Chapter 19 in *Optical and Digital Image Processing*, Wiley, 2010.

Visual Perception and Quality Assessment

Anush K. Moorthy[1], Zhou Wang[2] and Alan C. Bovik[1]

[1]The University of Texas at Austin, Austin, Texas, USA.
[2]The University of Waterloo, Ontario, Canada.

1 Introduction

'Quality' according to the International Standards Organization (ISO) is *the degree to which a set of inherent characteristics of a product fulfils customer requirements* [1]. Even though this definition seems relatively straightforward at first, introspection leads one to the conclusion that the ambiguity inherent in the definition makes the quality assessment task highly subjective and hence difficult to model. Indeed, over the years researchers in the field of visual quality assessment have found that judging the quality of an image or a video is a challenging task. The highly subjective nature of the task, coupled with the human visual systems' peculiarities make this an interesting problem to study and in this chapter we attempt to do just that.

This chapter is concerned with the algorithmic evaluation of quality of an image or video, which is referred to as objective quality assessment. What makes this task difficult is that the measure of quality produced by the algorithm should match up to that produced by a human assessor. In order to obtain a statistically relevant measure of what a human thinks the quality of an image or video is; a set of images or videos are shown to a group of human observers who are asked to rate the quality on a particular scale. The mean rating for an image or video is referred to as the mean opinion score (MOS) and is representative of the perceptual quality of that visual stimulus. Such assessment of quality is referred to as subjective quality assessment. In order to gauge the performance of an objective algorithm, the scores produced by the algorithm are correlated with MOS; a higher correlation is indicative of better performance. In this chapter, we focus on a subset of image/video quality assessment algorithms (IQA/VQA) which are referred to as full reference (FR) algorithms. In these algorithms; the original, pristine stimulus is available along with the stimulus whose quality is to be assessed. The FR IQA/VQA algorithm accepts as input the pristine reference stimulus and its distorted version and produces a score that is representative of the visual quality of the distorted stimulus [2].

One of the primary questions that arise when we talk of visual quality assessment is : 'why not use mean square error (MSE) for this purpose?'. MSE

between two N-dimensional vectors x and y is defined as:

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

A low MSE value indicates that the two vectors are similar. Generally, in order to follow a convention where a higher value indicates greater similarity, the peak signal to noise ratio (PSNR) is utilized. PSNR is defined as:

$$PSNR = 10log_{10}(\frac{L^2}{MSE}) \tag{2}$$

where, L is the dynamic range of the pixel values (eg. $L = 255$ for grayscale images).Through this chapter, we use MSE and PSNR interchangeably.

Let us now return to the valid question of why one should not use MSE for visual quality assessment. After all MSE has several elegant properties and is a prime candidate of choice to measure the deviation of one signal from another. How is visual quality assessment different from this task? The major difference for visual quality assessment (as for audio quality assessment) is the ultimate receiver. For images and videos, the ultimate receiver is the human observer. Immaterial of whether there exists a difference between the stimuli under test, the difference is not perceptually significant as long as the human is unable to observe the difference. This begs the question - are not all differences equally significant for a human? The answer is an emphatic NO! As vision researchers have observed, the human visual system (HVS) is replete with peculiarities. The properties of the HVS - as we shall see in the next section - govern the perceivability of the distortions and hence an algorithm that seeks to evaluate visual quality must be tuned human perception. MSE, as many researchers have argued, is not tuned to human perception and hence does not make for a good visual quality assessment algorithm [3, 4].

In this chapter we will begin with a short description of how visual stimulus is processed by the human. We shall then go on to describe various FR IQA/VQA algorithms. Our discussion then moves on to how one assess the performance of an IQA/VQA algorithm and we describe some standard databases that are used for this task. Finally, we conclude this chapter with a discussion of possible future research directions in the field of quality assessment.

2 The Human Visual System

The first stage of the human visual system (HVS) is the eye, where the visual stimulus passes through the optics of the eye and then on to the photoreceptors at the back of the eye. Even though the eye exhibits some peculiarities including lens aberrations; these are generally not modeled in HVS-based IQA/VQA algorithms. The optics of the eye are band-limited and act as a low-pass filter; hence some HVS-based IQA/VQA systems model this using a point-spread function (PSF).

The photoreceptors are classified as rods - which are responsible for vision under scotopic conditions and cones - which are responsible for vision under photopic conditions. Cones are also responsible for encoding color information. The distribution of rods and cones in the eye is not uniform. In fact, the number of photoreceptors are high at a region called the *fovea* and fall off as one moves away from the fovea [5]. Why this is important for IQA/VQA is because of the fact that the human does not assimilate the entire visual stimulus at the same 'resolution'. That part of the stimulus which is imaged on the fovea has the highest resolution and regions which are imaged farther away have lower resolution. In order to assimilate the entire stimulus, the human scans the image using a set of fixations followed by rapid eye movements called saccades. Little or no information is gathered during a saccade. This implies that for visual stimuli, certain regions may be of greater importance than others [6].

The information from the photoreceptors is then processed by the retinal ganglion cells. The ganglion cells are an interesting area of study and many researchers have devoted their energies toward such research. However, we will not dwell upon the ganglion cells here, the interested reader is referred to [5] for a thorough explanation. The information from the ganglion cells are passed onto the Lateral Geniculate Neucleus (LGN), which has been hypothesized to act as a 'relay' station [5, 7]. The LGN is the first location along the visual pathway where the information from the left and the right eye merges. The LGN receives not only the feed-forward information from the retinal cells, but also feed-back information from the next stage of processing - the primary visual cortex (area V1) [7]. The amount of feedback received leads one to believe that the LGN may not be *just* a relay station in the visual pathway [7]. Further, recent discoveries show that the LGN may perform certain normalization computations [8]. Further research in understanding the LGN and its functioning may be for import for visual quality assessment algorithm design.

Moving along, the information from the LGN is projected onto area V1 or the primary visual cortex. Area V1 is hypothesized to encompass two types of cells - simple cells and complex cells. Simple cells are known to be tuned to different orientations, scales and frequencies. This tuning of cells can be regarded as the HVS performing a scale-space-orientation decomposition of the visual stimulus (see Chapter 8). This is the rational behind many HVS based IQA/VQA systems performing a wavelet-like decomposition of the visual stimulus. Complex cells are currently modeled as receiving inputs from a set of simple cells [5, 7]. Complex cells are known to be direction selective. Even though V1 is connected to many other regions, one region of interest is area V5/MT which is hypothesized to play a key role in motion processing [9, 10]. Area MT along with its neighboring area MST are attributed with computing motion estimates. It is not surprising that motion processing is essential, since it allows us to perform many important tasks, including depth perception, tracking of moving objects, and so on. Humans are extremely good at judging velocities of approaching objects and in discriminating opponent velocities [5, 11]. A significant amount of neural activity is devoted to motion processing. Given that the HVS is sensitive to motion, it is imperative that objective measures of video quality take motion

into consideration.

Even though we have made progress in understanding the HVS, there is still a lot to be done. Indeed, some researchers have claimed that we have understood only a significantly small portion of the primary visual cortex [12]. Each of the above mentioned areas is an active field of research. Interested readers are directed to [5] and [7] for good overviews and [13] for an engineering perspective to understanding and analyzing the HVS.

Now that we have looked at the basics of human visual processing, let us list out some peculiarities of the HVS. These are relevant for IQA/VQA, since many of these peculiarities govern the discernibility of distortions and hence of quality.

Light Adaptation refers to the property that the HVS response depends much more upon the difference in the intensity between the object and the background than upon the actual luminance levels. This allows the human to see over a very large range of intensities.

Contrast sensitivity functions (CSFs) model the decreasing sensitivity of the HVS with increasing spatial frequencies. The HVS also exhibits varying sensitivity to temporal frequencies. Generally, most models for QA assume that the spatial and temporal responses are approximately separable. A thorough modeling would involve a spatio-temporal CSF [14].

Masking refers to the property of the HVS in which the presence of a 'strong' stimulus renders the weaker stimulus imperceptible. Types of masking include texture masking - where certain distortions are masked in the presence of strong texture; contrast masking - where regions with larger contrast mask regions with lower contrast and temporal masking - where the presence of a temporal discontinuity masks the presence of some distortions.

After having described the HVS briefly, we now run through some visual quality assessment algorithms. We broadly classify QA algorithms as i) those based on the HVS, ii) those that utilize a feature-based approach and iii) structural and information-theoretic approaches. For VQA, we also describe algorithms that utilize motion information explicitly - motion-modeling based approaches. While the HVS-based approaches seem the best way to evaluate visual quality, our limited understanding of the HVS leads to poor HVS models, which in turn do not function well as QA algorithms. Feature-based approaches employ heuristics and extracted features are generally only tenuously related to the HVS. Structural and information theoretic measures, on the other hand, utilize an approach based on Natural Scene Statistics (NSS) [15], which are hypothesized to be the inverse problem to that of modeling the HVS [16]. For VQA, explicit incorporation of motion is of prime importance and motion-modeling based approaches do just that.

3 Human Visual System based models

Human Visual System (HVS) based models for IQA/VQA generally follow a series of operations akin to those hypothesized to occur along the visual pathway

in humans. The first major component of these models in a linear decomposition of the stimulus over multiple scales and orientations. Contrast sensitivity is parameterized by a contrast sensitivity function (CSF). Generally, the spatial CSF is modeled using a low-pass filter (since the HVS is not as sensitive to higher frequencies) and the temporal CSF is modeled using band-pass filters. Parameters for the filters are estimated from psychovisual experiments. It is generally far more easier to model the spatial and temporal CSF's separately instead of modeling a spatio-temporal CSF [17]. The spatial and temporal responses of the HVS are approximately separable [5]. Masking is another HVS property that is taken into account for IQA. A good overview of HVS based models for IQA/VQA can be found in [18].

Visual Difference Predictor (VDP) Visual Difference Predictor (VDP) first applies a point-non linearity to the images to model the fact that visual sensitivity and perception of lightness are nonlinear functions of luminance, followed by a CSF [19]. A modified version of the Cortex transform [20] is then utilized to model the initial stage of the human *detection mechanisms*. Masking then follows. In order to account for the fact that the probability of detection increases with increase in stimulus contrast, VDP then applies a psychometric function followed by a probability summation.

Visual Discrimination Model (VDM) The Sarnoff Visual Discrimination Model(VDM) which was later modified to the popular Sarnoff JND metric for video [21] was proposed by Lubin in [22]. A PSF is first applied to the images, followed by a modeling of the retinal cone-sampling. A Laplacian pyramid performs a decomposition of the signal and a contrast energy measure is computed. This measure is then processed through a masking function and a just-noticeable-difference (JND) distance measure is computed to produce the quality index.

Teo and Heeger model Teo and Heeger proposed a model for IQA based on the HVS [23]. The model performs a linear decomposition of the reference and test images using a hex-quadrature mirror filter (QMF), and then squares each coefficient at the output. Contrast normalization is accomplished by computing:

$$R^{\theta} = k\frac{(A^{\theta})^2}{\sum_{\phi}(A^{\phi})^2 + \sigma^2} \tag{3}$$

where, A^{θ} is a coefficient at the output of the linear transform at orientation θ and k, σ^2 are the scaling and saturation constants. ϕ sums over all the possible orientations of the linear transform, thus performing a normalization. The final error measure is then the vector distance between the responses of the test and reference images.

Visual Signal to Noise Ratio (VSNR) Proposed by researchers at Cornell, Visual Signal to Noise Ratio (VSNR), which aims to evaluate the effect of

supra-threshold distortion, utilizes parameters for the HVS model derived from experiments where the stimulus was an actual image as against sinusoidal gratings or Gabor patches [24]. Many arguments that support the use of natural images/videos for estimating HVS parameters are enlisted in [13]. VSNR first computes a difference image from the reference and distorted images. This difference image is then subjected to a discrete wavelet transform. Within each subband, VSNR then computes the visibility of distortions, by comparing the contrast of the distortion to the detection threshold and then computes the RMS contrast of the error signal (d_{pc}). Finally, using a strategy inspired from what is termed as *global precedence* in the HVS, VSNR computes a global precedence preserving contrast (d_{gp}). The final index is a linear combination of d_{pc} and d_{gp}.

Digital Video Quality Metric (DVQ) Watson et. al. proposed the digital video quality metric (DVQ) which evaluates visual quality in the Discrete Cosine Transform (DCT) domain [25]. We note that even though the algorithm is labeled as a 'metric', it is not a metric in the true sense of the word. We continue its usage in this chapter for this and other metrics, however appropriate use of terminology for IQA/VQA metrics is essential (see [26] for relevant discussion). DVQ metric evaluates the quality in the YOZ opponent color space [27]. We note that this space is unusual in the sense that most researchers operate in the YUV color space. However the authors propose arguments for their choice [25]. A 8×8 block DCT is then performed on the reference and test videos. The ratio of the DCT (AC) amplitude to the DC amplitude is computed as the local contrast, which is then converted into just-noticeable-differences (JNDs) using thresholds derived from a small human study. Contrast masking follows. The error scores (which can be viewed as quality scores) are then computed using a Minkowski formulation.

Moving Picture Quality Metric (MPQM) The metric proposed by Van den Branden Lambrecht in [28] uses Gabor filters for spatial decomposition, two temporal mechanisms and a spatio-temporal CSF. It also models a simple intra-channel contrast masking. One difference in MPQM is the use of segmentation to identify regions - uniform areas, contours and textures - within an image and the error scores in each of these regions is pooled separately. An elementary evaluation of the metric was done to demonstrate its performance.

Scalable wavelet-based distortion metric for VQA This VQA algorithm filters the reference and test videos using a lowpass filter [29]. The Haar wavelet transform is then used to perform a spatial frequency decomposition. A subset of these coefficients is selected for distortion measurement. This model utilizes a CSF for weighting as well as masking. The reason this metric differs from other HVS-based metrics is that the parameters used for the CSF and masking function are derived from human responses to natural videos as against sinusoidal gratings as is generally the case (as in VSNR [24]). The algorithm then

computes a quality score for the distorted video using the differences in the responses from the reference and distorted videos.

4 Feature based models

Feature-based models generally extract features from images or videos, which are deemed to be of importance in visual quality assessment. For example, some algorithms extract the presence of edges and the relative edge strength using simple edge filters or a Fourier analysis. Some of the models extract elementary motion features for VQA. The primary argument against such models is the fact that the extracted features may not be correlated with the HVS. Some other arguments include the use of unmotivated thresholds and pooling strategies. However, as we shall see, some of these models perform well in terms of correlation with human perception.

A Distortion Measure Based on Human Visual Sensitivity Karunasekera and Kingsbury filter the difference (error) image (computed from the reference and distorted images) using direction filters - vertical, horizontal and 2 diagonal orientations to form oriented edge images [30]. The outputs of each of these orientations is processed independently, and then pooled to produce a distortion measure. Masking computation based on a activity measure and brightness is undertaken. A non-linearity is then applied to obtain the directional error. This model proposed in [30], is an extension of the authors' previous blocking measure proposed in [31] and incorporates ringing and blurring measures using the above described process.

Singular Value Decomposition and Quality Singular value decomposition (SVD) is a well known tool from linear algebra which has been used for a host of multimedia applications. In [32], the authors use SVD for image quality assessment. SVD is applied on 8×8 blocks in the reference and test images and then the distortion per block is computed as $D_i = \sqrt{\sum_{i=1}^{n}(s_i - \widehat{s_i})}$, where s_i and $\widehat{s_i}$ are the singular values for block i from the reference and test images. The final quality score is computed as: $M-SVD = \sum_{i \in all\_blocks} |D_i - D_{mid}|$, where D_{mid} represents the median of the block distortions. Even though the authors claim that the algorithm performs well, its relation to the HVS is unclear as is the significance of the SVD for IQA.

Curvature-based Image Quality Assessment The IQA index proposed in [33], first uses the discrete wavelet transform to decompose the reference and test images. In each subband, mean surface curvature maps are obtained as:

$$H = \frac{I_{uu} + I_{vv} + I_{uu}I_v^2 + I_{vv}I_u^2 - 2I_uI_vI_{uv}}{2(1 + I_u^2 + I_v^2)^{3/2}} \tag{4}$$

where I_{uu}, I_{vv}, I_u and I_v are the partial derivatives of the image I. The correlation coefficient between the curvatures of the original and distorted images

is then evaluated. These correlation coefficients are then collapsed across the subbands to produce a quality score.

Perceptual Video Quality Metric (PVQM) Proposed by Swisscom/KPN research, the perceptual video quality metric (PVQM) extracts three features from the videos under consideration [34]. *Edginess* is essentially indicated by a difference between the (dialated) edges of the reference and distorted frames computed using an approximation to the local gradient. The edginess indicator is supposed to reflect the loss in spatial detail. The *temporal indicator* for a frame is the correlation coefficient between adjacent frames subtracted from 1. Finally, a *chrominance indicator* based on color saturation is computed. Each of these indicators are pooled separately across the video and then a weighted sum of these pooled values is utilized as the quality measure. PVQM utilizes a large number of thresholds for each of the indicators as well as for the final pooling. Some of these thresholds are claimed to be based on psychovisual evaluation.

Video Quality Metric (VQM) Pinson and Wolf proposed the video quality metric (VQM) [35] which was the top performer in the video quality experts group (VQEG) phase-II studies [36]. Owing to its performance, VQM has also been standardized by the American National Standards Institute and the International Telecommunications Union (ITU) has included VQM as a normative measure for digital cable television systems [37]. VQM which was trained on the VQEG phase-I dataset [38], first performs a spatio-temporal alignment of the videos followed by gain and offset correction. This is followed by extraction of a set of features which are thresholded. The final quality score is computed as a weighted sum of these features. The computed features include a feature that describes the loss of spatial detail; one that describes a shift in the orientation of the edges; one that describes the spread of chrominance components as well as one to describe severe color impairments. VQM also includes elementary motion information in the form of the difference between frames and a quality improvement feature that accounts for improvements arising from (for example) sharpness operations. The quality score ranges from 0 to 1.

Temporal Variations of Spatial Distortion based VQA Ninassi et. al. proposed a VQA index recently [39]. They model temporal distortions like mosquito noise, flickering, jerkiness as an evolution of spatial distortions over time. A spatio-temporal tube consisting of a spatio-temporal chunk of the video computed from motion vector information is created, which is then evaluated for its spatial distortion. The spatial distortion is computed using the WQA quality index [40]. A temporal filtering of the spatial distortion is then undertaken, followed by a measurement of the temporal variation of the distortion. The quality scores are then pooled across the video to produce the final quality index.

Temporal Trajectory Aware Quality Measure One of the few algorithms that utilize motion information is the one proposed by Barkowsky et. al. - the Tetra Video Quality Metric [41]. Information from a block-based motion estimation algorithm for each (heuristically determined) shot [42] is utilized for temporal trajectory evaluation of the distorted video. This information is logged in a temporal information buffer, which is followed by a temporal visibility mask. Spatial distortion is evaluated by MSE - the authors claim that this is for reducing the complexity of the algorithm. A spatial-temporal distortion map is then created. The pooling stage first applies a mask to reflect human foveation and then a temporal summation is performed. The proposed algorithm also models the frame rate, pauses and skips. All of these indicators are then combined to form a quality score for the video.

5 Structural and Information Theoretic models

A structural approach for IQA was proposed by Wang and Bovik in [43]. This approach was later modified for VQA [44]. These models are based on the premise that the HVS extracts (and is hence sensitive to) structural information in the stimulus. Loss of structural information is hence related to perceptual loss of quality. Information theoretic models utilize natural scene statistics (NSS) in order to quantify loss of information in the wavelet domain. Recent research indicates how these two metrics are closely related to each other and to mutual masking hypothesized to occur in the HVS [45].

Single Scale Structural Similarity Index (SS-SSIM) For two image patches drawn from the same location of the reference and distorted images - $\mathbf{x} = \{x_i | i = 1, 2, \ldots, N\}$ and $\mathbf{y} = \{y_i | i = 1, 2, \ldots, N\}$ -respectively SS-SSIM computes three terms - luminance, contrast and structure as [43]:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{5}$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{6}$$

$$s(\mathbf{x}, \mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{7}$$

where C_1, C_2 and C_3 are small constants. The constants C_1, C_2 and C_3 ($C_3 = C_2/2$) are included to prevent instabilities from arising when the denominator tends to zero. μ_x, μ_y, σ_x^2, σ_y^2 and σ_{xy} are the means of $\mathbf{x}$, $\mathbf{y}$, the variances of $\mathbf{x}$, $\mathbf{y}$ and the covariance between $\mathbf{x}$ & $\mathbf{y}$ respectively, computed using a sliding window approach. The window used is a 11×11 circular-symmetric Gaussian weighting function $w = \{w_i | i = 1, 2, \ldots, N\}$, with standard deviation of 1.5 samples, normalized to sum to unity ($\sum_{i=1}^{N} w_i = 1$).

Finally,the SSIM index between signal $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$SSIM(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{8}$$

This index produces a map of quality scores having the same dimensions as that of the image. Generally, the mean of the scores is utilized as the quality index for the image. The SSIM index is an extension of the Universal Quality Index (UQI) [46], which is the SSIM index with $C_1 = C_2 = 0$.

Multi-scale Structural Similarity Index (MS-SSIM) Images are naturally multi-scales. Further, the perception of image quality depends upon a host of scale-related factors. In order to evaluate image quality at multiple resolutions, in [47], the Multi-Scale SSIM (MS-SSIM) index was proposed.

In MS-SSIM, quality assessment is accomplished over multiple scales of the reference and distorted image patches (the signals defined as $\mathbf{x}$ and $\mathbf{y}$ in the previous discussion on SS-SSIM) by iteratively low-pass filtering and downsampling the signals by a factor of 2. The original image scale is indexed as 1, the first down-sampled version is indexed as 2 and so on. The highest scale M is obtained after $M - 1$ iterations.

At each scale j, the contrast comparison (6) and the structure comparison (7) terms are calculated and denoted $c_j(\mathbf{x}, \mathbf{y})$ and $s_j(\mathbf{x}, \mathbf{y})$, respectively. The luminance comparison (5) term is computed only at scale M and is denoted $l_M(\mathbf{x}, \mathbf{y})$. The overall SSIM evaluation is obtained by combining the measurement over scales:

$$SSIM(\mathbf{x}, \mathbf{y}) = [l_M(\mathbf{x}, \mathbf{y})]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(\mathbf{x}, \mathbf{y})]^{\beta_j} \cdot [s_j(\mathbf{x}, \mathbf{y})]^{\gamma_j} \tag{9}$$

The highest scale used here is $M = 5$.

The exponents α_j, β_j, γ_j are selected such that $\alpha_j = \beta_j = \gamma_j$ and $\sum_{j=1}^{M} \gamma_j = 1$.

SSIM Variants SS-SSIM proposed in [43] was extended to the complex wavelet domain in [48] and the proposed index was labeled as the complex wavelet structural similarity index (CW-SSIM). CW-SSIM is computed as:

$$S(c_x, c_y) = \frac{2|\sum_{i=1}^{N} c_{x,i} c_{y,i}^*| + K}{\sum_{i=1}^{N} |c_{x,i}|^2 + \sum_{i=1}^{N} |c_{y,i}|^2 + K} \tag{10}$$

where, $c_x = \{c_{x,i} | i = 1, \ldots, N\}$ and $c_y = \{c_{y,i} | i = 1, \ldots, N\}$ are two sets of complex wavelet coefficients drawn from the same location in the reference and test images. CW-SSIM has been used for face recognition [49] and for a host of other applications [50].

Other variants of SSIM which modify the pooling strategy from the mean to fixation-based pooling [51], percentile pooling [6] and information-content-based pooling [52] have also been proposed. Modifications of SSIM include a gradient-based approach [53] and another technique based on pooling three perceptually

important parameters [54]. In [55], a classification based on the type of region is undertaken, after applying SSIM on the images. A weighted sum of the SSIM scores from each of the regions is combined to produce a score for the image.

Visual Information Fidelity (VIF) It is known that when images are filtered using oriented band-pass filters (eg. a wavelet transform), the distribution of resulting (marginal) coefficients are highly peaked around zeros and possess heavy tails [15]. Such statistical descriptions of natural scenes are labeled as natural scene statistics (NSS) and NSS has been an active area of research. Visual Information Fidely (VIF) [56] utilizes the Gaussian scale mixture (GSM) model for wavelet NSS [57]. VIF first performs a scale-space-orientation wavelet decomposition using the steerable pyramid [58] and models each subband in the source as $C = S \cdot U$, where S is a random field (RF) of scalars and U is a Gaussian vector RF. The distortion model is $D = GC + \nu$, where G is a scalar gain field and ν is additive Gaussian noise RF. VIF then assumes that the distorted and source images pass through the human visual system and the HVS uncertainty is modeled as *visual noise*: N and N' for the source and distorted image respectively; where N and N' are zero-mean uncorrelated multivariate Gaussians. It then computes $E = C + N$ and $F = D + N'$. The VIF criterion is then evalauted as:

$$VIF = \frac{\sum_{j \in allsubbands} I(C^j; F^j | s^j)}{\sum_{j \in allsubbands} I(C^j; E^j | s^j)}$$

where, $I(X;Y|Z)$ is the conditional mutual information between X and Y, conditioned on Z ; s^j is a realization of S^j for a particular image and the index j runs through all the sub bands in the decomposed image.

Structural Similarity for VQA SS-SSIM defined for images was applied on a frame-by-frame basis on videos for VQA and was shown to perform well [44]. Realizing the importance of motion information, the authors in [44] also utilized a simple motion-based weighting strategy that was used to weight the spatial SSIM scores. Video SSIM - the resulting algorithm was shown to perform well [44].

Video VIF VIF was extended to VQA in [59]. The authors justified the use of VIF for video by first motivating the GSM model for the spatio-temporal natural scene statistics. The model for video VIF then is essentially the same as that for VIF with the exception being the application of the model to the spatio-temporal domain as opposed to the spatial domain.

6 Motion Modeling Based Algorithms

As described before, the areas MT and MST in the human visual system are responsible for motion processing. Given that the HVS is sensitive to motion, it is imperative that objective measures of quality take motion into consideration.

An observant reader would have observed that the models for VQA described so far were essentially IQA algorithms applied on a frame-by-frame basis. Some of these algorithms utilized motion-information, however, the incorporation was ad-hoc. In this section, we describe some recent VQA algorithms that incorporate motion information. We believe that the importance of spatio-temporal quality assessment as against a spatial-only technique for VQA cannot be understated.

Speed-weighted Structural Similarity Index (SW-SSIM) Speed-weighted SSIM (SW-SSIM) first computes SS-SSIM at each pixel location in the video using SS-SSIM in the spatial domain [60]. Motion estimates are then obtained using Black and Anandan's optical flow computation algorithm [61]. Using a histogram based approach for each frame, a global motion vector for that frame is identified. Relative motion is then extracted as the difference between the absolute motion vectors (obtained from optical flow) and the computed global motion vectors. Then, a weight for each pixel is computed. This weight is a function of the computed relative and global motion and the stimulus contrast. The weight so obtained is then used to weight the SS-SSIM scores. The weighted scores are then pooled across the video and normalized to produce the quality index for the video. The described weighting scheme was inspired by the experiments into human visual speed perception by Stocker and Simoncelli [62].

Motion-based Video Integrity Evaluation (MOVIE) Motion-based Video Integrity Evaluation (MOVIE) first decomposes the reference and test videos using a multi-scale spatio-temporal Gabor filter-bank [63]. Spatial quality assessment is conducted using a technqiue similar to MS-SSIM. A modified version of the Jepson and Fleet algorithm for optical flow[64] is used to produce motion estimates. The same set of Gabor filters are utilized for optical flow computation and quality assessment. Translational motion in the spatio-temporal domain manifests itself as a plane in the frequency domain [65]. MOVIE assigns positive excitatory weights to the response of those filters which lie close to the spectra plane defined by the computed motion vectors and negative inhibitory weights to those filter responses which lie farther away from the spectral plane. Such weighting results in a strong response if the motion in the test and reference video coincide and a weak response is produced if the test video has motion deviant from that in the reference video. The mean-squared error between the responses from the test and reference filter banks then provides the temporal quality estimate. The final MOVIE index is the product of the spatial and temporal quality indices - a technique inspired from the spatial and temporal separability of the HVS.

We have described a handful of algorithms in this chapter. There exist many other algorithms for visual quality assessment that are not covered here. The reader is referred to [66, 18, 67, 68, 69, 2] for reviews of other such approaches

to FR QA. Further, we have covered only FR QA algorithms - partly due to the maturity of this field. Algorithms that do not require a reference stimulus for QA are called no-reference (NR) algorithms [2]. Some examples of NR IQA algorithms include [70, 71, 72, 73, 74, 75]. Some NR VQA algorithms can be found in [76, 77, 78]. There also exist another class of algorithms - reduced reference (RR) algorithms, in which the distorted stimulus contains some additional information about the pristine reference [2]. Recent RR QA algorithms can be found in [79, 80, 81].

Having described visual quality assessment algorithms, let us now move on to analyzing how performance of these algorithms can be computed.

7 Performance Evaluation & Validation

Now that we have gone through the nuts-and-bolts of a set of algorithms, the question of evaluating the performance of the algorithm follows. Simply demonstrating that the ranking of a handful of videos/images produced by the algorithm scores matches human perception is not an accurate measure of algorithm performance. In order to compare algorithm performance a common testing ground that is publicly available must be used. This testing ground, which we will refer to as a dataset, must contain a large number of images/videos which have undergone all possible kinds of distortions. For eg., for IQA, distortions should include - gaussian noise, blur, distortion due to packet-loss, distortion due to compression and so on. Further, each image in the dataset must be rated by a sizeable number of human observers in order to produce the subjective quality score for the image/video. This subjective assessment of quality is what forms the basis of performance evaluation for IQA/VQA algorithms. The International Telecommunication Union (ITU) has listed procedures and recommendations on how such subjective assessment is to be conducted [82]. After a group of human observers have rated the stimuli in the dataset, a mean opinion score (MOS) which is the mean of the ratings given by the observers (which is computed after subject rejection - see [82]) is computed for each of the stimuli. The MOS is representative of the perceived quality of the stimulus. In order to evaluate algorithm performance, each stimulus in the dataset is evaluated by the algorithm and receives an objective score. The objective and subjective scores are then correlated using statistical measures of correlation. The higher the correlation the better the performance of the algorithm.

Traditional measures of correlation that have been used to evaluate performance of visual quality assessment algorithms include - Spearman's rank ordered correlation coefficient (SROCC), linear (Pearson's) correlation coefficient (LCC), root mean squared error (RMSE) and outlier ratio (OR) [38]. SROCC is a measure of the prediction monotonicity and can directly be computed using algorithmic scores and MOS. OR is a measure of the prediction consistency. LCC and RMSE measure the prediction accuracy and are generally computed after transforming the algorithmic scores using a logistic function. This is because LCC assumes that the data under test are linearly related, and tries to quantify

this linear relationship. However, algorithmic quality scores may be non-linearly related to subjective MOS. After transforming the objective algorithmic scores using the logistic, we eliminate this non-linearity which allows us to evaluate LCC and RMSE. The logistic functions generally used are those proposed in [38] and [83]. The parameters of the logistic function are those that provide the best fit between the objective and subjective scores. In general, an algorithm with higher values (close to 1) for SROCC and LCC and lower values (close to 0) for RMSE and OR is considered to be a good visual quality assessment algorithm. One final statistical measure is that of statistical significance. This measure indicates whether, given a particular dataset, the obtained differences in correlation between algorithms is statistically significant. The F-Statistic and ANOVA [84] have been utilized in the past for this purpose [38, 83].

Even though the above described procedure is one that is currently adopted, there have been efforts directed at improving the performance evaluation of visual quality assessment algorithms. For example, a recently proposed approach by Wang and Simoncelli calls for the use of a technique called MAximum Differentiation competition (MAD) [85]. Images and video sequences live in a high-dimensional signal space, where the dimension equals the number of pixels. However, only hundreds or at most thousands of images can be tested in a practical subjective test, and thus their distribution is extremely sparse in the space of all possible images. Examining only a single sample from each orthant of an N-dimensional space would require a total of 2^N samples, an unimaginably large number for an image signal space with dimensionality on the order of thousands to millions. In [85] the authors propose an efficient methodology - MAD, where test images were synthesized to optimally distinguish competing perceptual quality models. Although this approach cannot fully prove the validity of a quality model, it offers an optimal means of selecting test images that are most likely to falsify it. As such, it provides a useful complementary approach to the traditional performance evaluation method. Other approaches based on Maximium Likelihood Difference Scaling (MLDS) [86] have been proposed as well [87, 88].

Now that we have an idea of the general procedure to evaluate algorithm performance, let us shift our attention to the databases which are used for these purposes. For IQA, one the most popular databases, which is currently the de facto standard for all IQA algorithms is the Laboratory for Image and Video Engineering (LIVE) image quality assessment database [83]. The LIVE image database was created at researchers at The University of Texas at Austin and consists of 29 reference images. Each reference image was distorted using five different distortion processes (eg. compression, noise etc.) and with varying level of severity for each distortion type. A subjective study was conducted as outlined earlier and each image was viewed by approximately 29 subjects. A host of leading IQA algorithms were then evaluated for their performance in [83]. MS-SSIM and VIF were shown to have correlated extremely well with human perception [83]. In order to demonstrate how the structural and information-theoretic approaches compare to the often-criticized PSNR, in table **??** we (partially) reproduce the SROCC values between these algorithms and DMOS on

	J2k#1	J2k#2	JPG#1	JPG#2	WN	Gblur	FF
PSNR	0.9263	0.8549	0.8779	0.7708	0.9854	0.7823	0.8907
MS-SSIM	0.9645	0.9648	0.9702	0.9454	0.9805	0.9519	0.9395
VIF	0.9721	0.9719	0.9699	0.9439	0.9282	0.9706	0.9649

Table 1: Spearman's rank ordered correlation coefficient (SROCC) on the LIVE image quality assessment database. J2k = JPEG2000 compression, JPG = JPEG compression, WN = white noise, Gblur = Gaussian blur, FF = Rayleigh fast-fading channel.

the LIVE IQA database. The reader is referred to [83] for other correlation measures and statistical analysis for these and other algorithms.

Other IQA datasets include the one from researchers at Cornell [89], the IVC dataset [90], the TAMPERE image dataset [91] and the one from researchers at Toyama university [92]. Readers interested in a package that encompasses some of the discussed algorithms are referred to [93].

For VQA, the largest known publicly available dataset is that from the video quality experts group (VQEG) and is labeled as the VQEG FRTV phase-I dataset[1] [38]. Even though the VQEG has conducted other studies, none of the data has been made publicly available [36, 94]. The VQEG dataset consists of a total of 320 distorted videos created by distorting 20 reference video sequences. Even though the VQEG dataset has been widely used, it is not without its flaws. The VQEG dataset (and the associated study) was specifically aimed at television and hence contains interlaced videos. De-interlacing algorithms used before VQA may add distortions of their own thereby possibly reducing the correlation of the algorithmic scores with subjective DMOS. Further the dataset consists of non-natural videos, and many VQA algorithms which rely on the assumption of natural scenes face a distinct disadvantage [2] [63]. Again, the perceptual separation of the dataset is such that humans and algorithms have difficulty in making consistent judgments. It is to alleviate many such problems and to provide for a common publicly available testing bed for future VQA algorithms that the researchers at LIVE have developed the LIVE video quality assessment and LIVE wireless video quality assessment databases [95, 96]. These databases have recently been made publicly available at no cost to researchers in order to further the field of VQA.

Before we conclude this chapter, we would like to stress on the importance of utilizing a common publicly available test-set for evaluating IQA and VQA algorithms. Even though a host of new IQA/VQA algorithms have been proposed, comparing algorithmic performance makes no sense if one is unable to see relative algorithmic performance. The LIVE image dataset and the associated scores for algorithms, for example, allow for an objective comparison of IQA algorithms. The publicly available dataset must encompass a range of distortions

[1] We refer to this as the VQEG dataset henceforth

[2] This is not only for those algorithms that explicitly utilize NSS models, like VIF; but also models like SSIM which have been shown to be equivalent to NSS-based models [45]

and perceptual distortion levels so as to test accutately the algorithm. Reporting results on small and/or private test sets fails to prove the performance of the proposed algorithm.

8 Conclusion

In this chapter we undertook a brief tour of the human visual system (HVS) and studied some HVS based algorithms for image and video quality assessment (IQA/VQA). We then looked at some feature-based approaches and moved on to describe structural and information theoretic measures for IQA/VQA. We then stressed the importance of motion modeling for VQA and described recent algorithms that incorporate motion information. This was followed by a description of performance evaluation techniques for IQA/VQA and relevant databases for this purpose.

The area of visual quality assessment has been an active area of research for a long period of time. With the advent of structural approaches (and its variants) for image quality assessment, the field of FR IQA seems to have found a winner [83]. This is mainly because the simplicity of the structural approach coupled with its performance make it an algorithm of choice in practical systems. Further simplifications of the index make it even more attractive [97]. Even though authors have tried to improve the performance of these approaches, the improvements have been minimal and in most cases do not justify the additional complexity [52, 51, 6]. What remains un-answered at this stage is the relationship between the structural approach and the human visual system. Even though some researchers have demonstrated a link [45], further research in understanding the index needs to be performed.

FR VQA is a far tougher problem to solve, since incorporation of motion information and temporal quality assessment are fields still in the nascent stage. Even though the structural approach does well on the VQEG dataset, improvements can be achieved by using appropriate motion modeling based techniques [60].

The field of RR/NR QA is one that has tremendous potential for research. RR/NR algorithms are particularly useful in applications where quality evaluation is of utmost importance, but the reference stimulus is unavailable - for example in systems which monitor quality of service (QoS) at the end-user. Further, FR/NR/RR algorithms that not only perform well but are computationally efficient need to be developed for real-time quality monitoring systems.

Having reviewed some algorithms in this chapter, we hope that the reader has obtained a general idea of approaches that are utilized in the design and analysis of visual quality assessment systems and that it has piqued his interest in this broad field which involves researchers from a host of areas ranging from engineering to psychology.

References

[1] International Standards Organization (ISO), *http://www.iso.org/iso/home.htm*.

[2] Wang, Z. and Bovik, A. C., *Modern Image Quality Assessment*, Morgan & Claypool Publishers, 2006.

[3] Girod, B., "What's wrong with mean-squared error?, Digital images and human vision, A. B. Watson, Ed." pp. 207–220, 1993.

[4] Wang, Z. and Bovik, A. C., "Mean squared error: Love it or leave it? - a new look at fidelity measures," IEEE Signal Processing Magazine, 2009.

[5] Sekuler, R. and Blake, R., *Perception*, Random House USA Inc, 1988.

[6] Moorthy, A. K. and Bovik, A. C., "Visual importance pooling for image quality assessment," *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment*, 3, pp. 193–201, 2009.

[7] Hubel, D., Wensveen, J. and Wick, B., "Eye, brain, and vision," 1988.

[8] Mante, V., Bonin, V. and Carandini, M., "Functional mechanisms shaping lateral geniculate responses to artificial and natural stimuli," *Neuron*, 58, pp. 625–638, 2008.

[9] Born, R. and Bradley, D., "Structure and Function of Visual Area MT." *Annual Review of Neuroscience*, 28, pp. 157–189, 2005.

[10] Rust, N. C. et al., "How mt cells analyze the motion of visual patterns," *Nature Neuroscience*, 9, pp. 1421–1431, 2006.

[11] Wandell, B., *Foundations of vision*, Sinauer Associates, 1995.

[12] Olshausen, B. and Field, D., "How close are we to understanding V1?" *Neural Computation*, 17, pp. 1665–1699, 2005.

[13] Carandini, M. et al., "Do we know what the early visual system does?" *Journal of Neuroscience*, 25, pp. 10577–10597, 2005.

[14] Daly, S., "Engineering observations from spatiovelocity and spatiotemporal visual models," *Proc. of SPIE*, 3299, pp. 180–191, 1998.

[15] Simoncelli, E. and Olshausen, B., "Natural Image Statistics and Neural Representation," *Annual Review of Neuroscience*, 24, pp. 1193–1216, 2001.

[16] Sheikh, H. R., Bovik, A. C. and De Veciana, G., "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Transactions on image processing*, 14, pp. 2117–2128, 2005.

[17] Kelly, D. H., "Spatiotemporal variation of chromatic and achromatic contrast thresholds," *Journal of the Optical Society of America*, 73(6), pp. 742–750, 1983.

[18] Nadenau, M. et al., "Human Vision Models for Perceptually Optimized Image Processing–A Review," *Proceedings of the IEEE*, 2000, 2000.

[19] Daly, S., "Visible differences predictor: an algorithm for the assessment of image fidelity," *Proceedings of SPIE*, 1666, p. 2, 1992.

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

[21] Lubin, J. and Fibush, D., "Sarnoff JND vision model," *T1A1*, 5, pp. 97–612, 1997.

[22] Lubin, J., "A visual discrimination model for imaging system design and evaluation," *Vision Models for Target Detection and Recognition: In Memory of Arthur Menendez*, p. 245, 1995.

[23] Teo, P. and Heeger, D., "Perceptual image distortion," *SID INTERNATIONAL SYMPOSIUM DIGEST OF TECHNICAL PAPERS*, 25, pp. 209–209, 1994.

[24] Chandler, D. M. and Hemami, S. S., "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, 16, pp. 2284–2298, 2007.

[25] Watson, A., Hu, J. and McGowan III, J., "Digital video quality metric based on human vision," *Journal of Electronic Imaging*, 10, p. 20, 2001.

[26] Bovik, A., "Meditations on Visual Quality," *IEEE COMPSOC E-LETTER, Technology Advances*, 2009.

[27] Peterson, H., Ahumada Jr, A. and Watson, A., "An improved detection model for DCT coefficient quantization," *Human Vision, Visual Processing, and Digital Display IV*, pp. 191–201.

[28] Van den Branden Lambrecht, C. and Verscheure, O., "Perceptual quality measure using a spatiotemporal model of the human visual system," *Proceedings of SPIE.*, pp. 450–461, 1996.

[29] Masry, M., Hemami, S. and Sermadevi, Y., "A scalable wavelet-based video distortion metric and applications," *IEEE Transactions on circuits and systems for video technology*, 16, pp. 260–273, 2006.

[30] Karunasekera, S. and Kingsbury, N., "A distortion measure for image artifacts based on human visualsensitivity," *1994 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1994. ICASSP-94.*, 1994.

[31] Karunasekera, S. and Kingsbury, N., "A distortion measure for blocking artifacts in images based on human visual sensitivity," *IEEE Transactions on image processing*, 4, pp. 713–724, 1995.

[32] Shnayderman, A., Gusev, A. and Eskicioglu, A., "A multidimensional image quality measure using singular value decomposition," 5294, pp. 82–92, 2003.

[33] Yao, S. et al., "Image quality measure using curvature similarity," *IEEE International Conference on Image Processing*, pp. 437–440, 2007.

[34] Hekstra, A. et al., "PVQM–a perceptual video quality measure," *Signal Processing: Image Communication*, 17, pp. 781–798, 2002.

[35] Pinson, M. H. and Wolf, S., "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting*, pp. 312–313, 2004.

[36] (VQEG), V. Q. E. G., "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment phase II," *http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII*, 2003.

[37] Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference, International Telecommunications Union, *Std. ITU-T Rec. J. 144*, 2004.

[38] Video Quality Experts Group (VQEG), "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment phase I," *http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI*, 2000.

[39] Ninassi, A. et al., "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment*, 3, pp. 253–265, 2009.

[40] Ninassi, A. et al., "On the performance of human visual system based image quality assessment metric using wavelet domain," *Proc. SPIE Human Vision and Electronic Imaging XIII*, 6806, 2008.

[41] Barkowsky, M. et al., "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing, Issue on Visual Media Quality Assessment*, 3, pp. 266–279, 2009.

[42] Cotsaces, C., Nikolaidis, N. and Pitas, I., "Video shot detection and condensed representation. a review," *IEEE signal processing magazine*, 23, pp. 28–37, 2006.

[43] Wang, Z. et al., "Image quality assessment: From error measurement to structural similarity," *IEEE Signal Processing Letters*, 13, pp. 600–612, 2004.

[44] Wang, Z., Lu, L. and Bovik, A. C., "Video quality assesssment based on structural distortion measurement," *Signal Processing: Image communication*, pp. 121–132, 2004.

[45] Seshadrinathan, K. and Bovik, A. C., "Unifying analysis of full reference image quality assessment," *15th IEEE International Conference on Image Processing, 2008. ICIP 2008*, pp. 1200–1203, 2008.

[46] Wang, Z. and Bovik, A. C., "A universal image quality index," *IEEE Signal Processing Letters*, 9, pp. 81–84, 2002.

[47] Wang, Z., Simoncelli, E. P. and Bovik, A. C., "Multi-scale structural similarity for image quality assessment," *Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, (Asilomar)*, 2003.

[48] Wang, Z. and Simoncelli, E., "Translation insensitive image similarity in complex wavelet domain," *IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05)*, 2, 2005.

[49] Gupta, S., Markey, M. K. and Bovik, A. C., "Advances and challenges in 3D and 2D+ 3D human face recognition," in "Pattern recognition in biology," Nova Science Publishers, 2007.

[50] Sampat, M. P. et al., "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, (to appear).

[51] Moorthy, A. K. and Bovik, A. C., "Perceptually significant spatial pooling techniques for image quality assessment," *Human Vision and Electronic Imaging XIV. Proceedings of the SPIE*, 7240, 2009.

[52] Wang, Z. and Shang, X., "Spatial pooling strategies for perceptual image quality assessment," *IEEE international conference on Image Processing*, 2006.

[53] Chen, G., Yang, C. and Xie, S., "Gradient-based structural similarity for image quality assessment," *IEEE International Conference on Image Processing*, pp. 2929–2932, 2006.

[54] Li, C. and Bovik, A., "Three-Component Weighted Structural Similarity Index," *Proceedings of SPIE*, 7242, p. 72420Q, 2009.

[55] Gao, X., Wang, T. and Li, J., "A content-based image quality metric," *Lecture Notes in Computer Science*, 3642, p. 231, 2005.

[56] Sheikh, H. R. and Bovik, A. C., "Image information and visual quality," *IEEE Transactions on Image Processing*, 15, pp. 430–444, 2006.

[57] Wainwright, M. and Simoncelli, E., "Scale mixtures of Gaussians and the statistics of natural images," *Advances in neural information processing systems*, 12, pp. 855–861, 2000.

[58] Simoncelli, E. et al., "Shiftable multiscale transforms," *IEEE Transactions on Information Theory*, 38, pp. 587–607, 1992.

[59] Sheikh, H. R. and Bovik, A. C., "A visual information fidelity approach to video quality assessment," *First International Workshop on Video Processing and Quality Metrics for Conusmer Electronics*, 2005.

[60] Wang, Z. and Li, Q., "Video quality assessment using a statistical model of human visual speed perception," *Journal of the Optical Society of America*, 24, pp. B61–B69, 2007.

[61] Black, M. J. and Anandan, P., "The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields," *Computer Vision and Image Understanding*, 63, pp. 75–104, 1996.

[62] Stocker, A. and Simoncelli, E., "Noise characteristics and prior expectations in human visual speed perception," *Nature neuroscience*, 9, pp. 578–585, 2006.

[63] Seshadrinathan, K., *Video quality assessment based on motion models*, Ph.D. thesis, The University of Texas at Austin, 2008.

[64] Fleet, D. and Jepson, A., "Computation of component image velocity from local phase information," *International Journal of Computer Vision*, 5, pp. 77–104, 1990.

[65] Watson, A. B. and Ahumada, A. J., "Model of human visual-motion sensing," *Journal of the Optical Society of America A*, 2, pp. 322–342, 1985.

[66] Ahumada Jr, A., "Computational image quality metrics: A review," *SID Digest*, 24, pp. 305–308, 1993.

[67] Wang, Z., Sheikh, H. and Bovik, A., "Objective video quality assessment," *The Handbook of Video Databases: Design and Applications*, pp. 1041–1078, 2003.

[68] Seshadrinathan, K. et al., *Image quality assessment in The Essential Guide to Image Processing*, chapter 20, Academic Press, 2009.

[69] Seshadrinathan, K. and Bovik, A. C., *Video Quality Assessment in The Essential Guide to Video Processing*, chapter 14, Academic Press, 2009.

[70] Wang, Z., Bovik, A. and Evans, B., "Blind measurement of blocking artifacts in images," in "Proc. IEEE Int. Conf. Image Proc," volume 3, pp. 981–984, Citeseer, 2000.

[71] Li, X. et al., "Blind image quality assessment," in "Intl. Conf. on Image Processing, New York, USA," 2002.

[72] Marziliano, P. et al., "A no-reference perceptual blur metric," in "Proceedings of the International Conference on Image Processing," volume 3, pp. 57–60, 2002.

[73] Gastaldo, P. et al., "Objective quality assessment of displayed images by using neural networks," *Signal Processing: Image Communication*, 20, pp. 643–661, 2005.

[74] Sheikh, H. R., Bovik, A. C. and Cormack, L. K., "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Transactions on Image Processing*, 14, pp. 1918–1927, 2005.

[75] Gabarda, S. and Cristobal, G., "Blind image quality assessment through anisotropy," *J. Opt. Soc. Am. A*, 24, pp. B42–B51, 2007.

[76] Winkler, S., Sharma, A. and McNally, D., "Perceptual video quality and blockiness metrics for multimedia streaming applications," in "Proceedings of the International Symposium on Wireless Personal Multimedia Communications," pp. 547–552, 2001.

[77] Farias, M. and Mitra, S., "No-reference video quality metric based on artifact measurements," in "IEEE International Conference on Image Processing," volume 3, pp. 141–144, 2005.

[78] Pastrana-Vidal, R. and Gicquel, J., "A no-reference video quality metric based on a human assessment model," in "Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics VPQM," volume 7, pp. 25–26, 2007.

[79] Wang, Z. et al., "Quality-aware images," *IEEE Transactions on Image Processing*, 15, pp. 1680–1689, 2006.

[80] Hiremath, B., Li, Q. and Wang, Z., "Quality-aware video," 3, 2007.

[81] Li, Q. and Wang, Z., "Reduced-Reference Image Quality Assessment Using Divisive Normalization-Based Image Representation," *IEEE Journal of Selected Topics in Signal Processing*, 3, pp. 202–211, 2009.

[82] BT., "500-11:Methodology for the subjective assessment of the quality of television pictures.," *International Telecommuncation Union, Geneva, Switzerland*, 2002.

[83] Sheikh, H. R., Sabir, M. F. and Bovik, A. C., "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, 15, pp. 3440–3451, 2006.

[84] Sheskin, D., *Handbook of parametric and nonparametric statistical procedures*, CRC Press, 2004.

[85] Wang, Z. and Simoncelli, E. P., "Maximum differentiation (mad) competition: A methodology for comparing computational models of perceptual quantities," *Journal of Vision*, 8, pp. 1–13, 2008.

[86] Maloney, L. and Yang, J., "Maximum likelihood difference scaling," *Journal of Vision*, 3, pp. 573–585, 2003.

[87] Charrier, C., Maloney, L. T. and Knoblauch, H. C. K., "Maximum likelihood difference scaling of image quality in compression-degraded images," *Journal of the Optical Society of America*, 24, pp. 3418 – 3426, 2007.

[88] Charrier, C. et al., "Comparison of image quality assessment algorithms on compressed images," *(submitted to) SPIE conference on Image quality and System Performance*, 2010.

[89] Chandler, D. M. and Hemami, S. S., "A57 database," *http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html*, 2007.

[90] Le Callet, P. and Autrusseau, F., "Subjective quality assessment irccyn/ivc database," 2005, http://www.irccyn.ec-nantes.fr/ivcdb/.

[91] Ponomarenko, N. et al., "Tampere image database," *http://www.ponomarenko.info/tid2008.htm*, 2008.

[92] "Toyama image database," *http://mict.eng.u-toyama.ac.jp/mict/index2.html*.

[93] Gaubatz, M., "Metrix mux visual quality assessment package," *http://foulard.ece.cornell.edu/gaubatz/metrix_mux/*.

[94] Video Quality Experts Group (VQEG), "Final report of video quality experts group multimedia phase I validation test, TD 923, ITU Study Group 9," 2008.

[95] Seshadrinathan, K. et al., "LIVE video quality assessment database," *http://live.ece.utexas.edu/research/quality/live_video.html*.

[96] Moorthy, A. K. and Bovik, A., "LIVE wireless video quality assessment database," *http://live.ece.utexas.edu/research/quality/live_wireless_video.html*.

[97] Rouse, D. and Hemami, S., "Understanding and simplifying the structural similarity metric," *15th IEEE International Conference on Image Processing, 2008. ICIP 2008*, pp. 1188–1191, 2008.