Multimed Tools Appl (2011) 51:163–186
DOI 10.1007/s11042-010-0625-9

Automatic prediction of perceptual quality of multimedia signals—a survey

Kalpana Seshadrinathan · Alan Conrad Bovik

Published online: 19 October 2010

Abstract We survey recent developments in multimedia signal quality assessment, including image, audio, video, and combined signals. Such an overview is timely given the recent explosion in all-digital sensory entertainment and communication devices pervading the consumer space. Owing to the sensory nature of these signals, perceptual models lie at the heart of multimedia signal quality assessment algorithms. We survey these models and recent competitive algorithms and discuss comparison studies that others have conducted. In this context we also describe existing signal quality assessment databases. We envision that the reader will gain a firmer understanding of the broad topic of multimedia quality assessment, of the various subdisciplines corresponding to different signal types, how these signals types co-relate in producing an overall user experience, and what directions of research remain to be pursued.

Keywords Survey · Quality assessment · Video quality · Image quality · Structural SIMilarity · Motion-based video integrity evaluation · Audio quality · Full reference · Perception

1 Introduction

Recent years have witnessed an explosion of visual and multimedia applications across the globe. Digital television and other home entertainment applications, mobile multimedia applications on cellular phones, social networking applications such

K. Seshadrinathan (✉)
Intel Corporation, 3600 Juliette Lane, M/S SC12–303, Santa Clara, CA 94086, USA
e-mail: kalpsesh@gmail.com

A. C. Bovik
Dept. of Electrical and Computer Engg., The University of Texas at Austin,
1 University Station C0803, Austin, TX 78712, USA
e-mail: bovik@ece.utexas.edu

as Facebook, personal multimedia collections, immersive multimedia and virtual reality applications, video tele-conferencing, gaming and educational multimedia presentations are just a few examples of multimedia applications that have become an integral, even indispensable, part of peoples' lives. The rising use of multimedia applications has been paralleled by a rising increase in the quality of experience that people demand from such applications. While significant strides are being made in offering new and improved multimedia services, the value of such services can only be assessed by the quality of experience that they offer to the end user. The end-user of the multimedia service is a human being in an increasingly large number of applications and determining the human's opinion of quality is critical in the design and deployment of a multimedia service.

"Quality" can be defined in a number of ways depending on the application of the multimedia service and the end-user of the video. For instance, measuring the "quality" of signals derived in applications such as laser range scanning or camera image acquisition often deal with aspects of the imaging system. The definition of "quality" in applications where the end-user is a human observer needs to consider *perception* of the signal by human sensory systems. Even within the application realm of human users, the interpretation of "quality" can depend on the multimedia service and the task defined for the human user. For instance, "quality" can mean detectability of image components that characterize disease in medical imaging or readability in document imaging or intelligibility in an audio service. In the overwhelming majority of digital multimedia entertainment applications, we are interested in defining "quality" as the overall Quality of Experience derived by the user from the service. This overall Quality of Experience is often a function of how good the image or video component of the multimedia signal looks or how good the audio component sounds, as well as the interactions between sensory perception. In this survey, we restrict our discussion to the quality assessment (QA) of image, video and audio components of multimedia signals.

The application realm of multimedia QA is enormous as evidenced by the increasing interest in this field over the last decade. We depict the universe of applications that can benefit from QA methods in Fig. 1. All the applications in Fig. 1 target a human end-user and can utilize QA methods in performance evaluation and benchmarking of the multimedia system. Additionally, QA methods can be used in *perceptual optimization* of the multimedia service. In other words, system configuration parameters of a multimedia system can be manipulated to maximize the Quality of Experience of the end-user of the system.

There are three categories of QA algorithms: Full-Reference QA, Reduced Reference QA and No-Reference QA, also known as blind QA. Full-Reference QA algorithms operate on distorted media signals while having a pristine, ideal "reference" signal (of the same content) available for comparison. The vast majority of QA algorithms fall into this category because of the relative simplicity of making quality judgments relative to a standard. Reduced-Reference QA algorithms operate without the use of a pristine reference, and instead, use additional (side) information along with the distorted signal. Reduced-Reference QA algorithms use features from the reference signal that are of lower bandwidth than using the entire reference signal to aid the QA task. No-Reference QA algorithms attempt to assess signal quality without using any other information than the distorted signal. This process has

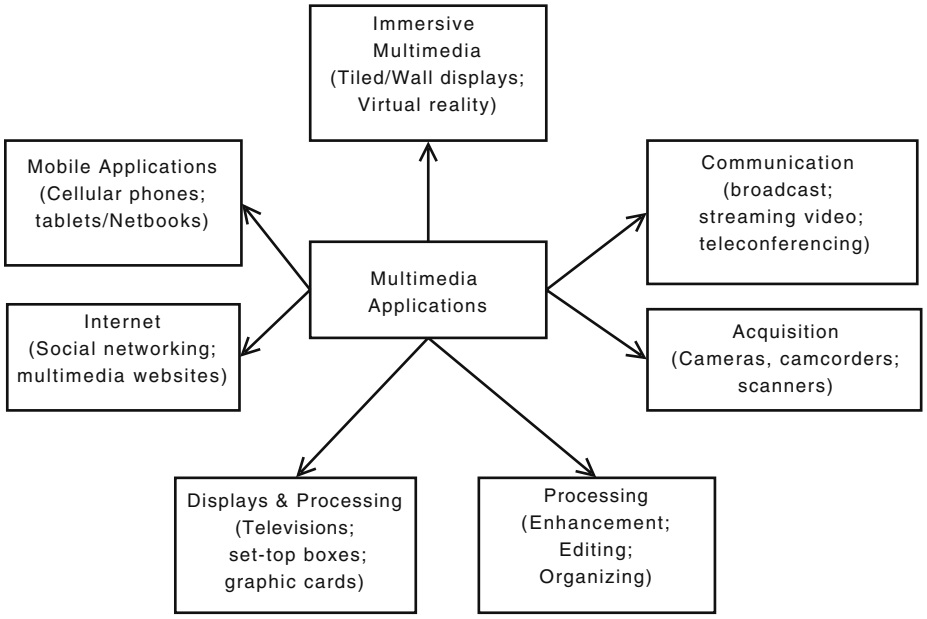


Fig. 1 Application realm of multimedia QA

proved daunting and there is little substantive work on this topic. Most of the work on No-Reference QA so far relies on the use of prior knowledge of the distortion process, such as degradation from compression that creates characteristic artifacts such as blocking, blurring, or ringing, to develop an algorithm. This approach does not generalize across different distortion types. Yet, humans perform the task almost instantaneously, which suggests that there is hope in this direction. It is our view that Full-Reference QA algorithms have reached a degree of maturity that make them suitable for widespread use and deployment, while much needs to be learned about human perception of quality before generic No-Reference QA algorithms reach desired levels of performance. In this survey, we will focus on Full-Reference QA algorithms for multimedia, which is an area that is of considerable practical significance. Full-Reference QA algorithms often require a registration stage to align the reference and test signals prior to QA. Registration methods are beyond the scope of this paper.

While a survey of QA methods can be found in [92], that reference studies methods developed in the 90's when interest in QA was primarily driven by display applications. A discussion of QA methods can also be found in several recent books [63, 65, 85, 95]. The goal of this survey is to provide the reader with a comprehensive view of current progress in QA methods in the context of multimedia applications. We review methods of objective QA of audio, image, video and audio-visual multimedia signals in Sections 2, 3, 4 and 5 respectively. We discuss benchmarking of objective QA methods, publicly available databases for researchers

and the performance of current methods in matching human perception in Section 6. We conclude this survey in Section 7.

2 Auditory perception and quality

Computational methods of assessing the quality of digital speech and audio signals have been studied since the 1970's in the context of telephone applications and remain important with the popularity of newer multimedia applications and Voice over IP (VoIP). Many of the approaches that have been proposed for full reference audio QA bear similarities to each other in that they use a psycho-acoustic model of auditory perception, which is applied to the reference and test audio signals whose quality difference is to be evaluated. The psycho-acoustic model accounts for various stages of processing that occur in the peripheral, middle and inner ear. These models are often constructed using psycho-acoustic experiments that study auditory responses of humans to carefully designed stimuli. Ground truth data from such experiments is used to construct a computational model of hearing sensitivity to the input signal that is often closely related to neuro-scientific models of how the human ear functions. The output of the psycho-acoustic model is considered representative of the neural activity patterns at the output of the auditory system, which is relayed by the vestibulocochlear nerve to the human brain via intermediate points along the auditory pathway. The outputs of the psycho-acoustic model are then further processed using cognitive models that account for higher stages of auditory processing in the human brain which are less understood. The outputs of the cognitive models are then compared using different metrics to predict the perceptual quality of the test signal with respect to the reference. Some of the early work on audio quality prediction focused on narrow-band speech signals and distortions introduced by speech coders, while later work has focused on the more difficult subject of QA for wide-band audio signals such as music.

A block diagram of a psycho-acoustically based audio QA algorithm is shown in Fig. 2 and this block diagram is intentionally simplified to illustrate the close similarities to psychovisually-based image QA algorithms discussed in Section 3. Typically, processing of an audio signal is performed in small intervals of time and a time interval of 20 ms duration has been often been used [60]. The transfer function of the outer and middle ear is modeled to account for differences in hearing sensitivities as a function of frequency. Sensitivity of hearing peaks in the region of 3 KHz and reduces for frequencies above and below this range, resulting in a bandpass characteristic [43, 72]. While bandpass models of the transfer function of the ear are an important component of most wideband audio QA systems, this step is often approximated using a lowpass filter that bandlimits the signal to less than 5 KHz in QA of narrowband speech signals [60]. The reference and test signals filtered using the transfer function is considered representative of the signal reaching the inner ear or cochlea, where a time-frequency transformation occurs. This is modeled by passing the audio signal through a bank of bandpass filters whose frequency response is designed to match the frequency analysis that the inner ear performs. The frequency analysis is often modeled using the Bark scale (or modifications of it) specified by Zwicker that has 24 bandpass filters, with center frequencies in the range of 50 Hz to 13.5 KHz, with better frequency resolution at lower frequencies

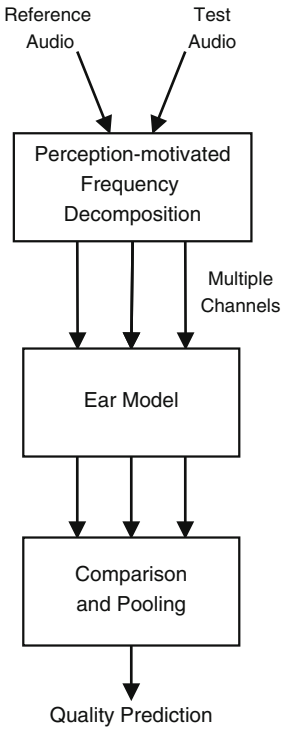


Fig. 2 Block diagram representing an audio quality metric

and increasing bandwidths at higher frequencies [97]. Other frequency analysis approaches, such as the short term FFT, have also been used for computational efficiency, while trading off accuracy in matching human auditory perception [56]. The output of the frequency analysis stage is passed through further stages that model downstream processing that occurs in the auditory system such as conversion to a loudness scale [60, 98]. One of the important and lesser understood components of downstream processing is masking, which refers to the reduction in loudness of a signal due to the presence of a second stronger signal. Masking is very important in QA since the same amount of noise or distortion in an audio signal may be masked to different degrees by the audio signal that carries the distortion, thereby modifying the degree of annoyance to the human listener. Several masking models have been proposed in the literature. The model often takes the form of a divisive normalization of the energy/intensity of the test or noise signal by the reference audio signal. A detailed discussion of the various masking models is beyond the scope of this paper, but can be found in [56, 60, 76]. The quality of the test signal with respect to the reference is then calculated using different metrics such as the noise to signal loudness ratio [60] or a combination of multiple features computed at the output of the psycho-acoustic model [56, 76]. Features that have been proposed for use in audio QA include noise-to-mask ratio, signal bandwidth, detection probability, perceived loudness and so on [76].

Several algorithms have been proposed for the quality assessment of speech and wideband audio signals using the psycho-acoustic modeling framework presented here [3, 5, 10, 27, 29, 34, 49, 55, 70, 75, 83]. The pioneering work of Schroeder et al.

is particularly notable [60], in addition to early and important work by Karjalainen and Brandenburg which has had an impact on this field [5, 34]. The International Telecommunications Union (ITU) has adopted two standards for measurement of full reference audio quality: the Perceptual Evaluation of Speech Quality (PESQ) for speech signals as recommendation ITU-T Rec. P.862 adopted in 2001 [50] and the Perceptual Audio Quality Evaluation (PEAQ) for wideband audio as recommendation ITU-R Rec. BS.1387 adopted in 1999 [44]. The PESQ and PEAQ algorithms are the result of consolidation of much of the early work on audio quality using the psycho-acoustic modeling framework . The PEAQ was developed as a collaboration of six audio QA algorithms [3, 10, 27, 49, 70, 75], which in turn built upon other early work [5, 34, 60, 98].

Other algorithms that do not utilize explicit modeling of the auditory processing in humans have also been proposed [11, 33]. These QA algorithms take an engineering approach to the problem and define features or mathematical entities that correlate with loss of quality, without incurring the complexity and computation of modeling the entire auditory perception pathway in humans. As an example, the Structural SIMilarity (SSIM) index that was originally developed for still images and described in Section 3 has been applied in audio quality evaluation with suitable modifications and has been demonstrated to correlate quite closely with subjective judgments [12, 33]. The Energy Equalization Quality Metric (EEQM) compares the spectrograms of the original and test audio signals to develop a quality metric [11]. It is worthwhile to note that such mathematically-based QA algorithms also utilize several components or concepts that are psycho-acoustically based, such as time-frequency decompositions or masking models, and there is no clear line of distinction between psycho-acoustic QA algorithms and mathematically-based QA algorithms.

3 Visual perception and still image quality

A substantial body of literature on image QA considers the psychophysics of human vision in constructing a quality index and are similar in vein to audio QA algorithms discussed in Section 2. Indeed, visual and auditory processing in the human being and psycho-physical modeling of the vision and hearing pathway share some remarkable and important commonalities, such as the bandpass nature of the sensitivity of the auditory and visual pathways to frequency; time-frequency decomposition of audio signals in hearing and space-frequency decomposition of still image signals in vision; increased frequency resolution of the decomposition at lower frequencies; increasing bandwidths at higher frequencies in the decomposition; masking effects and much more. As with audio, low level processing of the visual input is relatively well understood with established computational models of lower level processing in the retina and early stages of the visual cortex in the human brain. Higher level processing of these inputs in latter stages of the visual cortex, the extra-striate cortex and beyond in the visual pathway remains an active area of research. The increased dimensionality and the sheer number of neurons involved in processing the visual input makes vision modeling a much more complex problem than modeling of hearing.

The approach taken by most psycho-visual based models is to determine how the lower level physiology of the visual system limits visual sensitivity. Lower order processing occurs in the optics, retina, lateral geniculate nucleus, and striate cortex of the visual system [82]. Higher level processing, such as recognition and segmentation are, either too local in their effect, or not understood well enough to be effectively utilized. However, there have been attempts to model higher level motion processing in the extra-striate cortex in video QA, which we discuss in Section 4. In general, psycho-visual modeling based QA systems incorporate modeling of three types of processes that introduce sensitivity variations: light level, spatial frequency and signal content.

Most psycho-physically based methods for image QA are similar in philosophy to psycho-acoustically based audio QA algorithms discussed in Section 2 and construct a computational model of the response or sensitivity of the visual system as a function of the stimulus. Stimuli that best enable study of visual sensitivity to the stimulus characteristic of interest are carefully designed and displayed to human observers in psychophysical experiments. Thresholds of visibility are often measured for the stimuli to study various properties of vision and ground truth data from these studies are used to propose computational models that predict the observed responses as a function of the stimulus. Such models are often very closely related to computational models of how the human vision system processes the visual input, which is a wide open area of research in the field of neuro-science and visual psychology. A QA model is then constructed by passing the reference and test images through such a cognitive model to obtain a perceptually meaningful measure of quality. An extremely simplified block diagram of a generic psychovisually-based image QA system is illustrated in Fig. 5. Note the close similarity between Figs. 5 and 2. Many psycho-visually based methods incorporate elaborate models for calibration of the signal for viewing distances and display devices [13, 39], which are beyond the scope of this paper.

Most image QA algorithms include a frequency analysis stage that decomposes the reference and test images into different channels (usually called subbands) tuned to different spatial frequencies and orientations using a set of linear filters. This stage is intended to mimic similar processing that occurs in the human vision system: neurons in the visual cortex respond selectively to stimuli with particular spatial frequencies and orientations [82]. Different decompositions have been used in the literature including the Gabor decomposition, Cortex transform, steerable pyramid, wavelet transform and so on [15, 16, 69, 91]. While certain decompositions such as the Gabor and Cortex transforms are perceptually motivated, certain other transforms such as the wavelet transform are chosen for reasons of computational efficiency. Psycho-visually based quality metrics then model different properties of low level vision such as Weber's law or luminance masking, contrast masking and contrast sensitivity. A well known law governing perception known as the Weber-Fechner law stipulates that over a large dynamic range, and for many parameters, the threshold of discrimination between two stimuli increases linearly with stimulus intensity. This law was discovered by Ernst Weber in the 19th century and later, Gustav Fechner showed how Weber's law could be accounted for by postulating that the external stimulus is scaled into a logarithmic internal representation of sensation [18]. The sensitivity of the human eye for sinusoidal illuminance changes as a function of spatial frequency was systematically studied in [81] and it was shown that human perception of brightness follows Weber's law over a broad range of stimulus strength.

Interestingly, the Weber-Fechner law applies not only in vision, but also in the perception of sound and in particular, the perception of sound intensity, pitch and musical tempo has been found to follow the Weber's law. Spatial contrast sensitivity of vision refers to the differences in sensitivities to stimuli of varying spatial frequencies but with equal strength. The spatial contrast sensitivity function shows a bandpass shape with reduced sensitivity to low and high spatial frequencies [57, 59]. Contrast masking refers to the reduction in visibility of one signal component due to the presence of a similar signal component. In vision, contrast masking often occurs due to the presence of a masking signal at adjacent spatial locations, spatial frequency or orientation. Contrast masking has been studied extensively in the literature and psychophysical studies of this phenomenon that have influenced QA include [22, 38, 47, 58, 92]. Similar to masking in audio signals, the masking model often takes the form of a divisive normalization of the energy/intensity of the test or noise signal by the reference image and predicts the level of distortion to which an image can be exposed before the alteration is apparent to a human observer.

Several image QA algorithms have been proposed using a combination of computational models, often in cascade, of one or more of the luminance masking, contrast sensitivity and contrast masking properties of human vision. These include the pioneering work of Mannos and Sakrison [41], the Emmy award winning Sarnoff JNDMetrix technology based on the Lubin model [39], the Visible Differences Predictor (VDP) [13], DCTune [92], Teo and Heeger model [73], Moving Pictures Quality Metric (MPQM) [78], the Perceptual Distortion Metric (PDM) that builds upon MPQM [94], a scalable wavelet-based index [42] and the Visual Signal-to-Noise Ratio (VSNR) [7].

Recent trends in image QA have also seen a shift toward mathematically-based QA algorithms that take an engineering approach to the problem—a trend seen in audio QA also as described in Section 2. One of the prominent algorithms utilizing this approach is the Structural SIMilarity (SSIM) framework for image QA [84, 86]. Despite the apparent simplicity of the SSIM index, it has been shown to correlate quite closely with subjective judgments of quality [86]. The simplicity of the SSIM index also makes it fast, efficient, easy to use as a quality indicator in a variety of applications, as well as in optimizing various image processing systems for visual quality. The SSIM index has achieved enormous popularity over the last few years, and is now part of publicly available software packages such as the Wang Image Viewer, the MSU Video Quality Measurement Tool, the open source x.264 implementation of H.264/AVC, the JM reference software implementation of H.264 (http://iphome.hhi.de/suehring/tml/JM(JVT-X072).pdf) and VideoClarity's ClearView. SSIM has been utilized in a number of applications (not limited to image processing) such as audio QA as described in Section 2, image fusion, content retrieval/indexing, image/video compression, watermarking, denoising, chromatic image quality, retinal and see-through wearable displays, video hashing, wireless video, visual surveillance, radar imaging, digital camera design, infrared imaging, MRI imaging, remote sensing, target recognition, chromosome imaging, and industrial control.

SSIM hypothesizes that the visual quality of an image is related to the amount of structural information that the visual system can extract from it and attempts to avoid explicit modeling of visual processing. SSIM is computed locally between

corresponding patches from the reference and test images and the SSIM index between image patches $\tilde{\mathbf{f}}$ and $\tilde{\mathbf{g}}$ is defined as the product of three components:

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

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

$l(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ is a luminance comparison function:

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

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

$c(\tilde{\mathbf{f}}, \tilde{\mathbf{g}})$ is a contrast comparison function:

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

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

$s[\tilde{\mathbf{f}}, \tilde{\mathbf{g}}]$ is a structure comparison function:

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

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

The luminance and contrast comparison terms of the SSIM index account for variations in visual quality due to lighting changes in the test image with respect to the reference such as brightening, darkening, contrast enhancement etc. The key component of the SSIM index is the structure term that responds to distortions that alter the structure of the image patch and is quantified using the normalized cross correlation between the image patches. Although the SSIM index is defined by three terms, the structure term in the SSIM index is generally regarded as the most important, since variations in luminance and contrast of an image do not affect visual quality as much as structural distortions [84].

The SSIM index is computed locally at each pixel of the image and can be visualized as an image, often referred to as a SSIM map, which provides useful information on the localization of distortions. Examples of SSIM maps are shown in Fig. 3. This SSIM map is then pooled to obtain a single quality score for the entire image. The original paper on SSIM used the mean of the SSIM map to compute the overall SSIM index [86]. Several extensions of the SSIM index have also been proposed—most notably, the Multi-Scale SSIM (MS-SSIM) index that improves upon the SSIM index by decomposing the images into multiple scales before QA

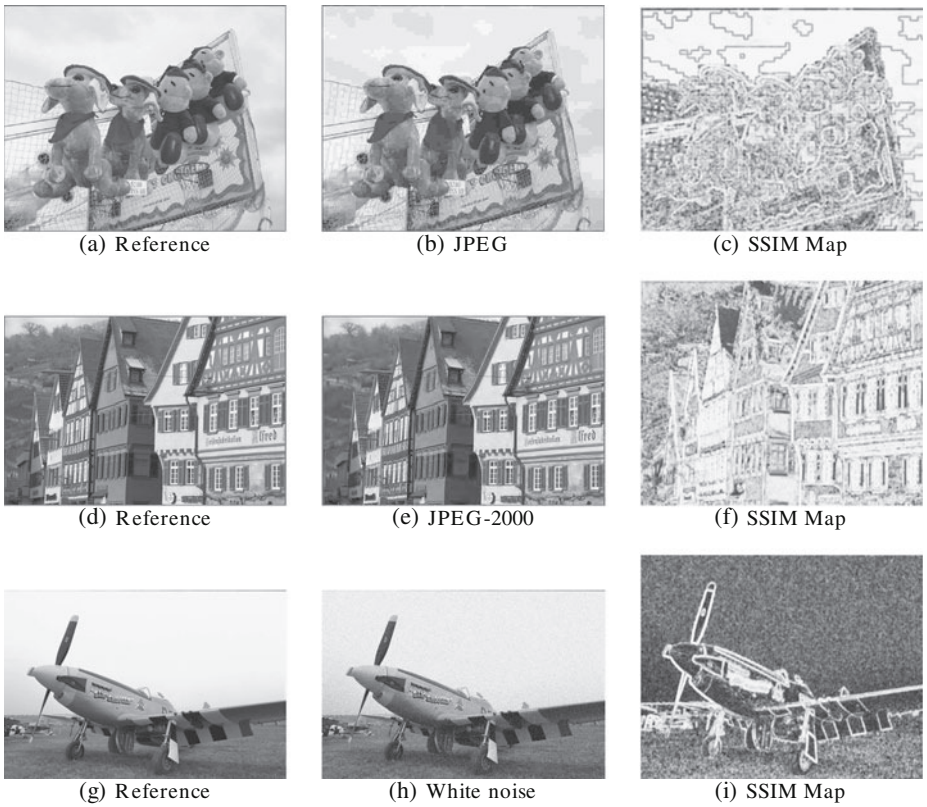


Fig. 3 Illustration of SSIM Maps. *Left column* shows reference images. *Middle column* shows distorted images obtained from the reference using JPEG compression, JPEG2000 compression and additive white Gaussian noise. *Right column* shows the SSIM index at each pixel displayed as an image. Bright regions correspond to better quality and dark regions correspond to worse quality. *SSIM Maps* clearly display the regions of the distorted image that are visually annoying to the human observer

[89]. Other extensions to SSIM incorporate color information [77], rotation and translation invariance [90] and so on.

Another recent IQA algorithm that has been shown to perform quite well in matching visual perception is known as the Visual Information Fidelity (VIF) criterion [68]. VIF uses information theoretic principles in defining the quality index. An image source communicates to a receiver through a channel that limits the amount of information that could flow through it, thereby introducing distortions. The output of the image source is the reference image and the output of the channel is the test image, as shown in Fig. 4. VIF utilizes two aspects of image information for quantifying perceptual quality: the information shared between the test and the reference image, and the information content of the reference image itself. Statistical models for signal sources and transmission channels are at the core of information fidelity methods, which attempt to exploit the relationship between statistical image information and visual quality. Recent work has shown important similarities and relationships between the VIF, SSIM and psycho-visually based indices for QA

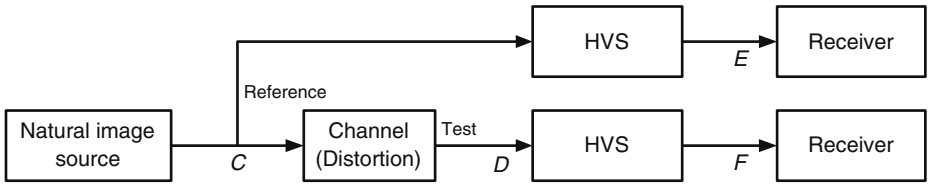


Fig. 4 An information-theoretic setup for quantifying visual quality using a distortion channel model. The visual system also acts as a channel that limits the flow of information from the source to the receiver. The VIF index is defined using a relative comparison of the information in the *upper path* of the figure and the information in the *lower path*

[62]. This re-emphasizes the notion that mathematically based QA algorithms are not that different from psycho-visually based QA algorithms and do incorporate modeling and concepts from psycho-physics, which is a necessary ingredient in achieving the end goal of predicting human responses. Other algorithms that take a mathematically-based approach to image QA include [1, 14, 79].

4 From images to motion pictures: video quality

Much of the research on video QA builds upon image QA algorithms with additional components to handle the temporal aspects of video. A psycho-visually based video QA system often consists of an entire psycho-visually based image QA system, with modifications to existing blocks or inclusion of additional blocks to account for the temporal dimension of video. For instance, video QA systems often utilize a temporal filtering stage in cascade with the spatial filtering that occurs in the "frequency decomposition" stage of image QA systems shown in Fig. 5. This is equivalent to filtering the videos using a spatio-temporal filterbank that is separable along the spatial and temporal dimensions. Temporal filtering typically models two kinds of temporal mechanisms that exist in the early stages of processing in the visual cortex that are often modeled using linear lowpass and bandpass filters applied along the temporal dimension of the videos [23]. Psycho-visually based video QA systems that utilize this approach include the Moving Pictures Quality Metric (MPQM) [78], the Perceptual Distortion Metric (PDM) [94], the Digital Video Quality (DVQ) metric [93] and a scalable wavelet based video distortion metric [42]. Typically, simple modifications are also made to the "Psychophysical Vision Model" block in Fig. 5 for video QA. For instance, video QA needs to account for the spatio-temporal contrast sensitivity function of human vision that measures the sensitivity of vision to different spatial and *temporal* frequencies of the stimulus. Spatio-temporal contrast sensitivity was first studied in early work on visual psychophysics [35]. A spatio-temporal model of contrast sensitivity was created using measurements of the contrast sensitivity function as a non-separable function of spatial and temporal frequencies using psychophysical experiments in [36]. Several psycho-visually based video QA systems have also been implemented in commercial products such as the Sarnoff JNDMetrix technology, the Picture Quality Analyzer (PQA) 200/500 systems from Tektronix (http://www.tek.com/products/video_test/pqa500/), the Cheetah V-Factor Quality of Experience (QoE) platform (formerly Symmetricom)

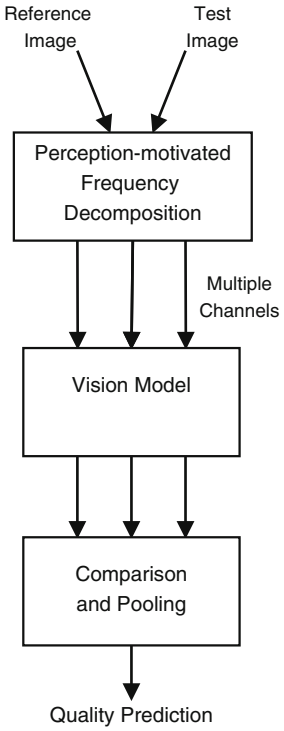


Fig. 5 Block diagram representing a psycho-visually based image quality metric

(http://www.cheetahtech.com/products/products-v-factor.html) and the Video Quality Analyzer (VQA) from AccepTV (http://www.acceptv.com/) [6].

Other mathematically based video QA algorithms utilize statistics and features computed from the reference and test videos to predict the visual quality of the test video. The Perceptual Evaluation of Video Quality (PEVQ) model from Opticom (http://www.opticom.de/technology/pevq_video-quality-testing.html) [2, 40] is based on an earlier model known as the Perceptual Video Quality Measure (PVQM) developed by Swisscom/KPN Research, Netherlands [26]. PVQM measures three different quantities from the reference and distorted videos to compute video quality—an edginess indicator, a temporal indicator and a chrominance indicator. These indicators are then combined to determine the overall perceived video quality. Another prominent video QA system was developed at the National Telecommunications and Information Administration (NTIA) and is known as the Video Quality Metric (VQM) or the NTIA General Model [51]. VQM and its associated calibration techniques have been adopted as a North American standard by the American National Standards Institute (ANSI) in 2003. The International Telecommunication Union (ITU) has also included VQM as a normative method for digital cable television systems [48, 74]. VQM contains seven parameters of which four are based on features extracted from spatial gradients of the luminance component of the video, two parameters are based on features extracted from the chrominance component, and one parameter is based on the product of features that measure contrast and temporal information (both of which are extracted from the

luminance component of the video). The SSIM index for still images has also been extended for video QA by applying it frame-by-frame on the video sequence and using motion vectors computed from the reference video to pool local SSIM indices into a single video quality score [87, 88].

The human eye is very sensitive to motion and can accurately judge the velocity and direction of motion of objects in a scene. This does not seem surprising in view of the fact that the ability to detect motion accurately is crucial to survival and performance of tasks such as navigating through the environment, avoiding danger and so on. Motion attracts visual attention and affects spatio-temporal aspects of human vision. These properties of vision are a consequence of processing in the visual system and in particular, Area MT/V5 of the extrastriate cortex, which plays an important role in motion perception [82]. Visual data, after lower level processing, is transported from the visual cortex along the ventral stream to Area MT/V5. Area MT/V5 appears to integrate local motion information computed in the cortex into global percepts of motion of complex patterns that typically occur in video sequences [46]. Area MT/V5 also plays a role in the guidance of some eye movements, segmentation and structure computation in 3-dimensional space [4]—properties of human vision that play an important role in visual perception of videos. The video QA algorithms discussed so far focus largely on spatial aspects of video and temporal processing in the visual cortex, which represents early stages of motion processing in the vision system. It is our view that their performance can be improved by modeling downstream processing in Area MT/V5 that plays an important role in motion perception in human vision and hence, visual perception of videos.

Towards this end, we have recently developed a framework for evaluating spatial and temporal (and spatio-temporal) aspects of distortions in video [61, 64], based on which an algorithm known as the MOtion based Video Integrity Evaluation or MOVIE index was defined. In this framework, video quality is evaluated not only in space and time, but also in space-time, by evaluating motion quality along computed motion trajectories. It is our view that using motion models in video QA represents a significant step forward in reaching the ultimate goal of matching human perception of videos. In our motion-based framework for VQA, separate components for spatial and temporal quality are defined [64]. First, the reference and test videos are decomposed into spatio-temporal bandpass channels using a Gabor filter family. Spatial quality measurement is accomplished by computing an error index between the bandpass reference and distorted Gabor channels using models of the contrast masking property of visual perception. This results in local estimates of spatial quality at each pixel of the test video which is known as the Spatial MOVIE map. Temporal quality is measured using optical flow fields computed from the reference video using our own multi-scale extension of the Fleet and Jepson phase-based optical flow estimation technique [21, 64]. To compute temporal quality, MOVIE computes *reference motion-tuned* responses from both the reference and distorted videos using a weighted sum of the Gabor outputs. The weights are designed based on the direction and speed of motion in the reference video such that the weighted sum responds strongly to similar motion trajectories in the test video as compared to the reference, with reduced responses whenever the test video motion trajectory deviates from the reference video motion. This computation in Temporal MOVIE bears similarities with computational models that have been proposed for the response of neurons in Area MT [69] and is perceptually motivated. Local estimates of Temporal MOVIE computed per pixel in the test video also results in a map known as the

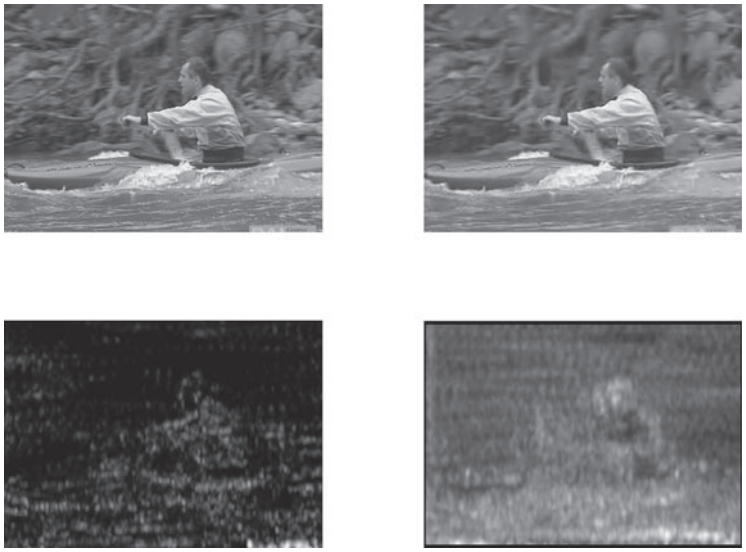


Fig. 6 Illustration of the performance of the MOVIE index. *Top left* shows a frame from the reference video. *Top right* shows the corresponding frame from the distorted video. *Bottom left* shows the Temporal MOVIE map that is logarithmically compressed for visibility. Note that only the central frame is shown in the *top row* for visualization purposes and that MOVIE computation uses several frames before and after this central frame. *Bottom right* shows the Spatial MOVIE map. Bright regions correspond to regions of poor quality predicted by MOVIE. Notice that the Spatial MOVIE map responds to the blur in the test video. The Temporal MOVIE map responds to motion compensation mismatches surrounding the man, the oar and the ripples in the water

Temporal MOVIE map. Figure 6 illustrates quality maps generated by MOVIE on a representative video sequence. First of all, it is evident that the kind of distortions captured by the spatial and temporal maps is different. The test video sequences in both examples suffer from significant blurring and the spatial quality map clearly reflects the loss of quality due to blur. The temporal quality map, however, shows poor quality along the edges of objects and in the water where motion compensation mismatches are evident. Of course, the spatial and temporal quality values are not completely independent.

Finally, the spatial and temporal quality scores are pooled to produce an overall video integrity score known as the MOVIE index. Temporal quality computation in MOVIE is based upon computational models of neurons in Area MT that play a critical role in motion perception and is capable of capturing the many motion-related distortions that occur in video. MOVIE has been shown to match human visual perception of video quality quite closely [64].

5 Combining sensory signals: audiovisual quality evaluation

Multimedia QA is the end goal of most communication systems and determines the overall Quality of Experience (QoE) derived by the end user of the system.

Multimedia QA examines the overall QoE, as opposed to the QA of individual components of the multimedia signals such as audio quality or video quality.

One critical component of assessing multimedia quality is the quality of synchronization between the individual media components of the multimedia signal. In particular, lip synchronization or the synchronization between the audio and video components of the multimedia stream has been extensively studied in the literature. Quality of lip synchronization is sometimes viewed as an alignment problem, which is a necessary step in most QA systems. However, alignment for synchronization is performed or assessed between the different streams in the multimedia signal which requires very different methods than the alignment that is performed between the reference and test signals in a single medium QA system. The degree of lack of synchronization between audio and video streams tolerated by humans has been studied [19, 54, 71]. It has been found that humans can tolerate about 80 ms delay between the audio and video signals without finding it disturbing, while a delay of about 160 ms can lead to considerable annoyance and loss of quality in conversational applications. The perception of asynchrony is not symmetric and people are more tolerant of video ahead of audio than when the audio stream leads the video stream [54, 71]. It has been hypothesized that this could be due to the fact that audio lagging video is common since light travels faster than sound [71].

Relations between the subjective quality of audio, video and multimedia signals have been studied and it has been found that audio and video qualities have a mutual influence. Changes in audio quality cause a change in perceived video quality and vice versa when humans are asked to evaluate the quality of individual component signals, as opposed to a presentation of a multimedia signal. The overall quality of multimedia as a function of the qualities of the component signals has been studied and it has been found that the overall quality can be predicted well using a function of the individual qualities of the audio and video signals [3, 25]. It has been found that video quality dominates audio quality in a number of situations such as high motion video, while audio quality dominates overall quality for specific stimuli such as "talking head" videos. Overall multimedia quality of an audiovisual stream has often been predicted using the following bilinear model.

$$\text{Multimedia quality} = \alpha \times \text{Audio quality} + \beta \times \text{Video quality} \tag{1}$$

$$+ \gamma \times \text{Audio quality} \times \text{Video quality} + \delta \tag{2}$$

It has been found that the multiplicative term, in particular, contributes significantly to predicting the overall multimedia quality [3, 25]. Although the end goal in most applications is the prediction of the overall Quality of Experience (QoE) derived by the end-user, the reliability of predicting overall audio-visual quality as a function of the quality of individual component media has resulted in the bulk of the work being performed on the subproblems of audio and video QA. While this is true of audiovisual quality, it is conceivable that future multimedia experiences that are not simply audio-visual, but include other media components, will require more extensive studies of the interaction between quality perceptions of different media components.

6 Benchmarking and public databases

The goal of any QA algorithm or system is to predict the perceptual quality derived by a user consuming a media signal and studying human responses to media quality is the only way to obtain ground truth data to assess the performance of a QA algorithm. Obtaining quality judgments of multimedia signals from humans is often referred to as subjective QA. Subjective QA is the only reliable means of assessing human quality judgments and continues to be used as the ultimate standard of performance of a multimedia communication system. However, subjective studies can be quite cumbersome and expensive due to human involvement in the process, which greatly limit the number of videos that can be accommodated in such an evaluation. However, subjective studies have great relevance in providing ground truth human data that enable benchmarking of objective QA algorithms. Subjective studies of video quality enable us to understand the level of maturity that QA algorithms have achieved and to understand how close we are to achieving the goal of correlating perfectly with human judgments and eliminating subjective studies completely. It is our view that we are a long way from achieving this goal, given the complexity of human visual processing and the challenges in understanding human perception of the enormous variety of multimedia (stereoscopic 3D, for example) that is constantly evolving.

Full reference algorithms are often benchmarked using subjective studies conducted in a double stimulus paradigm. Double stimulus studies present the reference and test signals to the subject, who is asked to rate his or her quality preference for each. Difference between scores assigned by the subject is indicative of the quality of the test signal with respect to the reference, which is considered "perfect" quality in full reference QA, thus accounting for any preferences subjects might have for the reference content. Since double stimulus studies can be quite time consuming due to the presentation of the reference signal along with each test signal, single stimulus studies obtain quality scores from a subject based on the presentation of the test signal alone. Often, subjective preferences for reference content is accounted for by including the reference signals in the study and obtaining quality scores for these to use in a subtractive manner—a procedure known as hidden reference removal.

Subjective quality scores can be obtained in a number of ways. A popular scale is known as the Absolute Category Rating (ACR) scale that obtains quality scores from a subject along a scale that consists of a fixed number of categories. A 5-point ACR scale is commonly used in audio, image, video and multimedia quality assessment [25, 31, 32] and is also recommended as part of several ITU standards for subjective quality assessment including ITU-T Recommendation P.800 for audio [32] and ITU-R Recommendation BT 500.11 for television [31]. The 5-point ACR scale consists of five labels "Bad", "Poor", "Fair", "Good" and "Excellent" corresponding to quality scores ranging from 1 to 5. Quality scales with finer resolution such as a sliding bar scale have also been used in subjective studies, since they allow for finer and better discrimination and statistical analysis of human preferences [20, 66, 67].

Typically, a large number of subjects are recruited in a subjective study to account for variations between subjects in the quality assessment task. The data obtained from individual subjects is then processed using multiple means to determine a Mean Opinion Score (MOS) for the audio/image/video/multimedia signal. Variation between subjects in assessing quality is also captured to characterize statistical limits

on agreement between subjects, which in turn determines the level of performance that an objective QA algorithm is expected to achieve since it does not aim to predict human variation. Multiple methods have been proposed to process subjective data obtained from individual human subjects [20, 31, 32, 44, 66, 68, 74, 80] and are beyond the scope of this paper. The results of a subjective study for full reference QA typically consists of reference signals, test signals, MOS scores that are obtained by processing the raw subjective scores and often, the variance in the MOS scores that captures inter-subject variability. Different performance metrics such as the Spearman Rank Order Correlation Coefficient (SROCC), Pearson Linear Correlation Coefficient (LCC) and Root Mean Square Error (RMSE) are then computed between algorithmic predictions of quality and the MOS scores from the subjective study to evaluate the performance of individual QA algorithms [20, 66, 67, 74]. Statistical analysis of the data to establish that the performance of one algorithm is statistically significantly better or worse than others is also performed to ensure that the differences between the performance of different algorithms is not within the scope of inter-subject variability [66, 67, 74].

Audio QA algorithms have often been benchmarked using the databased created by the ITU that are described in detail in [44]. PEAQ was standardized based on the good performance of this algorithm in predicting the subjective scores in this dataset. Several benchmarking studies of state-of-the-art image and video QA algorithms have been conducted. The LIVE image quality assessment database is publicly available and contains over 700 images suffering from multiple distortion types (JPEG and JPEG 2000 compression, Gaussian blur, additive white Gaussian noise, simulated errors of JPEG2000 bitstreams over lossy wireless networks), with associated MOS scores obtained through a large scale subjective study (http://live.ece.utexas.edu/research/quality/subjective.htm). The LIVE image quality database was used to study the performance of a number of publicly available image QA algorithms in terms of correlation coefficients and statistical significance of performance [67]. The VIF and SSIM indices emerged as the leading algorithms in matching human perception in this study and were shown to clearly outperform PSNR as a quality metric. The LIVE image quality assessment database has since emerged as a de-facto standard and is widely used in the literature to evaluate the performance of newer algorithms. Another publicly available image quality database is known as the Tampere Image Database (TID) [52] and contains all the distortion types in the LIVE image quality database, in addition to other distortion types such as different types of noise. Performance of several image QA algorithms were also tested on the TID database and the Multi-Scale SSIM index (MS-SSIM), SSIM and VIF indices emerged as the leading algorithms in this study as well [52].

One of the oldest and widely used databases for video QA is the publicly available VQEG FRTV Phase-I database as part of its FR-TV Phase 1 project in 2000 [20]. However, the VQEG database suffers from a number of drawbacks. There have been significant advances in video processing technology since 2000 and the test videos in the VQEG study are not representative of present generation encoders and communication systems. The VQEG study targeted secondary distribution of television and contains interlace videos. Interlaced videos are not typical of present generation applications such as multimedia, IPTV, video viewing on computer monitors, progressive High Definition Television (HDTV) standards and so on. Further, the VQEG database was designed to address the needs of secondary distribution of

television and hence, the database spans narrow ranges of quality scores—indeed, more than half of the sequences are of very high quality (MPEG-2 encoded at >3Mbps). Overall, the VQEG videos exhibit poor perceptual separation, making it difficult to distinguish the performance of VQA algorithms. More recently, the LIVE Video Quality Database was developed to overcome these limitations [66]. The LIVE Video Quality Database includes videos distorted by compression using more recent and advanced codecs such as H.264/AVC, as well as videos resulting from simulated transmission of H.264 packetized streams through error prone communication channels. Videos in the LIVE Video Quality Database are all captured in progressive scan formats. Further, the LIVE Video Quality Database spans a much wider range of quality—the low quality videos are designed to be of similar quality found in streaming video applications on the Internet (Youtube, wireless videos, live streaming of low bandwidth videos, etc.). A study of the performance of several state-of-the-art video QA algorithms was undertaken in [66]. The study established MOVIE as the leading algorithm for video QA and found that the performance of the MS-SSIM index and VQM from NTIA was competitive. Performance evaluation contests have also been performed as part of the ITU-T standardization process for video QA. However, the subjective data, video and results of these studies are often not released publicly, which makes it difficult to benchmark publicly available algorithms against the results of the ITU evaluations on proprietary algorithms from the industry. VQM from NTIA is a notable exception and is one of the few publicly available algorithms from the industry [51]. Other video QA databases that consist of application specific distortion types include [17] for IP video transmission and [45] for wireless video transmission.

Audiovisual QA studies obtain subjective scores from subjects experiencing a presentation of a synchronized audiovisual stream. Studies such as [3, 25] also study the subjective data from individual presentations of the audio and video substreams separately to study the interactions between perception of audio, video and audiovisual data.

7 Conclusions

This paper presented a survey of multimedia quality assessment with a focus on full reference methods for QA. We discussed audio, image, video and multimedia QA and attempted to bring out the similarities between models that have been proposed in each of these realms, and at the same time, highlight important differences that need to be accounted for in modeling human perception of these different media components. The similarities between perception-based approaches to image, audio and video QA are quite remarkable and the similarities extend down to individual components of these models (such as frequency sensitivity and masking) and computational models for these components. Indeed, given the similarities between audio and video QA, many of the companies such as British Telecom, Opticom, Swisscom/KPN Research etc. that develop tools for QA and participate in ITU standardization of QA methods are involved in the development of both audio and video QA methods.

While there has been significant progress in the areas of audio, image and video QA, much more work needs to be done to increase our understanding of the

many complexities of human perception. Recent work studies quality assessment of multi-channel audio signals [24, 96]. Another interesting area of research that has seen a lot of recent interest is the study of visual saliency and utilizing saliency models in improving image and video QA [30, 37, 53]. Rapid proliferation of devices enabled to display stereoscopic 3D videos has seen some nascent research in QA for this medium [28]. Applications of QA methods in the design of image and video processing systems is another nascent area of research that can help optimize video processing systems for perceptual quality [8, 9]. Finally, no-reference QA remains a wide open and much needed tool and while progress in this area remains limited and application-specific, the knowledge that humans can perform this task almost instantaneously gives us hope in reaching this objective in the future.

References

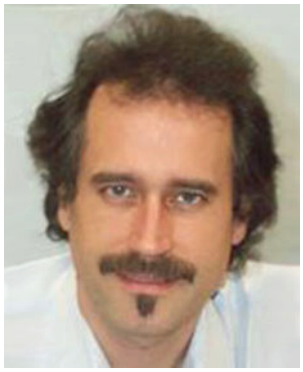
1. Avcibas I, Sankur B, Sayood K (2002) Statistical evaluation of image quality measures. J Electron Imaging 11(2):206–223
2. Barkowsky M, Bialkowski J, Bitto R, Kaup A (2007) Temporal registration using 3D phase correlation and a maximum likelihood approach in the perceptual evaluation of video quality. In: IEEE workshop on multimedia signal proc
3. Beerends JG, Stemerdink JA (1992) A perceptual audio quality measure based on a psychoacoustic sound representation. J Audio Eng Soc 40(12):963–978
4. Born RT, Bradley DC (2005) Structure and function of visual area MT. Annu Rev Neurosci 28:157–189
5. Brandenburg T, Sporer K (1992) NMR and masking flag: evaluation of quality using perceptual criteria. In: Audio engineering society conference: 11th international conference: test & measurement
6. Carnec M, Le Callet P, Barba D (2008) Objective quality assessment of color images based on a generic perceptual reduced reference. Signal Process Image Commun 23(4):239–256
7. Chandler DM, Hemami SS (2007) VSNR: a wavelet-based visual signal-to-noise ratio for natural images. IEEE Trans Image Process 16(9):2284–2298
8. Channappayya SS, Bovik AC, Caramanis C, Heath RW Jr (2008) Design of linear equalizers optimized for the structural similarity index. IEEE Trans Image Process 17(6):857–872
9. Channappayya SS, Bovik AC, Heath RW Jr (2008) Rate bounds on SSIM index of quantized images. IEEE Trans Image Process 17(9):1624–1639
10. Colomes C, Lever M, Rault J-B, Dehery Y-F, Faucon G (1995) A perceptual model applied to audio bit-rate reduction. J Audio Eng Soc 43(4):233–240
11. Creusere C (2003) Quantifying perceptual distortion in scalably compressed mpeg audio. In: Conference record of the thirty-seventh asilomar conference on signals, systems and computers, vol 1, pp 265–269
12. Creusere C, Hardin J (2010) Assessing the quality of audio containing temporally varying distortions. IEEE Trans Speech Audio Lang Process PP(99):1–1
13. Daly S (1993) The visible difference predictor: An algorithm for the assessment of image fidelity. In: Watson AB (ed) Digital images and human vision. The MIT, pp 176–206
14. Damera-Venkata N, Kite T, Geisler W, Evans B, Bovik A (2000) Image quality assessment based on a degradation model. IEEE Trans Image Process 9(4):636–650
15. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. Commun Pure Appl Math 41(7):909–996
16. Daugman JG (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. J Opt Soc Am A (Opt Image Sci) 2(7):1160–1169
17. De Simone F, Naccari M, Tagliasacchi M, Dufaux F, Tubaro S, Ebrahimi T (2009) Subjective assessment of H.264/AVC video sequences transmitted over a noisy channel. In: International workshop on quality of multimedia experience, pp 204–209
18. Dehaene S (2003) The neural basis of the weber-fechner law: a logarithmic mental number line. Trends Cogn Sci 7(4):145–147

19. Dixon NF, Spitz L (1980) The detection of auditory visual desynchrony. Perception 9(6): 719–721
20. Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment (2000) Available online: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI/COM-80E_final_report.pdf. Accessed June 2000
21. Fleet DJ, Jepson AD (1990) Computation of component image velocity from local phase information. Int J Comput Vis 5(1):77–104
22. Foley J (1994) Human luminance pattern-vision mechanisms: masking experiments require a new model. J Opt Soc Am A (Opt Image Sci) 11(6):1710–1719
23. Fredericksen RE, Hess RF (1997) Temporal detection in human vision: dependence on stimulus energy. J Opt Soc Am A (Opt Image Sci Vis) 14(10):2557–2569
24. George S, Zielinski S, Rumsey F (2006) Feature extraction for the prediction of multichannel spatial audio fidelity. IEEE Trans Speech Audio Lang Process 14(6):1994–2005
25. Hands DS (2004) A basic multimedia quality model. IEEE Trans Multimedia 6(6):806–816
26. Hekstra AP, Beerends JG, Ledermann D, de Caluwe FE, Kohler S, Koenen RH, Rihs S, Ehrsam M, Schlauss D (2002) PVQM—A perceptual video quality measure. Signal Process Image Commun 17:781–798
27. Herre J, Eberlein E, Schott H, Schmidmer C (1992) Analysis tool for realtime measurements using perceptual criteria. In: Audio engineering society conference: 11th international conference: test & measurement
28. Hewage CTER, Worrall ST, Dogan S, Kondoz AM (2008) Prediction of stereoscopic video quality using objective quality models of 2-d video. Electron Lett 44(16):963–965
29. Huber R, Kollmeier B (2006) PEMO-Q—A new method for objective audio quality assessment using a model of auditory perception. IEEE Trans Speech Audio Lang Process 14(6): 1902–1911
30. Itti L, Koch C (2001) Computational modelling of visual attention. Nat Rev Neurosci 2(3): 194–203
31. ITU-R Recommendation BT.500-11 (2000) Methodology for the subjective assessment of the quality of television pictures. International Telecommunications Union, Tech Rep
32. ITU-T Recommendation P.800 (1996) Methods for subjective determination of transmission quality. International Telecommunications Union, Tech Rep
33. Kandadai S, Hardin J, Creusere C (2008) Audio quality assessment using the mean structural similarity measure. In: IEEE international conference on acoustics, speech and signal processing, pp 221–224
34. Karjalainen M (1985) A new auditory model for the evaluation of sound quality of audio systems. In: IEEE international conference on acoustics, speech, and signal processing, vol 10, pp 608–611
35. Kelly DH (1984) Retinal inhomogeneity. i. spatiotemporal contrast sensitivity. J Opt Soc Am A 1(1):107–113
36. Lambrecht CJvdB, Kunt M (1998) Characterization of human visual sensitivity for video imaging applications. Signal Process 67(3):255–269
37. Le Meur O, Le Callet P, Barba D, Thoreau D (2006) A coherent computational approach to model bottom-up visual attention. IEEE Trans Pattern Anal Mach Intell 28(5):802–817
38. Legge GE, Foley JM (1980) Contrast masking in human vision. J Opt Soc Am 70(12): 1458–1471
39. Lubin J (1993) The use of psychophysical data and models in the analysis of display system performance. In: Watson AB (ed) Digital images and human vision. The MIT, pp 163–178
40. Malkowski M, Claben D (2008) Performance of video telephony services in UMTS using live measurements and network emulation. Wirel Pers Commun 1:19–32
41. Mannos J, Sakrison D (1974) The effects of a visual fidelity criterion of the encoding of images. IEEE Trans Inf Theory 20(4):525–536
42. Masry M, Hemami SS, Sermadevi Y (2006) A scalable wavelet-based video distortion metric and applications. IEEE Trans Circuits Syst Video Technol 16(2):260–273
43. Mehrgardt S, Mellert V (1977) Transformation characteristics of the external human ear. J Acoust Soc Am 61(6):1567–1576
44. Method for objective measurements of perceived audio quality. ITU Std. BS. 1387, 1999
45. Moorthy A, Seshadrinathan K, Soundararajan R, Bovik AC (2010) Wireless video quality assessment: a study of subjective scores and objective algorithms. IEEE Trans Circuits Syst Video Technol 20(4):587–599

46. Movshon JA, Newsome WT (1996) Visual response properties of striate cortical neurons projecting to Area MT in macaque monkeys. J Neurosci 16(23):7733–7741
47. Nachmias J, Sansbury RV (1974) Grating contrast: discrimination may be better than detection. Vis Res 14(10):1039–1042
48. Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference (2004) International Telecommunications Union Std. ITU-T Rec J 144
49. Paillard B, Mabilleau P, Morissette S, Soumagne J (1992) PERCEVAL: Perceptual evaluation of the quality of audio signals. J Audio Eng Soc 40(1/2):21–31
50. Perceptual evaluation of speech quality (PESQ): an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. International Telecommunications Union Std., p 862, 2001
51. Pinson MH, Wolf S (2004) A new standardized method for objectively measuring video quality. IEEE Trans Broadcast 50(3):312–322
52. Ponomarenko N, Lukin V, Zelensky A, Egiazarian K, Carli M, Battisti F (2009) TID2008—a database for evaluation of full-reference visual quality assessment metrics. Adv Modern Radio-Electronics 10:30–45
53. Rajashekar U, van der Linde I, Bovik AC, Cormack LK (2008) GAFFE: a gaze-attentive fixation finding engine. IEEE Trans Image Process 17(4):564–573
54. Rihs S (1995) The influence of audio on perceived picture quality and subjective audio-video delay tolerance. RACE MOSAIC deliverable R211 180CESR007.B1, Tech. Rep
55. Rix AW, Beerends JG, Kim D-S, Kroon P, Ghitza O (2006) Objective assessment of speech and audio quality—technology and applications. IEEE Trans Speech Audio Lang Process 14(6):1890–1901
56. Rix AW, Hollier MP, Hekstra AP, Beerends JG (2002) Perceptual evaluation of speech quality (PESQ): the new ITU standard for end-to-end speech quality assessment part I–time-delay compensation. J Audio Eng Soc 50(10):755–764
57. Robson JG (1966) Spatial and temporal contrast-sensitivity functions of the visual system. J Opt Soc Am 56(8):1141–1142
58. Ross J, Speed HD (1991) Contrast adaptation and contrast masking in human vision. Proc Biol Sci 246(1315):61–70
59. Schober HAW, Hilz R (1965) Contrast sensitivity of the human eye for square-wave gratings. J Opt Soc Am 55(9):1086–1090
60. Schroeder MR, Atal BS, Hall JL (1978) Optimizing digital speech coders by exploiting masking properties of the human ear. J Acoust Soc Am 64(S1):S139–S139
61. Seshadrinathan K, Bovik AC (2007) A structural similarity metric for video based on motion models. In: IEEE intl. conf. on acoustics, speech, and signal proc
62. Seshadrinathan K, Bovik AC (2008) Unifying analysis of full reference image quality assessment. In: IEEE intl. conf. on image proc. San Diego, CA, pp 1200–1203
63. Seshadrinathan K, Bovik AC (2009) Video quality assessment. In: Bovik AC (ed) The essential guide to video processing, chapter 14. Academic, pp 417–436
64. Seshadrinathan K, Bovik AC (2010) Motion tuned spatio-temporal quality assessment of natural videos. IEEE Trans Image Process 19(2):335–350
65. Seshadrinathan K, Safranek RJ, Chen J, Pappas TN, Sheikh HR, Simoncelli EP, Wang Z, Bovik AC (2009) Image quality assessment. In: Bovik AC (ed) The essential guide to image processing, chapter 21. Academic, pp 553–596
66. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) Study of subjective and objective quality assessment of video. IEEE Trans Image Process 19(6):1427–1441
67. Sheikh HR, Bovik AC (2006) An evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process 15(11):3440–3451
68. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–444
69. Simoncelli EP, Heeger DJ (1998) A model of neuronal responses in visual area MT. Vis Res 38(5):743–761
70. Sporer T (1997) Objective audio signal evaluation-applied psychoacoustics for modeling the perceived quality of digital audio. In: Audio engineering society convention 103
71. Steinmetz R (1996) Human perception of jitter and media synchronization. IEEE J Sel Areas Commun 14(1):61–72
72. Terhardt E (1979) Calculating virtual pitch. Hear Res 1(2):155–182
73. Teo PC, Heeger DJ (1994) Perceptual image distortion. In: Proceedings of the IEEE international conference on image processing, vol 2. IEEE, pp 982–986

74. The Video Quality Experts Group (2003) Final VQEG report on the validation of objective models of video quality assessment. Available online: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII . Accessed 25 August 2003
75. Thiede E, Kabot T (1996) A new perceptual quality measure for bit-rate reduced audio. In: Audio engineering society convention 100
76. Thiede T, Treurniet WC, Bitto R, Schmidmer C, Sporer T, Beerends JG, Colomes C (2000) PEAQ—the ITU standard for objective measurement of perceived audio quality. J Audio Eng Soc 48(1/2):3–29
77. Toet A, Lucassen MP (2003) A new universal colour image fidelity metric. Displays 24(4–5):197–207
78. van den Branden Lambrecht CJ, Verscheure O (1996) Perceptual quality measure using a spatiotemporal model of the human visual system. In: Proc. SPIE, vol 2668, no. 1. SPIE, San Jose, pp 450–461
79. Van der Weken D, Nachtegael M, Kerre EE (2004) Using similarity measures and homogeneity for the comparison of images. Image Vis Comput 22(9):695–702
80. van Dijk AM, Martens J-B, Watson AB (1995) Quality asessment of coded images using numerical category scaling. In: Proc. SPIE—advanced image and video communications and storage technologies
81. van Nes FL, Bouman MA (1967) Spatial modulation transfer in the human eye. J Opt Soc Am 57(3):401–406
82. Wandell BA (1995) Foundations of vision. Sinauer Associates Inc., Sunderland
83. Wang S, Sekey A, Gersho A (1992) An objective measure for predicting subjective quality of speech coders. IEEE J Sel Areas Commun 10(5):819–829
84. Wang Z, Bovik AC (2002) A universal image quality index. IEEE Signal Process Lett 9(3): 81–84
85. Wang Z, Bovik AC (2006) Modern image quality assessment. Morgan and Claypool Publishing Co., New York
86. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
87. Wang Z, Li Q (2007) Video quality assessment using a statistical model of human visual speed perception. J Opt Soc Am A Opt Image Sci Vis 24(12):B61–B69
88. Wang Z, Lu L, Bovik AC (2004) Video quality assessment based on structural distortion measurement. Signal Process Image Commun 19(2):121–132
89. Wang Z, Simoncelli E, Bovik A, Matthews M (2003) Multiscale structural similarity for image quality assessment. In: IEEE asilomar conference on signals, systems and computers, pp 1398–1402
90. Wang Z, Simoncelli EP (2005) Translation insensitive image similarity in complex wavelet domain. In: IEEE international conference on acoustics, speech, and signal processing, pp 573–576
91. Watson AB (1987) The cortex transform: rapid computation of simulated neural images. Comput Vis Graph Image Process 39(3):311–327
92. Watson AB (ed) (1993) Digital images and human vision. The MIT
93. Watson AB, Hu J, McGowan JF III (2001) Digital video quality metric based on human vision. J Electron Imaging 10(1):20–29
94. Winkler S (1999) Perceptual distortion metric for digital color video. In: Proc. SPIE human vision and electronic imaging, vol 3644, no 1. San Jose, CA, pp 175–184
95. Winkler S (2005) Digital video quality. Wiley, New York
96. Zielinski SK, Rumsey F, Kassier R, Bech S (2005) Development and initial validation of a multichannel audio quality expert system. J Audio Eng Soc 53(1/2):4–21
97. Zwicker E (1961) Subdivision of the audible frequency range into critical bands (frequenzgruppen). J Acoust Soc Am 33(2):248–248
98. Zwicker E, Scharf B (1965) A model of loudness summation. Psychol Rev 72(1):3–26

Kalpana Seshadrinathan received the B.Tech. degree from the University of Kerala, India in 2002 and the M.S. and Ph.D. degrees in Electrical Engineering from the University of Texas at Austin, in 2004 and 2008, respectively. She is currently a Research Scientist with Intel Corporation in Santa Clara, CA. Her research interests include image and video quality assessment, computational aspects of human vision, motion estimation and applications, computational photography and statistical modeling of images and video. She is a recipient of the 2003 Texas Telecommunications Engineering Consortium Graduate Fellowship and the 2007 Graduate Student Professional Development Award from the University of Texas at Austin. She was Assistant Director of the Laboratory for Image and Video Engineering (LIVE) at the University of Texas at Austin from 2005-2008. She is a member of the IEEE.



Alan Conrad Bovik is the Curry/Cullen Trust Endowed Chair Professor at The University of Texas at Austin, where he is the Director of the Laboratory for Image and Video Engineering (LIVE). He is a faculty member in the Department of Electrical and Computer Engineering and the Center for Perceptual Systems in the Institute for Neuroscience. His research interests include image and video processing, computational vision, and visual perception. He has published over 500 technical articles in these areas and holds two U.S. patents. He is the author of The Handbook of Image and Video Processing (Academic Press, 2005), Modern Image Quality Assessment (Morgan & Claypool, 2006), and two recent books, The Essential Guide to Image Processing and The Essential Guide to Video Processing (Academic Press, 2009). Dr. Bovik has received a number of major awards from the IEEE Signal Processing Society, including: the Best Paper Award (2009); the Education Award (2007); the Technical Achievement Award (2005), and the Meritorious Service Award (1998). He is also a recipient of the Hocott Award for Distinguished Engineering Research at the University of Texas at Austin; received the Distinguished Alumni Award from the University of Illinois at Champaign-Urbana (2008), the IEEE Third Millennium Medal (2000) and two journal paper awards from the international Pattern Recognition Society (1988 and 1993). He is a Fellow of the IEEE, a Fellow of

the Optical Society of America, and a Fellow of the Society of Photo-Optical and Instrumentation Engineers. He has been involved in numerous professional society activities, including: Board of Governors, IEEE Signal Processing Society, 1996-1998; Editor-in-Chief, IEEE Transactions on Image Processing, 1996-2002; Editorial Board, The Proceedings of the IEEE, 1998-2004; Series Editor for Image, Video, and Multimedia Processing, Morgan and Claypool Publishing Company, 2003-present; and Founding General Chairman, First IEEE International Conference on Image Processing, held in Austin, Texas, in November, 1994. Dr. Bovik is a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial and academic institutions.