This is accomplished by simulating the perceptual quality related functional components of the HVS. However, the HVS is an extremely complicated, highly non-linear system, and the current understanding of the HVS is limited. How far the error sensitivity based framework can reach is a question that may need many years to answer.

It is worth noting that most error sensitivity based approaches, explicitly or implicitly, make a number of assumptions. The following is an incomplete list (Note: a specific model may use a subset of these assumptions):

1. The reference signal is of perfect quality.
2. Light adaptation follows the Weber's law.
3. After light adaptation, the optics of the eye can be modelled as a linear time-invariant system characterized by a PSF.

4. There exist frequency, orientation and temporal selectivity visual channels in the HVS, and the channel responses can be modelled by a discrete set of linear decompositions.
5. Although the contrast definitions of simple patterns used in psychovisual experiments and the contrast definitions of complex natural images may be different, they are consistent with each other.
6. The relative visual error sensitivity between different spatial and/or temporal frequency channels can be normalized using a bandpass or lowpass CSF.
7. The channel decomposition is lossless or nearly lossless in terms of visual importance, in the sense that the transformed signals maintain most of the information needed to assess the image quality.
8. The channel decomposition effectively decorrelates the image structure, such that the inter- and intra-channel interactions between transformed coefficients can be modelled using a masking model, in which the strength of the mask is determined by the magnitudes (not structures) of the coefficients. After masking, the perceived error of each coefficient can be evaluated individually.
9. For a single coefficient in each channel, after error normalization and masking, the relationship between the magnitude of the error, $e_{l,k}$, and the distortion perceived by the HVS, $d_{l,k}$, can be modelled as a non-linear function: $d_{l,k} = \left| e_{l,k} \right|^{\beta}$.
10. The overall perceived distortion monotonically increases with the summation of the perceived errors of all coefficients in all channels.
11. The overall framework covers a complete set of dominant factors (light adaptation, PSF of the eye optics, CSF of the frequency responses, masking effects, etc.) that affect the perceptual quality of the observed image.
12. Higher level processes happening in the human brain, such as pattern matching with memory and cognitive understanding, are less important for predicting perceptual image quality.
13. Active visual processes, such as the change of fixation points and the adaptive adjustment of spatial resolution because of attention, are less important for predicting perceptual image quality.

Depending on the application environment, some of the above assumptions are valid or practically reasonable. For example, in image and video compression and communication applications, assuming a perfect original image or video signal (Assumption 1) is acceptable. However, from a more general point of view, many of the assumptions are arguable and need to be validated. We believe that there are several problems that are critical for justifying the usefulness of the general error-sensitivity based framework.

The Suprathreshold Problem

Most psychophysical subjective experiments are conducted near the threshold of visibility, typically using a 2-Alternative Forced-Choice (2AFC)

method [14,59]. The 2AFC method is used to determine the values of stimuli strength (also called the *threshold strength*) at which the stimuli are *just visible*. These measured threshold values are then used to define visual error sensitivity models, such as the CSF and the various masking effect models. However, there is not sufficient evidence available from vision research to support the presumption that these measurement results can be generalized to quantify distortions much larger than *just visible*, which is the case for a majority of image processing applications. This may lead to several problems with respect to the framework. One problem is that when the error in a visual channel is larger than the threshold of visibility, it is hard to design experiments to validate Assumption 9. Another problem is regarding Assumption 6, which uses the just noticeable visual error threshold to normalize the errors between different frequency channels. The question is: when the errors are much larger than the thresholds, can the relative errors between different channels be normalized using the visibility thresholds?

The Natural Image Complexity Problem

Most psychovisual experimental results published in the literature are conducted using relatively simple patterns, such as sinusoidal gratings, Gabor patches, simple geometrical shapes, transform basis functions, or random noise patterns. The CSF is obtained from threshold experiments using single frequency patterns. The masking experiments usually involve two (or a few) different patterns. However, all such patterns are much simpler than real world images, which can usually be thought of as a superposition of a large number of different simple patterns. Are these simple-pattern experiments sufficient for us to build a model that can predict the quality of complex natural images? Can we generalize the model for the interactions between a few simple patterns to model the interactions between tens or hundreds of patterns?

The Minkowski Error Pooling Problem

The widely used Minkowski error summation formula (41.3) is based on signal differencing between two signals, which may not capture the structural changes between the two signals. An example is given in Figure 41.7, where two test signals, test signals 1 (up-left) and 2 (up-right), are generated from the original signal (up-center). Test signal 1 is obtained by adding a constant number to each sample point, while the signs of the constant number added to test signal 2 are randomly chosen to be positive or negative. The structural information of the original signal is almost completely lost in test signal 2, but preserved very well in test signal 1. In order to calculate the Minkowski error metric, we first subtract the original signal from the test signals, leading to the error signals 1 and 2, which have very different structures. However, applying the absolute operator on the error signals results in exactly the same absolute error signals. The final Minkowski error measures of the two test signals are equal, no matter how the β value is selected. This example not only demonstrates that "structure-preservation" ability is an important factor in measuring the similarity between signals, but also shows that Minkowski error pooling is inefficient in capturing the structures of errors and is a "structural information lossy" metric. Obviously, in this specific example, the problem may be solved by applying a spatial frequency channel decomposition on the error signals and weighting the errors differently in different channels with a CSF. However, the decomposed signals may still

exhibits different structures in different channels (for example, assume that the test signals in Figure 41.7 are from certain channels instead of the spatial domain), then the "structural information lossy" weakness of the Minkowski metric may still play a role, unless the decomposition process strongly decorrelates the image structure, as described by Assumption 8, such that the correlation between adjacent samples of the decomposed signal is very small (in that case, the decomposed signal in a channel would look like random noise). However, this is apparently not the case for a linear channel decomposition method such as the wavelet transform. It has been shown that a strong correlation or dependency exists between intra- and inter-channel wavelet coefficients of natural images [60,61]. In fact, without exploiting this strong dependency, state-of-the-art wavelet image compression techniques, such as embedded zerotree wavelet (EZW) coding [62], set partitioning in hierarchical trees (SPIHT) algorithm [63], and JPEG2000 [64] would not be successful.

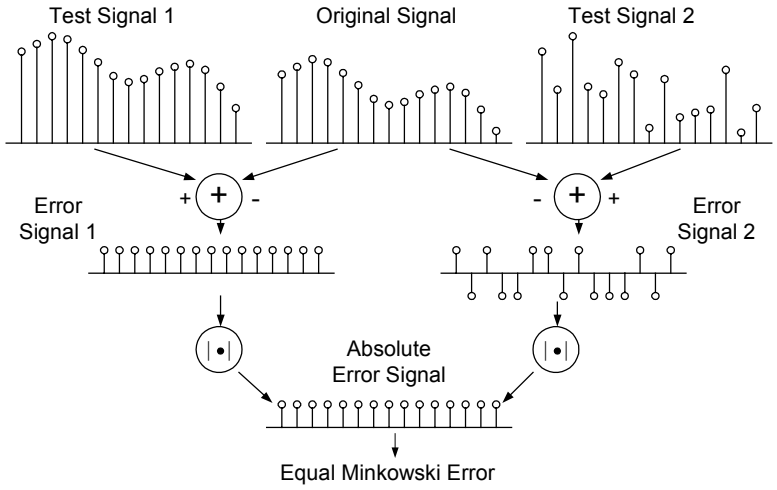


Figure 41.7 Illustration of Minkowski error pooling.

The Cognitive Interaction Problem

It is clear that cognitive understanding and active visual process (e.g., change of fixations) play roles in evaluating the quality of images. For example, a human observer will give different quality scores to the same image if s/he is instructed with different visual tasks [2,65]. Prior information regarding the image content, or attention and fixation, may also affect the evaluation of the image quality [2,66]. For example, it is shown in [67] that in a video conferencing environment, "the difference between sensitivity to foreground and background degradation is increased by the presence of audio corresponding to speech of the foreground person" [67]. Currently, most image and video quality metrics do not consider these effects. How these effects change the perceived image quality, how strong these effects compare with other HVS features employed in the current quality assessment models, and how to incorporate these effects into a quality assessment model have not yet been deeply investigated.

3. FULL-REFERENCE QUALITY ASSESSMENT USING STRUCTURAL DISTORTION MEASURES

The paradigm of error sensitivity based image and video quality assessment considers any kind of image distortions as being certain types of *errors*. Since different error structures will have different effects on perceived image quality, the effectiveness of this approach depends on how the structures of the errors are understood and represented. Linear channel decomposition is the most commonly used way to decompose the error signals into a set of elementary components, and the visual error sensitivity models for these elementary components are relatively easily obtained from psychovisual experiments. As described in Section 2.5, because linear channel decomposition methods cannot fully decorrelate the structures of the signal, the decomposed coefficients still exhibit strong correlations with each other. It has been argued in Section 2.5 that the Minkowski error metric cannot capture these structural correlations. Therefore, the error sensitivity based paradigm relies on a very powerful masking model, which must cover various kinds of intra- and inter-channel interactions between the decomposed coefficients. Current knowledge about visual masking effects is still limited. At this moment, it is not clear whether building a comprehensive masking model is possible or not, but it is likely that even if it were possible, the model would be very complicated.

In this section, we propose an alternative way to think about image quality assessment: it is not necessary to consider the difference between an original image and a distorted image as a certain type of error. What we will now describe as structural distortion measurement may lead to more efficient and more effective image quality assessment methods.

3.1 NEW PHILOSOPHY

In [8] and [68], a new philosophy in designing image and video quality metrics has been proposed:

The main function of the human visual system is to extract structural information from the viewing field, and the human visual system is highly adapted for this purpose. Therefore, a measurement of structural distortion should be a good approximation of perceived image distortion.

The new philosophy can be better understood by comparison with the error sensitivity based philosophy:

First, a major difference of the new philosophy from the error sensitivity based philosophy is the switch from *error measurement* to *structural distortion measurement*. Although error and structural distortion sometimes agree with each other, in many circumstances the same amount of error may lead to significantly different structural distortion. A good example is given in Figures 41.8 and 41.9, where the original "Lena" image is altered with a wide variety of distortions: impulsive salt-pepper noise, additive Gaussian noise, multiplicative speckle noise, mean shift, contrast stretching, blurring, and heavy JPEG compression. We tuned all the distorted images to yield the same MSE relative to the original one, except for the JPEG compressed image, which has a slightly smaller MSE. It is interesting to see that images with nearly identical MSE have drastically different perceptual quality. Our

subjective evaluation results show that the contrast stretched and the mean shifted images provide very high perceptual quality, while the blurred and the JPEG compressed images have the lowest subjective scores [7,68]. This is no surprise with a good understanding of the new philosophy since the structural change from the original to the contrast stretched and mean shifted images is trivial, but to the blurred and JPEG compressed images the structural modification is very significant.

Figure 41.8 Evaluation of "Lena" images with different types of noise. Top-left: Original "Lena" image, 512×512, 8bits/pixel; Top-right: Impulsive salt-pepper noise contaminated image, MSE=225, Q=0.6494; Bottom-left: Additive Gaussian noise contaminated image, MSE=225, Q=0.3891; Bottom-right: Multiplicative speckle noise contaminated image, MSE=225, Q=0.4408.

Second, another important difference of the new philosophy is that it considers image degradation as *perceived structural information loss*. For example, in Figure 41.9, the contrast stretched image has a better quality than the JPEG compressed image simply because almost all the structural information of the original image is preserved, in the sense that the original image can be recovered via a simple pointwise inverse linear luminance transform. Apparently, a lot of information in the original image is permanently lost in the JPEG compressed image. The reason that a structural information loss measurement can be considered as a prediction

of visual perception is based on the assumption that the HVS functions similarly — it has adapted to extract structural information and to detect changes in structural information. By contrast, an error sensitivity based approach estimates *perceived errors* to represent image degradation. If it works properly, then a significant perceptual error should be reported for the contrast stretched image because its difference (in terms of error) from the original image is easily discerned.

Figure 41.9 Evaluation of "Lena" images with different types of distortions. Top-left: Mean shifted image, MSE=225, Q=0.9894; Top-right: Contrast stretched image, MSE=225, Q=0.9372; Bottom-left: Blurred image, MSE=225, Q=0.3461; Bottom-right: JPEG compressed image, MSE=215, Q=0.2876.

Third, the new philosophy uses a *top-down* approach, which starts from the very top level — simulating the hypothesized functionality of the overall HVS. By comparison, the error sensitivity based philosophy uses a *bottom-up* approach, which attempts to simulate the function of each relevant component in the HVS and combine them together, in the hope that the combined system will perform similarly to the overall HVS.

How to apply the new philosophy to create a concrete image and video quality assessment method is an open issue. There may be very different implementations, depending on how the concepts of "structural information" and "structural distortion" are interpreted and quantified. Generally

speaking, there may be two ways of implementing a quality assessment algorithm using the new philosophy. The first is to develop a feature description framework of natural images, which covers most of the useful structural information of an image signal. Under such a description framework, structural information changes between the original and the distorted signals can be quantified. The second is to design a structure comparison method that can compare structural similarity or structural difference between the original and the distorted signals directly. As a first attempt to implement this new philosophy, a simple image quality indexing approach was proposed in [7,68], which conforms to the second approach.

3.2 AN IMAGE QUALITY INDEXING APPROACH

Let $\boldsymbol{x} = \{x_i \mid i = 1,2,\cdots N\}$ and $\boldsymbol{y} = \{x_i \mid i = 1,2,\cdots N\}$ be the original and the test image signals, respectively. The proposed quality index is defined as:

$$Q = \frac{4\sigma_{xy}\,\bar{x}\,\bar{y}}{(\sigma_x^2 + \sigma_y^2)[(\bar{x})^2 + (\bar{y})^2]}, \tag{41.4}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$,

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2,\ \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2,$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}).$$

The dynamic range of Q is [−1, 1]. The best value 1 is achieved if and only if $y_i = x_i$ for all $i = 1,2,\cdots N$. The lowest value of −1 occurs when $y_i = 2\bar{x} - x_i$, for all $i = 1,2,\cdots N$.

This quality index models any distortion as a combination of three factors: loss of correlation, mean distortion, and contrast distortion. In order to understand this, we rewrite the definition of Q as the product of three components:

$$Q = \frac{\sigma_{xy}}{\sigma_x\,\sigma_y} \cdot \frac{2\bar{x}\,\bar{y}}{(\bar{x})^2 + (\bar{y})^2} \cdot \frac{2\sigma_x\,\sigma_y}{\sigma_x^2 + \sigma_y^2} \tag{41.5}$$

The first component is the correlation coefficient between $\boldsymbol{x}$ and $\boldsymbol{y}$, which measures the degree of linear correlation between $\boldsymbol{x}$ and $\boldsymbol{y}$, and its dynamic range is [−1, 1]. The best value 1 is obtained when $y_i = ax_i + b$ for all $i = 1,2,\cdots N$, where a and b are constants and $a > 0$. We consider the linear correlation coefficient as a very important factor in comparing the structures of two signals. Notice that a pointwise linearly changed signal can be recovered exactly with a simple pointwise inverse linear transform. In this sense, the "structural information" is preserved. Furthermore, a decrease in the linear correlation coefficient gives a quantitative measure of how much

the signal is changed nonlinearly. Obviously, even if $\boldsymbol{x}$ and $\boldsymbol{y}$ are linearly correlated, there still may be relative distortions between them, which are evaluated in the second and third components. The second component, with a range of [0, 1], measures how similar the mean values of $\boldsymbol{x}$ and $\boldsymbol{y}$ are. It equals 1 if and only if $\bar{x} = \bar{y}$. σ_x and σ_y can be viewed as rough estimate of the contrast of $\boldsymbol{x}$ and $\boldsymbol{y}$, so the third component measures how similar the contrasts of the images are. Its range of values is also [0, 1], where the best value 1 is achieved if and only if $\sigma_x = \sigma_y$.

Image signals are generally non-stationary and image quality is often spatially variant. In practice it is usually desired to evaluate an entire image using a single overall quality value. Therefore, it is reasonable to measure statistical features locally and then combine them together. We apply our quality measurement method to local regions using a sliding window approach. Starting from the top-left corner of the image, a sliding window of size B×B moves pixel by pixel horizontally and vertically through all the rows and columns of the image until the bottom-right corner is reached. At the *j*-th step, the local quality index Q_j is computed within the sliding window. If there are a total of *M* steps, then the overall quality index is given by

$$Q = \frac{1}{M}\sum_{j=1}^{M} Q_j \tag{41.6}$$

It has been shown that many image quality assessment algorithms work consistently well if the distorted images being compared are created from the same original image and the same type of distortions (e.g., JPEG compression). In fact, for such comparisons, the MSE or PSNR is usually sufficient to produce useful quality evaluations. However, the effectiveness of image quality assessment models degrades significantly when the models are employed to compare the quality of distorted images originating from different types of original images with different types of distortions. Therefore, cross-image and cross-distortion tests are very useful in evaluating the effectiveness of an image quality metric.

The images in Figures 41.8 and 41.9 are good examples for testing the cross-distortion capability of the quality assessment algorithm. Obviously, the MSE performs very poorly in this case. The quality indices of the images are calculated and given in Figures 41.8 and 41.9, where the sliding window size is fixed at B=8. The results exhibit surprising consistency with the subjective measures. In fact, the ranks given by the quality index are the same as the mean subjective ranks of our subjective evaluations [7,68]. We noticed that many subjects regard the contrast stretched image to have better quality than the mean shifted image and even the original image. This is no surprise because contrast stretching is often an image enhancement process, which often increases the visual quality of the original image. However, if we assume that the original image is the perfect one (as our quality measurement method does), then it is fair to give the mean shifted image a higher quality score.

Figure 41.10 Evaluation of blurred image quality. Top-left: Original "Woman" image; Top-right: Blurred "Woman" image, MSE=200, Q=0.3483; Middle-left: Original "Man" image; Middle-right: Blurred "Man" image, MSE=200, Q=0.4123; Bottom-left: Original "Barbara" image; Bottom-right: Blurred "Barbara" image, MSE=200, Q=0.6594.

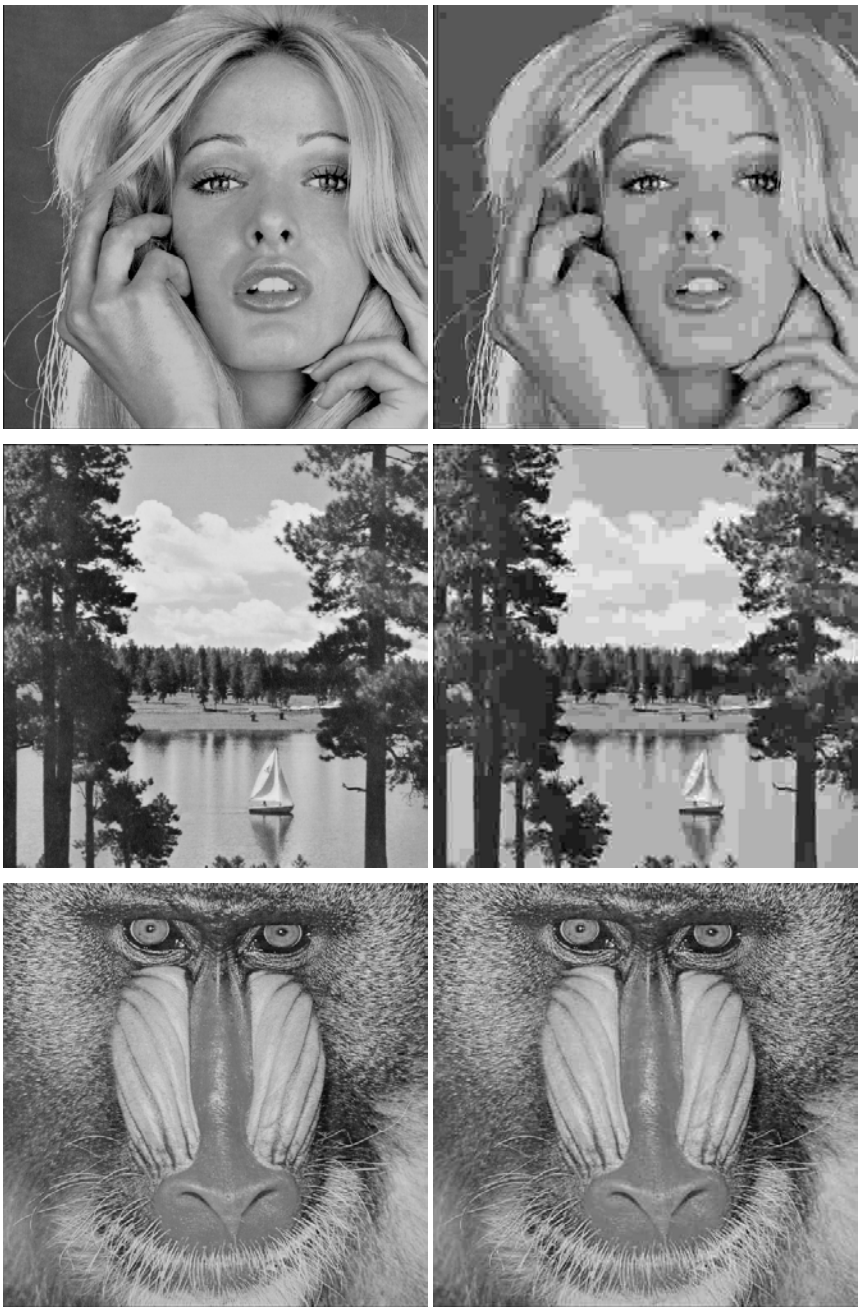


Figure 41.11 Evaluation of JPEG compressed image quality. Top-left: Original "Tiffany" image; Top-right: compressed "Tiffany" image, MSE=165, Q=0.3709; Middle-left: Original "Lake" image; Middle-right: compressed "Lake" image, MSE=167, Q=0.4606; Bottom-left: Original "Mandrill" image; Bottom-right: compressed "Mandrill" image, MSE=163, Q=0.7959.

In Figures 41.10 and 41.11, different images with the same distortion types are employed to test the cross-image capability of the quality index. In Figure 41.10, three different images are blurred, such that they have almost the same MSE with respect to their original ones. In Figure 41.11, three other images are compressed using JPEG, and the JPEG compression quantization steps are selected so that the three compressed images have similar MSE in comparison with their original images. Again, the MSE has very poor correlation with perceived image quality in these tests, and the proposed quality indexing algorithm delivers much better consistency with visual evaluations.

Interested users may refer to [69] for more demonstrative images and an efficient MATLAB implementation of the proposed quality indexing algorithm.

The proposed quality indexing method is only a rudimentary implementation of the new paradigm. Although it gives promising results under the current limited testings, more extended experiments are needed to validate and optimize the algorithm. More theoretical and experimental connections with respect to human visual perception need to be established. Another important issue that needs to be explored is how to apply it for video quality assessment. In [70], the quality index was calculated frame by frame for a video sequence and combined with other image distortion features such as blocking to produce a video quality measure.

4. NO-REFERENCE AND REDUCED-REFERENCE QUALITY ASSESSMENT

The quality metrics presented so far assume the availability of the reference video to compare the distorted video against. This requirement is a serious impediment to the feasibility of video quality assessment metrics. Reference videos require tremendous amounts of storage space, and in many cases, are impossible to provide for most applications.

Reduced-reference (RR) quality assessment does not assume the complete availability of the reference signal, only that of partial reference information that is available through an ancillary data channel. Figure 41.12 shows how a RR quality assessment metric may be deployed. The server transmits side-information with the video to serve as a reference for an RR quality assessment metric down the network. The bandwidth available to the RR channel depends upon the application constraints. The design of RR quality assessment metrics need to look into what information is to be transmitted through the RR channel so as to provide minimum prediction errors. Needless to say, the feature extraction from the reference video at the server would need to correspond to the intended RR quality assessment algorithm.

Perhaps the earliest RR quality assessment metric was proposed by Webster *et al.* [71] and is based on extracting localized spatial and temporal activity features. A spatial information (SI) feature measures the standard deviation of edge-enhanced frames, assuming that compression will modify the edge statistics in the frames. A temporal information (TI) feature is also extracted, which is the standard deviation of difference frames. Three comparison metrics are derived from the SI and TI features of the reference and the distorted videos. The features for the reference video are transmitted over the

RR channel. The metrics are trained on data obtained from human subjects. The size of the RR data depends upon the size of the window over which SI & TI features are calculated. The work was extended in [72], where different edge enhancement filters are used, and two activity features are extracted from 3D windows. One feature measures the strength of the edges in the horizontal/vertical directions, while the second feature measures the strength of the edges over other directions. Impairment metric is defined using these features. Extensive subjective testing is also reported.

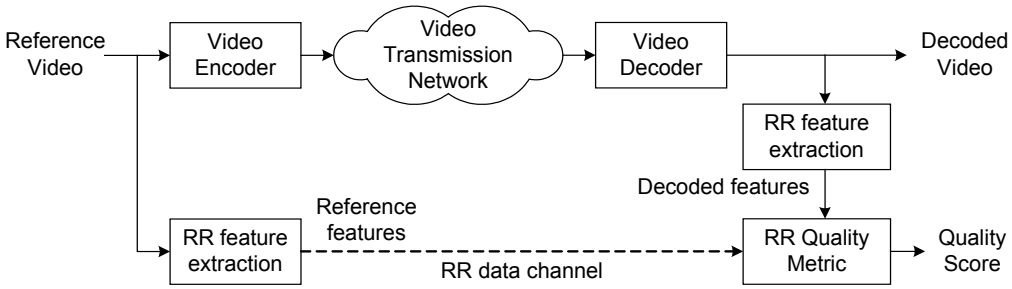


Figure 41.12 Deployment of a reduced-reference video quality assessment metric. Features extracted from the reference video are sent to the receiver to aid in quality measurements. The video transmission network may be lossy but the RR channel is assumed to be lossless.

Another approach that uses side-information for quality assessment is described in [73], in which marker bits composed of random bit sequences are hidden inside frames. The markers are also transmitted over the ancillary data channel. The error rate in the detection of the marker bits is taken as an indicator of the loss of quality. In [74], a watermarking based approach is proposed, where a watermark image is embedded in the video, and it is suggested that the degradation in the quality of the watermark can be used to predict the degradation in the quality of the video. Strictly speaking, both these methods are not RR quality metrics since they do not extract any features from the reference video. Instead, these methods gauge the distortion processes that occur during compression and the communication channel to estimate the perceptual degradation incurred during transmission in the channel.

Given the limited success that FR quality assessment has achieved, it should come as no surprise that designing objective no-reference (NR) quality measurement algorithms is very difficult indeed. This is mainly due to the limited understanding of the HVS and the corresponding cognitive aspects of the brain. Only a few methods have been proposed in the literature [75-80] for objective NR quality assessment, yet this topic has attracted a great deal of attention recently. For example, the video quality experts group (VQEG) [81] considers the standardization of NR and RR video quality assessment methods as one of its future working directions, where the major source of distortion under consideration is block DCT-based video compression.

The problem of NR quality assessment (sometimes called *blind* quality assessment) is made even more complex due to the fact that many unquantifiable factors play a role in human assessment of quality, such as aesthetics, cognitive relevance, learning, visual context, etc., when the reference signal is not available for MOS evaluation. These factors introduce variability among human observers based on each individual's subjective impressions. However, we can work with the following philosophy for NR quality assessment: *all images and videos are perfect unless distorted during acquisition, processing or reproduction*. Hence, the task of blind quality measurement simplifies into blindly measuring the distortion that has possibly been introduced during the stages of acquisition, processing or reproduction. The reference for measuring this distortion would be the statistics of "perfect" natural images and videos, measured with respect to a model that best suits a given distortion type or application. This philosophy effectively decouples the unquantifiable aspects of image quality mentioned above from the task of objective quality assessment. All "perfect images" are treated equally, disregarding the amount of cognitive information in the image, or its aesthetic value [82,83].

The NR metrics cited above implicitly adhere to this philosophy of quantifying quality through blind distortion measurement. Assumptions regarding statistics of "perfect natural images" are made such that the distortion is well separated from the "expected" signals. For example, natural images do not contain blocking artifacts, and any presence of periodic edge discontinuity in the horizontal and vertical directions with a period of 8 pixels, is probably a distortion introduced by block-DCT based compression techniques. Some aspects of the HVS, such as texture and luminance masking, are also modelled to improve prediction. Thus NR quality assessment metrics need to model not only the HVS but also natural scene statistics.

Certain types of distortions are quite amenable to blind measurement, such as blocking artifacts. In wavelet-based image coders, such as the JPEG2000 standard, the wavelet transform is often applied to the entire image (instead of image blocks), and the decoded images do not suffer from blocking artifact. Therefore, NR metrics based on blocking artifacts would obviously fail to give meaningful predictions. The upcoming H.26L standard incorporates a powerful de-blocking filter. Similarly post-processing may reduce blocking artifacts at the cost of introducing blur. In [82], a statistical model for natural images in the wavelet domain is used for NR quality assessment of JPEG2000 images. Any NR metric would therefore need to be specifically designed for the target distortion system. More sophisticated models of natural images may improve the performance of NR metrics and make them more robust to various distortion types.

5. VALIDATION OF QUALITY ASSESSMENT METRICS

Validation is an important step towards successful development of practical image and video quality measurement systems. Since the goal of these systems is to predict perceived image and video quality, it is essential to build an image and video database with subjective evaluation scores associated with each of the images and video sequences in the database. Such a

database can then be used to assess the prediction performance of the objective quality measurement algorithms.

In this section, we first briefly introduce two techniques that have been widely adopted in both the industry and the research community for subjective evaluations for video. We then review the quality metric comparison results published in the literature. Finally, we introduce the recent effort by the video quality experts group (VQEG) [81], which aims to provide industrial standards for video quality assessment.

5.1 SUBJECTIVE EVALUATION OF VIDEO QUALITY

Subjective evaluation experiments are complicated by many aspects of human psychology and viewing conditions, such as observer vision ability, translation of quality perception into ranking score, preference for content, adaptation, display devices, ambient light levels etc. The two methods that we will present briefly are single stimulus continuous quality evaluation (SSCQE) and double stimulus continuous quality scale (DSCQS), which have been demonstrated to have repeatable and stable results, provided consistent viewing configurations and subjective tasks, and have consequently been adopted as parts of an international standard by the international telecommunications union (ITU) [84]. If the SSCQE and DSCQS tests are conducted on multiple subjects, the scores can be averaged to yield the mean opinion score (MOS). The standard deviation between the scores may also be useful to measure the consistency between subjects.

Single Stimulus Continuous Quality Evaluation

In the SSCQE method, subjects continuously indicate their impression of the video quality on a linear scale that is divided into five segments, as shown in Figure 41.13. The five intervals are marked with adjectives to serve as guides. The subjects are instructed to move a slider to any point on the scale that best reflects their impression of quality at that instant of time, and to track the changes in the quality of the video using the slider.

Figure 41.13 SSCQE sample quality scale.

Double Stimulus Continuous Quality Scale

The DSCQS method is a form of discrimination based method and has the extra advantage that the subjective scores are less affected by adaptation and contextual effects. In the DSCQS method, the reference and the distorted videos are presented one after the other in the same session, in small

segments of a few seconds each, and subjects evaluate both sequences using sliders similar to those for SSCQE. The difference between the scores of the reference and the distorted sequences gives the subjective impairment judgement. Figure 41.14 demonstrates the basic test procedure.

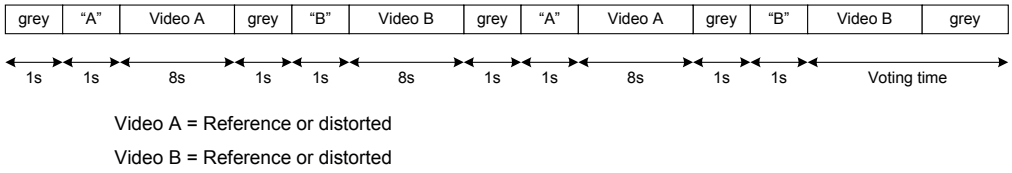


Figure 41.14 DSCQS testing procedure recommended by VQEG FR-TV Phase-II test.

5.2 COMPARISON OF QUALITY ASSESSMENT METRICS

With so many quality assessment algorithms proposed, the question of their relative merits and demerits naturally arises. Unfortunately, not much has been published in comparing these models with one another, especially under strict experimental conditions over a wide range of distortion types, distortion strengths, stimulus content and subjective evaluation criterion. This is compounded by the fact that validating quality assessment metrics comprehensively is time-consuming and expensive, not to mention that many algorithms are not described explicitly enough in the literature to allow reproduction of their reported performance. Most comparisons of quality assessment metrics are not broad enough to be able to draw solid conclusions, and their results should only be considered in the context of their evaluation criterion.

In [3] and [65], different mathematical measures of quality that operate without channel decompositions and masking effect modelling are compared against subjective experiments, and their performance is tabulated for various test conditions. Li *et al.* compare Daly's and Lubin's models for their ability to detect differences [85] and conclude that Lubin's model is more robust than Daly's given their experimental procedures. In [86] three metrics are compared for JPEG compressed images: Watson's DCT based metric [87], Chou and Li's method [50] and Karanusekera and Kingsbury's method [49]. They conclude that Watson's method performed best among the three.

Martens and Meesters have compared Lubin's metric (also called the Sarnoff model) with the root mean squared error (RMSE) [9] metric on transformed luminance images. The metrics are compared using subjective experiments based on images corrupted with noise and blur, as well as images corrupted with JPEG distortion. The subjective experiments are based on *dissimilarity measurements*, where subjects are asked to assess the dissimilarity between pairs of images from a set that contains the reference image and several of its distorted versions. Multidimensional scaling (MDS) technique is used to compare the metrics with the subjective experiments. MDS technique constructs alternate spaces from the dissimilarity data, in which the positions of the images are related to their dissimilarity (subjective or objective) with the rest of the images in that set. Martens and Meesters then compare RMSE and Lubin's method with subjective experiments, with and

without MDS, and report that "in none of the examined cases could a clear advantage of complicated distance metrics (such as the Sarnoff model) be demonstrated over simple measures such as RMSE" [9].

5.3 VIDEO QUALITY EXPERTS GROUP

The video quality experts group [81] was formed in 1997 to develop, validate and standardize new objective measurement methods for video quality. The group is composed of experts from various backgrounds and organizations around the world. They are interested in FR/RR/NR quality assessment for various bandwidth videos for television and multimedia applications.

VQEG has completed its Phase I test for FR video quality assessment for television in 2000 [10,11]. In Phase I test, 10 proponent video quality models (including several well-known models and PSNR) were compared with the subjective evaluation results on a video database, which contains video sequences with a wide variety of distortion types and stimulus content. A systematic way of evaluating the prediction performance of the objective models was established, which is composed of three components:

1. Prediction accuracy — the ability to predict the subjective quality ratings with low error. (Two metrics, namely the variance-weighted regression correlation [10] and the non-linear regression correlation [10], were used.)
2. Prediction monotonicity — the degree to which the model's predictions agree with the relative magnitudes of subjective quality ratings. (The Spearman rank order correlation [10] was employed.)
3. Prediction consistency — the degree to which the model maintains prediction accuracy over the range of video test sequences. (The outlier ratio [10] was used.)

The result was, in some sense, surprising, since except for 1 or 2 proponents that did not perform properly in the test, the other proponents performed statistically equivalent, including PSNR [10,11]. Consequently, VQEG did not recommend any method for an ITU standard [10]. VQEG is continuing its work on Phase II test for FR quality assessment for television, and RR/NR quality assessment for television and multimedia.

Although it is hard to predict whether VQEG will be able to supply one or a few successful video quality assessment standards in the near future, the work of VQEG is important and unique from a research point of view. First, VQEG establishes large video databases with reliable subjective evaluation scores (the database used in the FR Phase I test is already available to the public [81]), which will prove to be invaluable for future research on video quality assessment. Second, systematic approaches for comparing subjective and objective scores are being formalized. These approaches alone could become widely accepted standards in the research community. Third, by comparing state-of-the-art quality assessment models in different aspects, deeper understanding of the relative merits of different methods will be achieved, which will have a major impact on future improvement of the models. In addition, VQEG provides an ideal communication platform for the researchers who are working in the field.

6. CONCLUSIONS AND FUTURE DIRECTIONS

There has been increasing interest in the development of objective quality measurement techniques that can automatically predict the perceptual quality of images and video streams. Such methods are useful tools for video database systems and are also desired for a broad variety of applications, such as video acquisition, compression, communication, displaying, analysis, watermarking, restoration and enhancement. In this chapter, we reviewed the basic concepts and methods in objective image and video quality assessment research and discussed various quality prediction approaches. We also laid out a new paradigm for quality assessment based on structural distortion measures, which has some promising advantages over traditional perceptual error sensitivity based approaches.

After decades of research work, image and video quality assessment is still an active and evolving research topic. An important goal of this chapter is to discuss the challenges and difficulties encountered in developing objective quality assessment approaches, and provide a vision for future directions.

For error sensitivity based approaches, the four major problems discussed in Section 2.5 also laid out the possible directions that may be explored to provide improvements of the current methods. One of the most important aspects that requires the greatest effort is to investigate various masking/facilitation effects, especially the masking and facilitation phenomena in the suprathreshold range and in cases where the background is composed of broadband natural image (instead of simple patterns). For example, the contrast matching experimental study on center-surround interactions [88] may provide a better way to quantitatively measure the image distortions at the suprathreshold [89] level. The Modelfest phase one dataset [90] collected simple patterns such as Gabor patches as well as some more complicated broadband patterns including one natural image. Comparing and analysing the visual error prediction capability of different error sensitivity based methods with these patterns may help researchers to better understand whether HVS features measured with simple patterns can be extended to predict the perceptual quality of complex natural images.

The structural distortion based framework is at a very preliminary stage. The newly proposed image quality indexing approach is attractive not only because of its promising results, but also its simplicity. However, it is perhaps too simple and the combination of the three factors is *ad-hoc*. More theoretical analysis and subjective experimental work is needed to provide direct evidence on how it is connected with visual perception and natural image statistics. Many other issues may also be considered, such as multiscale analysis, adaptive windowing and space-variant pooling using a statistical fixation model. Furthermore, under the new paradigm of structural distortion measurement, other approaches may emerge that could be very different from the proposed quality indexing algorithm. The understanding of "structural information" would play a key role in these innovations.

Another interesting point is the possibility of combining the advantages of the two paradigms. This is a difficult task without a deeper understanding of both. One possible connection may be as follows: use the structural distortion based method to measure the amount of structural information

loss, and the error sensitivity based approach to help determine whether such information loss can be perceived by the HVS.

The fields of NR and RR quality assessment are very young, and there are many possibilities for the development of innovative metrics. The philosophy of doing NR or RR quality assessment will continue to be that of blind distortion measurement with respect to features that best separate the undistorted signals from the distortion. The success of statistical models for natural scenes that are more suited to certain distortion types and applications will drive the success of NR and RR metrics in the future. A combination of natural scene models with HVS models may also prove beneficial for NR and RR quality assessment.

REFERENCES

[1] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Processing*, pp. 982-986, 1994.

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

[3] A. M. Eskicioglu and P. S. Fisher, “Image quality measures and their performance,” *IEEE Trans. Communications*, vol. 43, pp. 2959-2965, Dec. 1995.

[4] B. Girod, “What's wrong with mean-squared error,” in *Digital Images and Human Vision*, A. B. Watson, ed., pp. 207-220, MIT Press, 1993.

[5] S. Winkler, "A perceputal distortion metric for digital color video," *Proc. SPIE*, vol. 3644, pp. 175-184, 1999.

[6] Y. K. Lai and C.-C. J. Kuo, "A Haar wavelet approach to compressed image quality measurement," *Journal of Visual Communication and Image Understanding,* vol. 11, pp. 17-40, Mar. 2000.

[7] Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81-84, March 2002.

[8] Z. Wang, A. C. Bovik and L. Lu, “Why is image quality assessment so difficult?” *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Proc.*, vol. 4, pp. 3313-3316, May 2002.

[9] J.-B. Martens and L. Meesters, “Image dissimilarity,” *Signal Processing*, vol. 70, pp. 1164-1175, Aug. 1997.

[10] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," *http://www.vqeg.org/*, Mar. 2000.

[11] P. Corriveau, *et al.*, "Video quality experts group: Current results and future directions," *Proc. SPIE Visual Comm. and Image Processing*, vol. 4067, June 2000.

[12] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Information Theory*, vol. 4, pp. 525-536, 1974.

[13] W. S. Geisler and M. S. Banks, "Visual performance," in *Handbook of Optics* (M. Bass, ed.), McGraw-Hill, 1995.

[14] B. A. Wandell, *Foundations of Vision*, Sinauer Associates, Inc., 1995.

[15] L. K. Cormack, "Computational models, of early human vision," in *Handbook of Image and Video Processing* (A. Bovik, ed.), Academic Press, May 2000.

[16] A. C. Bovik, M. Clark, and W. S. Geisler, "Multichannel texture analysis using localized spatial filters," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, pp. 55-73, Jan. 1990.

[17] 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, Massachusetts: The MIT Press, 1993.

[18] J. Lubin, "A visual discrimination model for image system design and evaluation," in *Visual Models for Target Detection and Recognition*, E. Peli, ed., pp. 207-220, Singapore: World Scientific Publisher, 1995.

[19] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video quality assessment," *IEEE Trans. Multimedia*, vol. 4, pp. 129-132, Mar. 2002.

[20] Z. Wang and A. C. Bovik, "Embedded foveation image coding," *IEEE Trans. Image Processing*, vol. 10, pp. 1397-1410, Oct. 2001.

[21] C. J. van den Branden Lambrecht, *Perceptual models and architectures for video coding applications*, PhD thesis, Swiss Federal Institute of Technology, Aug. 1996.

[22] C. J. van den Branden Lambrecht, "A working spatio-temporal model of the human visual system for image restoration and quality assessment applications," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, pp. 2291-2294, 1996.

[23] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. SPIE*, vol. 2668, (San Jose, LA), pp. 450461, 1996.

[24] A. B. Watson, J. Hu, and J. F. III. McGowan, "DVQ: A digital video quality metric based on human vision," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 20-29, 2001.

[25] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. SPIE*, vol. 2179, pp. 127-141, 1994.

[26] E. Peli, "Contrast in complex images," *Journal of Optical Society of America*, vol. 7, pp. 2032-2040, Oct. 1990.

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

[28] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Proc. SPIE*, vol. 1616, pp. 2-15, 1992.

[29] S. Daly, "The visible difference predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision* (A. B. Watson, ed.) pp. 179-206, Cambridge, Massachusetts: The MIT Press, 1993.

[30] D. J. Heeger and T. C. Teo, "A model of perceptual image fidelity," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 343-345, 1995.

[31] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. SPIE*, vol. 2668, (San Jose, LA), pp. 450461, 1996.

[32] C. J. van den Branden Lambrecht, D. M. Costantini, G. L. Sicuranza, and M. Kunt, "Quality assessment of motion rendition in video

coding," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 9, pp. 766-782. Aug. 1999.

[33] A. B. Watson and J. A. Solomon, "Model of visual contrast gain control and pattern masking," *Journal of Optical Society of America*, vol. 14, no. 9, pp. 2379-2391, 1997.

[34] W. Xu and G. Hauske, "Picture quality evaluation based on error segmentation", *Proc. SPIE*, vol. 2308, pp. 1454-1465, 1994.

[35] W. Osberger, N. Bergmann, and A. Maeder, "An automatic image quality assessment technique incorporating high level perceptual factors," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 414-418, 1998.

[36] T. N. Pappas and R. J. Safranek, "Perceptual criteria for image quality evaluation," in *Handbook of Image and Video Processing* (A. Bovik, ed.), Academic Press, May 2000.

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

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

[39] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Information Theory*, vol. 38, pp. 587-607, 1992.

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

[41] 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. IEEE Int. Conf. Acoust. Speech, and Signal Processing*, pp. 1945-1948, May 1989.

[42] J. W. Woods and S. D. O'Neil, "Subband coding of images," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 34, pp. 1278-1288, Oct. 1986.

[43] A. P. Bradley, "A wavelet difference predictor," *IEEE Trans. Image Processing*, vol. 5, pp. 717-730, May 1999.

[44] 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.

[45] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using the wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 205-220, Apr. 1992.

[46] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, "Image quality assessment based on a degradation model," *IEEE Trans. Image Processing*, vol. 4, pp. 636-650, Apr. 2000.

[47] T. D. Kite, N. Damera-Venkata, B. L. Evans, and A. C. Bovik, "A high quality, fast inverse halftoning algorithm for error diffused halftones," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 2, pp. 59-63, Oct. 1998.

[48] S. A. Karunasekera and N. G. Kingsbury, "A distortion measure for blocking artifacts in images based on human visual sensitivity," *IEEE Trans. Image Processing*, vol. 4, pp. 713-724, June 1995.

[49] S. A. Karunasekera and N. G. Kingsbury, "A distortion measure for image artifacts based on human visual sensitivity," in *Proc. IEEE Int.*

Conf. Acoust., Speech, and Signal Processing, vol. 5, pp. 117-120, 1994.

[50] C. H. Chou and Y. C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 5, pp. 467-476, Dec. 1995.

[51] M. Miyahara, K. Kotani, and V. R. Algazi, "Objective picture quality scale (PQS) for image coding," *IEEE Trans. Communications*, vol. 46, pp. 1215-1225, Sept. 1998.

[52] T. Yamashita, M. Kameda and M. Miyahara, "An objective picture quality scale for video images (PQS_{video}) - definition of distortion factors," *Proc. SPIE*, vol. 4067, pp. 801-809, 2000.

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

[54] K. T. Tan and M. Ghanbari, "A multi-metric objective picture-quality measurement model for MPEG video," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 10, pp. 1208-1213, Oct. 2000.

[55] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, pp. 247-278, Nov. 1998.

[56] S. Winker, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, pp. 231-252, 1999.

[57] A. B. Watson, "Toward a perceptual video quality metric," in *Proc. SPIE Human Vision and Electronic Imaging III*, vol. 3299, pp. 139-147, Jan. 1998.

[58] Z. Yu, H. R. Wu, S. Winkler, and T. Chen, "Vision-model-based impairment metric to evaluate blocking artifact in digital video," *Proceedings of the IEEE*, vol. 90, pp. 154-169, Jan. 2002.

[59] N. Graham, J. G. Robson, and J. Nachmias, "Grating summation in fovea and periphery," *Vision Research*, vol. 18, pp. 815-825, 1978.

[60] E. P. Simoncelli, "Modeling the joint statistics of images in the wavelet domain," in *Proc. SPIE*, vol. 3813, pp. 188-195, July 1999.

[61] J. Liu and P. Moulin, "Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients," *IEEE Trans. Image Processing*, vol. 10, pp. 1647-1658, Nov. 2001.

[62] J. M. Shapiro, "Embedded image coding using zerotrees of wavelets coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 2445-3462, Dec. 1993.

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

[64] D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals*, Standards, and Practice, Kluwer Academic Publishers, 2001.

[65] D. R. Fuhrmann, J. A. Baro, and J. R. Cox Jr., "Experimental evaluation of psychophysical distortion metrics for JPEG-encoded images," *Journal of Electronic Imaging*, vol. 4, pp. 297-406, Oct. 1995.

[66] W. F. Good, G. S. Maitz, and D. Gur, "Joint photographic experts group (JPEG) compatible data compression of mammograms," *Journal of Digital Imaging*, vol. 7, no. 3, pp. 123-132, 1994.

[67] M. R. Frater, J. F. Arnold, and A. Vahedian, "Impact of audio on subjective assessment of video quality in videoconferencing applications," *IEEE Journal of Selected Areas in Comm.*, vol. 11, pp. 1059-1062, Sept. 2001.

[68] Z. Wang, *Rate Scalable Foveated Image and Video Communications*, PhD thesis, Dept. of ECE, The University of Texas at Austin, Dec. 2001.

[69] Z. Wang, "Demo imaegs and free software for 'A Universal Image Quality Index'," in *http://anchovy.ece.utexas.edu/zwang/research/quality_index/demo.html*, 2001.

[70] Z. Wang, L. Lu and A. C. Bovik, "Video quality assessment using structural distortion measurement," *Proc. IEEE Int. Conf. Image Proc.*, Sept. 2002.

[71] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "An objective video quality assessment system based on human perception," *Proc. SPIE*, vol. 1913, pp. 15-26, 1993.

[72] S. Wolf and M. H. Pinson, "Spatio-temporal distortion metrics for in-service quality monitoring of any digital video system," *Proc. SPIE*, vol. 3845, pp. 266-277, 1999.

[73] O. Sugimoto, R. Kawada, M. Wada, and S. Matsumoto, "Objective measurement scheme for perceived picture quality degradation caused by MPEG encoding without any reference pictures," *Proc. SPIE*, vol. 4310, pp. 932-939, 2001.

[74] M. C. Q. Farias, S. K. Mitra, M. Carli, and A. Neri, "A comparison between an objective quality measure and the mean annoyance values of watermarked videos," in *Proc. IEEE Int. Conf. Image Proc.*, Sept. 2002.

[75] Z. Wang, A. C. Bovik and B. L. Evans, "Blind measurement of blocking artifacts in images," *Proc. IEEE Int. Conf. Image Proc.*, vol. 3, pp. 981-984, Sept. 2000.

[76] A. C. Bovik and S. Liu, "DCT-domain blind measurement of blocking artifacts in DCT-coded images," *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, vol. 3, pp. 1725-1728, May 2001.

[77] P. Gastaldo, S. Rovetta and R. Zunino, "Objective assessment of MPEG-video quality: a neural-network approach," in *Proc. IJCNN*, vol. 2, pp. 1432-1437, 2001.

[78] M. Knee, "A robust, efficient and accurate single-ended picture quality measure for MPEG-2," available at *http://www-ext.crc.ca/vqeg/frames.html*, 2001.

[79] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Processing Letters*, vol. 4, pp. 317-320, Nov. 1997.

[80] J. E. Caviedes, A. Drouot, A. Gesnot, and L. Rouvellou, "Impairment metrics for digital video and their role in objective quality assessment," *Proc. SPIE*, vol. 4067, pp. 791-800, 2000.

[81] VQEG: The Video Quality Experts Group, *http://www.vqeg.org*.

[82] H. R. Sheikh, Z. Wang, L. Cormack and A. C. Bovik, "Blind quality assessment for JPEG2000 compressed images," *Proc. IEEE Asilomar Conference on Signals, Systems and Computers,* Nov. 2002.

[83] Z. Wang, H. R. Sheikh and A. C. Bovik, “No-reference perceptual quality assessment of JPEG compressed images,” *Proc. IEEE Int. Conf. Image Proc.*, Sept. 2002..

[84] ITU-R Rec. BT. 500-10, *Methodology for the Subjective Assessment of Quality for Television Pictures*.

[85] B. Li, G. W. Meyer, and R. V. Klassen, "A comparison of two image quality models," in *Proc. SPIE*, vol. 3299, pp. 98-109, 1998.

[86] A. Mayache, T. Eude, and H. Cherifi, "A comprison of image quality models and metrics based on human visual sensitivity," in *Proc. IEEE Int. Conf. Image Proc.*, pp. 409-413, 1998.

[87] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proc. SPIE*, vol. 1913, pp. 202-216, 1993.

[88] J. Xing and D. J. Heeger, "Measurement and modeling of center-surround suppression and enhancement," *Vision Research*, vol. 41, pp. 571-583, 2001.

[89] J. Xing, "An image processing model of contrast perception and discrimination of the human visual system," in *SID Conference*, (Boston), May 2002.

[90] A. B. Watson, “Visual detection of spatial contrast patterns: Evaluation of five simple models,” *Optics Express*, vol. 6, pp. 12-33, Jan. 2000.