Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison

Shyamprasad Chikkerur, Vijay Sundaram, *Member, IEEE*, Martin Reisslein, and Lina J. Karam

***Abstract*—With the increasing demand for video-based applications, the reliable prediction of video quality has increased in importance. Numerous video quality assessment methods and metrics have been proposed over the past years with varying computational complexity and accuracy. In this paper, we introduce a classification scheme for full-reference and reduced-reference media-layer objective video quality assessment methods. Our classification scheme first classifies a method according to whether natural visual characteristics or perceptual (human visual system) characteristics are considered. We further subclassify natural visual characteristics methods into methods based on natural visual statistics or natural visual features. We subclassify perceptual characteristics methods into frequency- or pixel-domain methods. According to our classification scheme, we comprehensively review and compare the media-layer objective video quality models for both standard resolution and high definition video. We find that the natural visual statistics based MultiScale-Structural SIMilarity index (MS-SSIM), the natural visual feature based Video Quality Metric (VQM), and the perceptual spatio-temporal frequency-domain based MOtion-based Video Integrity Evaluation (MOVIE) index give the best performance for the LIVE Video Quality Database.**

***Index Terms*—Full-reference metric, objective video quality, perceptual video quality, reduced-reference metric.**

I. Introduction

THE advent of high performance video compression standards [1]–[3] in conjunction with efficient and ubiquitous transmission systems [4]–[8], and a myriad of consumer video technologies have brought the contemporary world closer to digital videos than ever before. According to recent forecasts, e.g., [9], video transmitted to and from mobile devices will account for 66% of the global mobile data traffic by 2014. This has increased the onus on video service providers to match the video quality expectations of the end user. The reliable assessment of video quality plays an important role in meeting the promised quality of service (QoS) and in improving the end user's quality of experience (QoE) [10]. More specifically, in a video transport system, it is important to monitor the network transport QoS through network QoS parameters, such as packet delay and packet loss rates [11], as well as the QoS of the video service through video related parameters, including start-up delay of the video playback and video quality, which ultimately contribute to the user's QoE [12]. Moreover, the video quality can be used in gauging the performance of the various components of a video transport system, including compression, processing, and transmission components. Controlling and monitoring the QoS parameters of the individual system components by appropriately selecting system parameters (such as compression ratios and reserved network bandwidth) is important for efficiently achieving high overall system performance and user QoE.

The traditional video quality metrics[1], such as signal-to-noise ratio (SNR), peak-signal-to-noise ratio (PSNR), and mean squared error (MSE), though computationally simple, are known to disregard the viewing conditions and the characteristics of human visual perception [13]. Subjective video quality assessment methods are able to reliably measure the video quality that is perceived by the Human Visual System (HVS) and are crucial for evaluating the performance of objective visual quality assessment metrics. The subjective video quality methods are based on groups of trained/untrained users viewing the video content, and providing ratings for quality [14]. Also, to meet the ITU-T recommendations for subjective quality evaluation, the tests have to follow strict evaluation conditions, including conditions on viewing distance, room illumination, test duration, and evaluators' selection [15], [16]. Though subjective video quality evaluation methods can capture reliably the perceived video quality, they are unable to provide instantaneous measurement of video quality and they are time consuming, laborious and expensive. This has led to a growing interest in developing objective quality assessment algorithms. Similar to traditional subjective metrics, objective quality metrics are required to produce video quality scores that reflect the perceived video quality, and they should highly correlate with the subjective assessments provided by human evaluators.

The Video Quality Experts Group (VQEG) is the principal forum that validates objective video quality metric models that result in International Telecommunication Union (ITU) recommendations and standards for objective quality models for both television and multimedia applications [17]. Our

Manuscript received August 09, 2010; revised December 12, 2010; accepted December 20, 2010 Date of publication February 10, 2011; date of current version May 25, 2011. This work was supported in part by the National Science Foundation under Grant CRI-0750927.

The authors are with the School of Electrical, Computer, and Energy Engineering, Arizona State University, Tempe, AZ 85287-5706 USA (e-mail: schikker@asu.edu; vijays@asu.edu; reisslein@asu.edu; karam@asu.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2011.2104671

[1]Throughout this article we use the term "video quality metric" to mean a "measure of video quality".

review includes the top-performing methods from the VQEG assessments, which have been incorporated as normative models in ITU recommendations for objective video quality measurements.

As per the ITU standardization activities, the objective quality measurement methods have been classified into the following five main categories [18] depending on the type of input data that is being used for quality assessment:

(1) Media-layer models—These models use the speech or video signal to compute the Quality of Experience (QoE). These models do not require any information about the system under testing, hence can be best applied to scenarios such as codec comparison and codec optimization.
(2) Parametric packet-layer models—Unlike the media-layer models, the parametric packet-layer models predict the QoE only from the packet-header information and do not have access to media signals. But this forms a lightweight solution for predicting QoE as it does not have to process the media signals.
(3) Parametric planning models—These models make use of quality planning parameters for networks and terminals to predict the QoE. As a result they require a priori knowledge about the system that is being tested.
(4) Bitstream-layer models—These models use encoded bitstream information and packet-layer information that is used in parametric packet-layer models for measuring QoE.
(5) Hybrid models—These models mainly combine two or more of the preceding models.

As illustrated in Fig. 1, the media-layer objective quality assessment methods can be further categorized as full-reference (FR), reduced-reference (RR), and no-reference (NR) [19] depending on whether a reference, partial information about a reference, or no reference is used in assessing the quality, respectively. Full- and reduced-reference methods are important for the evaluation of video systems in non-real-time scenarios where both *(i)* the original (reference) video data or a reduced feature data set, and *(ii)* the distorted video data are available. For instance, during the development and prototyping process of video transport systems, the original video can be delivered offline for full-reference quality assessment at the receiver, or the received distorted video data can be reliably (without any further bit loss or modifications) delivered back to the sender. In contrast, for real-time quality assessments at the receiver without availability of the original video data, low-complexity reduced-reference or no-reference methods are needed. The objective methods can also be classified in terms of their usability in the context of adaptive streaming solutions [20], [21] as out-of-service methods and in-service methods. In the out-of-service methods, no time constraints are imposed and the original sequence can be available. Full-reference visual quality assessment metrics and high-complexity non real-time RR and NR metrics fall within this class. On the other hand, the in-service methods place strict time constraints on the quality assessment and are performed during streaming applications.

In this article, we provide an up-to-date classification, review, and performance comparison of existing and contemporary media-layer full-reference and reduced-reference objective video quality metrics. For parametric-packet layer, parametric planning, and bitstream-layer models, we refer to [22]–[29]. For no-reference video quality measurement methods, we refer to [30]–[37]. In one of the earliest works, Olsson *et al.* [38] presented a survey on objective quality models for both image and video quality, and mainly for MPEG-2 compressed video. Further, fundamentals of perceptual models for video quality metrics and overviews of metrics developed prior to 2005 are provided in [39], [40].

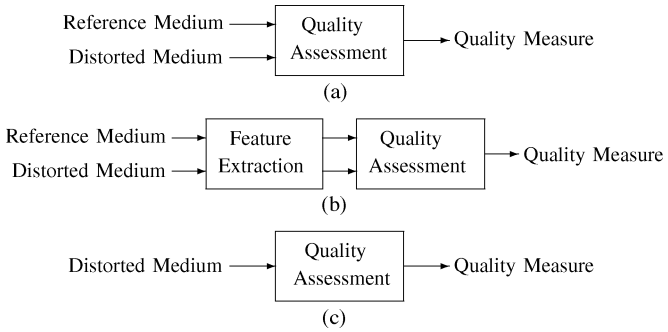


Fig. 1. Overview of media layer models [30].

The paper is organized as follows. We briefly review the factors affecting the perceived video quality in Section II. We briefly explain the subjective video quality assessments and the metrics for assessing how closely an objective metric predicts subjective quality ratings in Section III. In Section IV, we introduce our classification scheme of the full-reference and reduced-reference media-layer models and review the methods. In Section VI, we compare the performance of state-of-the-art methods from the various categories of our classification scheme. We summarize our findings in Section VII.

II. Factors Affecting Perceived Visual Quality

Many factors can affect and/or impair the quality of visual media including, but not limited to, acquisition, processing, compression, transmission, display and reproduction systems. Most of the contemporary video coding standards use motion compensation and block-based coding schemes for compression. As a result, the decoded video suffers from one or more of the compression artifacts, such as blockiness, blurriness, color bleeding, ringing, false edges, jagged motion, chrominance mismatch, and flickering. Transmission errors such as damaged or lost packets can further degrade the video quality. Furthermore, the pre- or post-processing stages in the video transmission system, such as domain conversion (analog to digital or vice-versa), frame rate conversion, and de-interlacing degrade the video.

It has been also shown that the perceived quality heavily depends upon the viewing distance, display size, resolution of video, brightness, contrast, sharpness, color, content (faces versus other objects), and naturalness [41]. Studies [41] show that some viewers may prefer more colorful images, while this might actually reduce the naturalness of the video content. In [42], it was observed that test scenes accompanied by good audio quality masked to some extent the perceived video

degradation. Moreover, adverse environmental conditions, such as turbulence, atmospheric particles and fog, as well as motion and vibrations, can degrade the perceived video quality.

Though tedious, when conducted properly, the subjective video quality assessment approaches are more accurate than the objective ones. Accounting for various degradations and other important factors is a challenging task for objective video quality models. Thus, in the recent years, there has been a growing interest in the development of advanced objective video quality models that can closely match the performance of subjective video quality evaluation.

III. Performance Evaluation of Objective Video Quality Metrics

Subjective video models serve as a benchmark for the performance evaluation of objective models. The perceptual video quality predicted by objective models is always compared for degree of closeness with the perceptual quality measured with traditional subjective models. The prominent subjective tests used from ITU-R Rec. BT.500-11 [14] and ITU-T Rec.P.910 [16] are:

(a) Double Stimulus Continuous Quality Scale (DSCQS) [ITU-R Rec. BT.500-11]—In this test, the reference and processed video sequence are presented twice to the evaluators in alternating fashion, with randomly chosen order (Example: reference, degraded, reference, degraded). At the end of the screening, the evaluators are asked to rate the video quality on a continuous quality scale of 0–100 (with 0 being *Bad* and 100 *Excellent*). Multiple pairs of reference and processed video sequences and of rather short durations (around 10 seconds) are used. The evaluators are not told which video sequence is the reference and which is the processed.

(b) Double Stimulus Impairment Scale (DSIS) [ITU-R Rec. BT.500-11]—Unlike the DSCQS, in the DSIS, the evaluators are aware of the presentation sequence, and each sequence is showed only once. The reference video sequence is shown first followed by the processed video sequence. (In DSIS variant II, this presentation sequence is repeated once.) The evaluators rate the sequences on a discrete five-level scale ranging from *very annoying* to *imperceptible* after watching the video sequences. ITU-T Rec.P.910 has an identical method called Degradation Category Rating (DCR).

(c) Single Stimulus Continuous Quality Evaluation (SSCQE) [ITU-R Rec. BT.500-11]—As the name suggests, the evaluators are only shown the processed video sequence, usually of long duration (typically 20–30 minutes). The evaluators rate the instantaneous perceived quality on the DSCQS scale of *bad* to *excellent* using a slider.

(d) Absolute Category Rating (ACR) [ITU-T Rec.P. 910]—This is also a single stimulus method similar to SSCQE with only the processed video being shown to the evaluators. The evaluators provide one rating for the overall video quality using a discrete five-level scale ranging from *Bad* to *Excellent*

(e) Pair Comparison (PC) [ITU-T Rec.P.910]—In this method, test clips from the same scene but under varying conditions, are paired in all possible combinations and screened to the evaluators for preference judgment about each pair.

We briefly note that these subjective test scales have been extensively studied. For instance, a general methodology for creating valid scales is examined in [43]. The DSCQS and DSIS II scales have been compared in [44], revealing that the DSCQS scale is robust with respect to the level of video impairment, while the DSIS II scale exhibited high sensitivity to the impairment level. A multiple reference impairment scale (MRIS) that overcomes the impairment sensitivity of the DSIS II scale is proposed and examined in [45].

For all these methods, the perceptual video quality ratings obtained from the evaluators are averaged to obtain the Mean Opinion Score (MOS). In the case of DSCQS, the Difference Mean Opinion Score (DMOS) is used. The DMOS consists of the mean of differential subjective scores. For each subject and each video sequence, a differential subjective score is computed by subtracting the score assigned by the subject to the processed video sequence from the score assigned by the same subject to the corresponding reference video sequence. The differential scores of a given subject can be further normalized using the mean and the standard deviation of all the differential scores given by the considered subject to obtain Z-scores. The DMOS can then be computed by averaging the obtained Z-scores.

One of the responsibilities of the VQEG is to provide standardized test data and evaluation methodologies to test new video quality metrics. The performance of a perceptual quality metric depends on its correlation with subjective results. The performance of the objective models is evaluated with respect to the prediction accuracy, the prediction monotonicity, and the prediction consistency in relation to predicting the subjective assessment of video quality over the range of the considered video test sequences. In addition, by choosing a set of video sequences that include various impairments that are of interest, the robustness of an objective quality assessment metric can be tested with respect to a variety of video impairments.

As described in [46], there are four commonly used metrics that are used for evaluating the performance of objective video quality metrics (see for instance [47] for general background on correlation statistics). These include the following:

- The Pearson correlation coefficient (PCC) is the linear correlation coefficient between the predicted MOS (DMOS) and the subjective MOS (DMOS). It measures the prediction accuracy of a metric, i.e., the ability to predict the subjective quality ratings with low error. For N data pairs (x_i, y_i), with $\bar{x}$ and $\bar{y}$ being the means of the respective data sets, the PCC is given by:

$$\mathrm{PCC} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}. \tag{1}$$

  Typically, the PCC is computed after performing a nonlinear regression using a logistic function, as described in [48], in order to fit the objective metric quality scores to the subjective quality scores.
- The Spearman rank order correlation coefficient (SROCC) is the correlation coefficient between the predicted MOS (DMOS) and the subjective MOS (DMOS). It measures

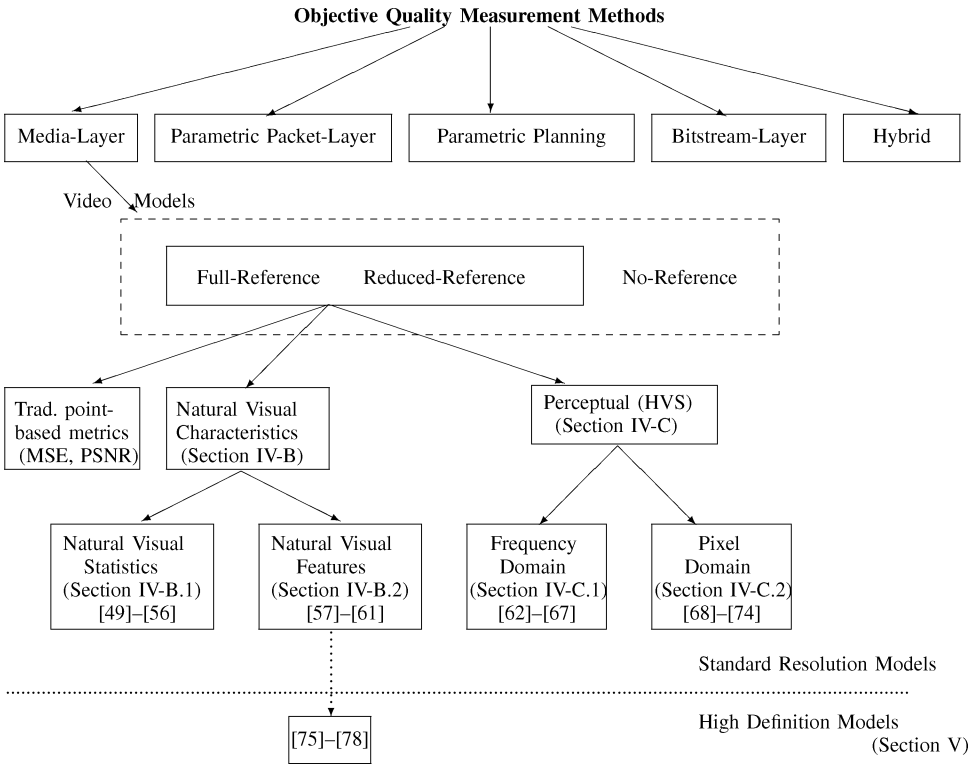


Fig. 2. Classification of media-layer objective video quality models. In this article, we focus on full-reference and reduced-reference models, which we classify into natural visual characteristics based models and perceptual (HVS) based models.

the prediction monotonicity of a metric, i.e., the degree to which the predictions of a metric agree with the relative magnitudes of the subjective quality ratings. The SROCC is defined as:

$$\mathrm{SROCC} = \frac{\sum(X_i - X')(Y_i - Y')}{\sqrt{\sum(X_i - X')^2}\sqrt{\sum(Y_i - Y')^2}}. \quad (2)$$

where X_i is the rank of x_i and Y_i the rank of y_i for the ordered data series and X' and Y' denote the respective midranks.

- The Outlier Ratio (OR) is defined as the percentage of the number of predictions outside the range of ± 2 times the standard deviations of the subjective results. It measures the prediction consistency, i.e., the degree to which the metric maintains the prediction accuracy. If N is the total number of data points and N' is the number of determined outliers, the outlier ratio is defined as:

$$\mathrm{OR} = \frac{N'}{N}. \quad (3)$$

- The Root Mean Square Error (RMSE) for N data points $x_i, i = 1, \ldots, N$, with $\bar{x}$ being the mean of the data set, is defined as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum(x_i - \bar{x})^2}. \quad (4)$$

The fidelity of an objective quality assessment metric to the subjective assessment is considered high if the Pearson and Spearman correlation coefficients are close to 1 and the outlier ratio is low. Some studies use the Root Mean Square Error (RMSE) to measure the degree of accuracy of the predicted objective scores. For the 95% confidence interval, it is desirable that the RMSE be less than 7.24 [39].

IV. Media-Layer Objective Video Quality Assessment Methods

A. Classification Overview

We classify and review the existing full-reference and reduced-reference video quality assessment methods in this section. As illustrated in Fig. 2, we classify the full-reference (FR) and reduce-reference (RR) video quality metrics into traditional point-based metrics (e.g., MSE and PSNR), Natural Visual Characteristics oriented metrics, and Perceptual (HVS) oriented metrics. We do not examine in further detail the traditional point-based metrics. We further classify the Natural Visual Characteristics metrics into Natural Visual Statistics and Natural Visual Features based methods. Similarly, we further classify the HVS methods into DCT domain, DWT domain, and pixel domain models. In Tables I and II, we highlight the key concepts behind the surveyed methods, the test details, and their comparative performance.

B. Natural Visual Characteristics

In Sections IV-B-1 and IV-B-2, we cover models that are based on statistical features and visual features, respectively. The statistical models use statistical measures, such as mean, variance, covariance, and distributions, in modeling their respective quality metrics. The visual features based models employ measurements of blurring and blocking in video as well as image segmentation for extracting significant visual

TABLE I
COMPARISON OF NATURAL VISUAL CHARACTERISTICS ORIENTED OBJECTIVE VIDEO QUALITY MODELS

Method	Approach	Test Details	Subj. Model	Performance
Natural Visual Statistics				
Wang et al. [49], SSIM	structural distortion measurement	VQEG Phase I, LIVE Image database	–	–
Wang et al. [50], VSSIM	structural distortion measurement based on SSIM	VQEG Phase I	–	PCC = 0.864 /w weighted regr., 0.849 /w non-lin. regr., SROCC = 0.812, OR = 0.578
Wang et al. [51], MS-SSIM	structural distortion measurement based on SSIM	VQEG Phase I, LIVE Image database	–	PCC = 0.969, SROCC = 0.966, RMSE = 4.91, and OR = 1.16
Sheikh and Bovik [53], VIF	stat. visual model in wavelet domain, distortion and HVS modeling, visual information	29 test images, five distortion types	–	SROCC = 0.949, RMSE = 5.08, and OR = 0.013
Lu et al. [54]	block-DCT and region classification (plane, edge, and textured)	VQEG Phase I	–	95% confidence interval error
Tao & Eskicioglu [56]	singular value decomposition (SVD)	VQEG Phase I test data set for FT-TV video qu. measurement	–	–
Natural Visual Features				
Pessoa et al. [57]	segments image/frame to plane, edge and textured region	MPEG-2 coded, five natural 2 sec. scene clips (*Garden, Mobile, Tennis, Diva* and *Kiel*)	DSIS	MAE less than 4% for each scene
Pinson and Wolf [58], VQM	edge impairment filter	VQEG FRTV Phase II	DSCQS	PCC = 0.938, OR = 0.46 for 525-line videos. PCC = 0.886, OR = 0.31 for 625-line vid.
Okamoto et al. [59]; NTT full ref. meth. [74]	PSNR, edge energy difference, moving energy of blocks	36 vids from ITU-R BT.802 and BT.1210 in 640x480 resol., encoded with Window Media 8 Encoder; VQEG MM Phase I test set	DSCQS; ACR-HR, DMOS	95% confidence interval; avg. PCC = 0.777, RMSE = 0.604, OR = 0.538 for CIF vid.
Lee and Sim [60]	degradation feature values of edges, boundary and blur	H.263 and H.264/AVC coded 140 video clips (CIF and QCIF resolution)	DSCQS	sum of absolute errors = 5.09 for training vid., 11.50 for test vid.
Bhat et al. [61]	MSE, edge information	*Carphone, Foreman, Mobile, News, Bus, Paris, Coastguard* CIF sequences H.264 compressed at different bit-rates	–	PCC = 0.947, and OR = 0.402

features, and edge detection to capture the edge, plane, and texture properties.

1) Natural Visual Statistics: Wang *et al.* [50] proposed the Video Structural Similarity (VSSIM) index which uses structural distortions to estimate perceptual distortions. The VSSIM technique intends to exploit the strong dependencies between samples of the signal. The degradations are considered to be due to perceptual structural information loss in the human visual system. The authors base the VSSIM metric on their previously proposed Structural Similarity Index (SSIM) [49] which was specific to still image quality assessment. SSIM defines the luminance comparison

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

where μ_x and μ_y denote the mean luminance intensities of the compared image signals $\mathbf{x}$ and $\mathbf{y}$. For an image with a dynamic range L, the stabilizing constant is set to $C_1 = (K_1 L)^2$ where K_1 is a small constant such that C_1 takes effect only when $(\mu_x^2 + \mu_y^2)$ is small. Similarly, SSIM defines a contrast comparison function

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

with σ_x and σ_y denoting the standard deviations of the luminance samples of the two images and C_2 is a stabilizing constant similar to C_1. Further, a structure comparison function is defined with the covariance of the luminance samples σ_{xy} as

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

The SSIM index is defined as

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha} \cdot [c(\mathbf{x},\mathbf{y})]^{\beta} \cdot [s(\mathbf{x},\mathbf{y})]^{\gamma}, \tag{8}$$

whereby the positive parameters α, β, and γ adjust the relative importance of the three comparison functions. Setting $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$ gives the specific form

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{9}$$

examined in [49]. The overall quality of the image is defined as the average of the quality map, i.e., the mean SSIM (MSSIM) index.

For video sequences, the VSSIM metric measures the quality of the distorted video in three levels, namely the local region

TABLE II
COMPARISON OF PERCEPTUAL (HVS) ORIENTED OBJECTIVE VIDEO QUALITY MODELS

Method	Approach	Test Details	Subj. Model	Performance
Frequency Domain				
Lukas and Budrikis [62]	visual thresholds	12 frames from *Judy* sequence	NA	coeff. quadr. regr.: raw $= 0.69$, filtered $= 0.80$, masked $= 0.88$.
Lambrecht & Verscheure [63], MPQM	contrast sensitivity	*Mobile, Calendar, Flower Garden, Basket Ball, Carphone*, *LTS* seq	–	–
Watson et al. [64], DVQ	visual thresholds using DCT transform	65 seq. (five orig. and 60 proc.) ITU-601 PAL Format (576x720, interl., 4:2:2 sampl.)	DSCQS	RMSE $= 14.61$
Xiao et al. [65]	JNDs with spatial contrast sensitivity function	*carphone* sequence	–	better than trad. point-based root MSE
Lee and Kwon [66]	discrete wavelet transform, segmenting frame to plane, edge, and textured region	Ten 8-sec test seq. x 16 reference circuits using H.263 and MPEG-2 enc. in 525/50 Hz and 625/60 Hz.	DMOS	correlation of 0.94
Seshadrinathan & Bovik [67], MOVIE	Gabor filter bank	VQEG FRTV Phase 1	DMOS	PCC $= 0.821$, SROCC $= 0.833$ and OR $= 0.644$
Pixel Domain				
Hekstra et al. [68], PVQM	edginess, color error, temporal decorrelation	20 video seq., processed by 16 video systems using H.263, MPEG2, ETSI codecs.	DMOS	correlation of 0.934
Lu et al [69], PQSM	visual distortion	VQEG [79] test data for SDTV sequences	–	PCC and SROCC upto 0.83 and 0.81, resp.
Ong et al. [70], [71]	distortion-invisibility, blockiness and content fidelity	*Container, Coast Guard, Japan League, Forem., News, Temp.*, CIF and QCIF, MPEG-4 coded	DSIS variant II	PCC, SROCC within 95% confidence interval
Nya et al. [72]	block and blur errors	MPEG data sets, QCIF, 10 Hz and 15 Hz, 10 s, 32 kbps and 64 kbps for MPEG-2, H.26L. QVGA, 12.5 Hz, 10 s, HHI VBR data sets	DMOS	PCC same as [82], higher than [70], [71], PSNR; SROCC higher than [70], [71], [82]
Chandler & Hemami [73], VSNR	visual masking and visual summation	LIVE Image database	–	PCC $=$ SROCC $=$ 0.889; RMSE $= 7.39$
PEVQ [74]	edginess in lumin. and chromin., temporal variability, frame delay/loss/freezing	VQEG Multimedia Phase I test data set for full-reference multimedia video quality measurement	ACR-HR, DMOS	avg. PCC $= 0.808$, RMSE $= 0.562$, OR $=$ 0.513 for CIF video
Psytechnics [74]	analysis of spatial frequency, edge distortion, blur, block distortion, spatial/temporal distortion	VQEG Multimedia Phase I test data set for full-reference multimedia video quality measurement	ACR-HR, DMOS	avg. PCC $= 0.836$, RMSE $= 0.526$, OR $=$ 0.507 for CIF video

level, the frame level, and the sequence level. The local quality index is obtained as a function of the SSIM indices for the Y, Cb, and Cr components as

$$\mathrm{SSIM}_{ij} = W_Y \cdot \mathrm{SSIM}_{ij}^{Y} + W_{Cb} \cdot \mathrm{SSIM}_{ij}^{Cb} + W_{Cr} \cdot \mathrm{SSIM}_{ij}^{Cr}, \tag{10}$$

where W_Y, W_{Cb}, and W_{Cr} are weights for the Y, Cb, and Cr components. Based on the reasoning that the luminance distortion measure has more impact on the video quality than the chroma distortion, Wang *et al.* fix $W_Y = 0.8$ and $W_{Cb} = W_{Cr} = 0.1$ [50]. At the second level, the local level quality values are weighted to give a frame level quality measure which is in turn weighted to obtain the overall quality of the video sequence. The metric was tested on the VQEG Phase 1 test data set with the Pearson correlation, the Spearman correlation, and the Outlier ratio. In addition to its simplicity, the VSSIM was shown in [50] to provide reasonably good results as compared to the PSNR, the KPN/Swisscom CT (the best metric of VQEG Phase 1 in terms of performance [48], [79]).

In addition to the SSIM and the VSSIM, the MultiScale-SSIM (MS-SSIM) [51] and the Speed SSIM [52] metrics have been proposed. The MS-SSIM is an extension of the single-scale approach used in SSIM and provides more flexibility by incorporating the variations of the image resolution and viewing conditions. At every stage (also referred to as scale), the MS-SSIM method applies a low pass filter to the reference and distorted images and downsamples the filtered images by a factor of two. At the mth scale, contrast and structure comparisons are evaluated according to Eqns. (6) and (7) and denoted as $c_m(\mathbf{x},\mathbf{y})$ and $s_m(\mathbf{x},\mathbf{y})$, respectively. The luminance comparison (5) is computed at scale M (i.e., the highest scale obtained after $M-1$ iterations) and denoted as $l_M(\mathbf{x},\mathbf{y})$. Combining the scales gives

$$\mathrm{MS-SSIM}(\mathbf{x},\mathbf{y}) = [l_M(\mathbf{x},\mathbf{y})]^{\alpha} \cdot \prod_{m=1}^{M} [c_m(\mathbf{x},\mathbf{y})]^{\beta} \cdot [s_m(\mathbf{x},\mathbf{y})]^{\gamma}, \tag{11}$$

which has been shown to outperform the SSIM index and many other still image quality assessment algorithms [80].

The MS-SSIM index can be extended to video by applying it frame-by-frame on the luminance component of the video and the overall MS-SSIM index for the video is computed as the average of the frame level quality scores. The Speed SSIM is the VQA model proposed in [52] and uses the SSIM index in conjunction with statistical models of visual speed perception described in [81]. Using models of visual speed perception with the SSIM index was shown in [52] to improve the performance as compared to PSNR and SSIM.

The Visual Information Fidelity (VIF) [53] is based on visual statistics combined with HVS modeling. VIF models natural images as realizations of Gaussian Scale Mixtures in the wavelet domain. VIF first models the distorted image through signal attenuation and additive Gaussian noise in the wavelet domain. Then, the masking and sensitivity aspects of the HVS are modeled through a zero mean, additive white Gaussian noise model in the wavelet domain that is applied to both the reference image and the distorted image model. The visual information of both images is quantified by the mutual information between the input natural image and the respective images at the output of the HVS model. The ratio of the visual information of the distorted image to the visual information of the reference image is defined as the VIF measure.

Similar to the VSSIM, Lu *et al.* [54] proposed a full reference video quality assessment model based on structural distortion measurements. The first stage evaluates the MSSIM by randomly selecting localized areas and computing statistical features, such as mean and variance, to obtain the local quality and the frame quality measure (as in VSSIM). The authors then adjust the frame quality value by measuring the blockiness and blurriness as well as the motion factor. Blocking and blurring, which are measured from the power spectrum of the signal, as well as the relative motion, which is measured using a block-based motion compensation algorithm, are incorporated adaptively based on the quality index of the frame. The final frame quality index is obtained as a weighted sum of the results for Y, Cr, and Cb. Averaging over all frames gives the overall quality value for the test sequence. The metric was tested with the VQEG Phase 1 data set and showed consistency with subjective measurements when evaluated using the Spearman and the Pearson coefficients [54]. Applications such as low bit rate MPEG coding suit the metric.

Shnayderman *et al.* [55] developed a distortion measure called M-SVD for image quality assessment based on the concept of singular value decomposition. Singular Value Decomposition is a way of factoring matrices into a series of linear approximations that expose the underlying structure of the matrix. The M-SVD measures distortion as a function of the distance between the original and distorted image block singular values, given by

$$D = \sqrt{\sum_{i=1}^{n} (s_i - \hat{s}_i)^2}, \tag{12}$$

where s_i and $\hat{s}_i$ represent the singular values of the original and distorted block, and n represents the block size. Once the distance measures are computed for all blocks, a global measure is derived by averaging the differences between the distance measure for each block and the median of all block distance measures. This global error is used to derive the M-SVD measure. Using this concept of distortion measure, Tao and Eskicioglu [56] developed a full-reference objective video quality model. Initially, both the original and degraded video sequences are converted to the 4:4:4 YCbCr format, and the frames are decomposed into 8×8 blocks. Then, the distance measures are computed for all the blocks in each frame. To account for the HVS sensitivity to high frequency regions, edge detection for each block is conducted using a local gradient filter, such as Sobel. Each block is assigned an edge index based on the degree of edge content. The M-SVD is derived as a function of distance measures of each block and their respective edge indices. The error index for a frame is expressed as a linear weighted sum of M-SVDs computed for both the luminance and chroma components, with weights derived experimentally from test video sequences. The overall quality of the video sequence is expressed as an average of the error indices across all frames. The performance evaluation for this method was performed using video sequences from the VQEG Phase I test data set for FT-TV video quality measurement. A variance-weighted regression analysis correlation score of 0.893, non-linear regression analysis correlation score of 0.877, SROCC of 0.799 and OR of 0.486 were observed, when objective video quality was measured using both the luma and chroma components with edge detection. The performance of the model was observed to be better when both the luma and chroma components were used with edge detection, as compared to using only the luma component, or both the luma and chroma components without edge detection.

2) Natural Visual Features: Pessoa *et al.* [57] presented a video quality model that segments images into plane, edge, and texture regions. The region segmentation helps in capturing the degree of perceived distortion. For example, blockiness is more noticeable in plane (flat) regions, and blurriness is more noticeable in edge and textured regions. Pessoa *et al.* [57] evaluated the model using three different segmentation algorithms: *(i)* segmentation based on edge detection using recursive filtering and a median filter, *(ii)* fuzzy image segmentation based on spatial features, and *(iii)* a watershed algorithm. After segmentation, for each region, error measures including the Mean Square Error (MSE), Positive Sobel Difference (PSD), Negative Sobel Difference (NSD), and Absolute Sobel Difference (ASD) are computed for both the luminance and chrominance components from the reference and processed video signal. The ASD is the sum of PSD and NSD. For a given region, if $R(x,y)$ is the pixel value of the original frame and $D(x,y)$ is the pixel value of the distorted frame, and $R_m(x,y)$ and $D_m(x,y)$ are the corresponding pixel values after median filtering, then the PSD and NSD are given by:

$$\text{PSD} = \max_{x,y}[\text{sobel}\{R_m(x,y)\} - \text{sobel}\{D_m(x,y)\}, 0] \tag{13}$$

$$\text{NSD} = -\max_{x,y}[\text{sobel}\{D_m(x,y)\} - \text{sobel}\{R_m(x,y)\}, 0]. \tag{14}$$

For each impairment objective parameter, weights are computed such as to satisfy a statistical reliability constraint. The statistical reliability is defined to be inversely proportional to the mean

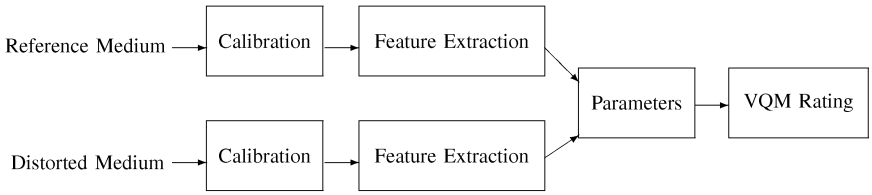


Fig. 3. Block diagram of the NTIA VQM general model.

squared error between the objective parameter and the normalized subjective score. The final objective score is obtained as a weighted linear combination of all these objective parameters. Five 2-second clips of MPEG-2 coded natural scenes and the DSIS subjective quality test were used for the performance evaluation. The objective test results showed a mean absolute error (MAE) of less than 4% for each individual scene and a global MAE of 1.8%, when the first segmentation algorithm was used. The second and third segmentation methods resulted in no significant drop in objective quality estimation accuracy. Pessoa *et al.* [57] note that the results could be improved if temporal details are also considered as the method does not use any temporal information for video quality evaluation.

The Video Quality Metric (VQM) software tools [58] developed by the Institute for Telecommunication Services (ITS), the research and engineering branch of the National Telecommunications and Information Administration (NTIA), provide standardized (for digital cable TV applications) as well as non-standardized (expanded) methods to measure the perceived video quality of digital video systems. The expanded methods can be used to measure the perceived video quality for various video applications, including direct broadcast satellites (DBS), standard definition television (SDTV), high definition television (HDTV), video teleconferencing (VTC), and wireless or IP-based video streaming systems. The NTIA VQM provides several quality models, such as the Television model, the General Model, and the Video Conferencing Model, based on the video sequence under consideration and with several calibration options prior to feature extraction in order to produce highly efficient quality ratings. We provide here a brief review of the NTIA General Model, which is illustrated in Fig. 3, and which provides objective video quality ratings for video sequences that span a wide range of quality levels. The main impairments considered in the General Model include blurring, block distortion, jerky/unnatural motion, noise in luminance and chrominance channels, and error blocks (e.g., transmission errors). The blurring information is computed using a 13-pixel information filter (SI13). The SI13 is a perceptually significant edge impairment filter defined in [82], with a peak response around 4.5 cycles/degree and that makes use of 13 $\times$ 13 horizontal and vertical filter masks. Jerky/unnatural motion is detected by considering the shift of horizontal and vertical edges with respect to diagonal orientation due to high blurring. The output of the SI13 filter is used to measure this unnatural motion by considering edge angles. Also, using the SI13 filter, the shift of edges from the diagonal to horizontal and vertical orientations due to tiling or blocking artifacts is considered. Then, the distribution of chroma spaces is computed to consider the color impairments by dividing both the chroma planes into 8 pixels $\times$ 8 lines $\times$ 1 frame spatio-temporal regions. In addition, the model also considers a quality improvement parameter that might result from edge sharpening or enhancements. The amount of perceived temporal impairment is influenced by the amount of spatial detail. Using the features derived from the product of contrast information and amount of spatial detail, the temporal distortion is computed. Finally, using the same color features as used in computing the chroma spread earlier, localized color impairments such as those caused by digital transmission errors are accounted for. A weighted linear combination of all the impairments metrics is used to arrive at the VQM rating. The NTIA VQM General Model was the only model that broke the 0.9 threshold of the Pearson correlation coefficient on the VQEG FRTV Phase II test database [46] and, as a result, was standardized by ANSI in July 2003 (ANSI T1.801.03-2003) and included as normative model in ITU Recommendations ITU-T J.144 and ITU-R BT.1683 (both adopted in 2004).

Okamoto *et al.* [59] proposed a video quality metric that considers visual distortions including blurring of the edge sections, generation of new edges, and deterioration in the temporal direction. Using the Average Edge Energy Difference metric presented in ANSI T1.801.03 [83], Okamoto *et al.* investigate the quality prediction accuracy of this metric in relation to the prediction of deteriorations in edge regions. The Average Edge Energy Difference metric is the difference in the number of edges between the original and degraded video per frame divided by the number of edges in the original video frame. This metric is found to be insufficient to account for deteriorations, such as mosquito noise and blurring in the edge regions, and degradations in the temporal domain, and is also found to treat the entire frame uniformly without accounting for the local nature of deteriorations. To account for mosquito noise and blurring distortions around the edge regions, a minimum edge ratio metric is used. To identify blocking distortions, the amount of distortion between the horizontal/vertical edges and the other edges is calculated. The average moving energy of blocks is computed to account for the temporal and local nature of degradations. A weighted sum of these measures is used to predict the video quality, with weighting coefficients arrived at using dual regression analysis using a subjective training dataset. When compared with the DSCQS subjective quality prediction scores, the RMSE is found to be 6.43 which falls within the 95% confidence interval. The tests were done using 36 videos selected from ITU-R BT.802 and BT.1210 recommendations. All the test videos were 640 $\times$ 480 in spatial resolution, with the Windows Media 8 Encoder used as the codec. Based on the good performance in [84], as summarized in Table I, this NTT full reference method was included as normative model in [74].

Lee and Sim [60] measure visual features at the edges and the block boundary regions. Their proposed KVQM metric computes feature values that indicate the visual degradation of the image, namely the edginess, blockiness, and the blurriness. The final quality metric score is obtained by a weighted linear combination of the three feature metrics as:

$$\mathrm{KVQM} = w_1 \cdot M_{\mathrm{edge}} + w_2 \cdot M_{\mathrm{block}} + w_3 \cdot G_{\mathrm{diff}} + \mathrm{offset} \quad (15)$$

where w_1, w_2, and w_3 represent the weights that are derived based on linear regression analysis on a training test video set of 50 clips. The performance of the model is evaluated by comparing the Sum of Absolute Error (SAE) values between the subjective model (DSCQS) and the KVQM using a training data set. The aim of the KVQM was to assess the objective quality of digital mobile videos.

More recently, Bhat *et al.* [61] presented a method exploiting the correlation between objective and subjective results. Bhat *et al.* determine the correlation between the predicted Mean Opinion Score MOS_p and the Mean Square Error (MSE) using the linear correlation model

$$\mathrm{MOS}_p = 1 - k(\mathrm{MSE}), \quad (16)$$

where k is the slope of the regression line. The authors train this MOS_p model with a variety of video sequences. Since the visibility of artifacts is low in highly detailed regions, the spatial edge information is extracted using edge filters and is fit into the linear model to determine k as follows:

$$k = 0.03585 \cdot \exp(-0.02439 \cdot \mathrm{SequenceEdgeStrength}). \quad (17)$$

Similar to VSSIM, the MOS_p metric is calculated first at the macroblock level, and subsequently the macroblock level MOS_p scores are averaged out to obtain the frame level quality measure and then the overall quality of the video sequence. The metric of [61] is evaluated using the Pearson correlation coefficient and the Outlier ratio for a variety of video sequences with low and high levels of detail. Compared to the PSNR, SSIM, and PSNRplus [85], it was reported in [61] that the MOS_p metric performs better in terms of both subjective results as well as speed on the tested video sequences.

C. Perceptual (HVS)

In this section, we discuss metrics which have been modeled based on Human Visual System (HVS) characteristics, both in the frequency as well as pixel domains. In the frequency domain, transforms such as DCT, wavelets, and Gabor filter banks are used to measure the impairments in different frequency regions. In the pixel domain, the impairments are measured using change in local gradient strength around a pixel or based on perceptually significant visual features. In these models, perceptual features motivated from computational models of low level vision are extracted to provide a reduced description of the image.

1) Frequency Domain: While one of the earliest color image quality metrics was proposed by Faugeras [86], one of the earliest video quality metrics based on a vision model was developed by Lukas and Budrikis [62]. In [62], the first stage of the model constitutes a nonlinear spatio-temporal model of a visual filter describing threshold characteristics on uniform background fields. The second stage incorporates a masking function in the form of a point-by-point weighting of the filtered error based on the spatial and temporal activity in the immediate surroundings in order to account for the non-uniform background fields. The processed error, averaged over the picture, is then used as a prediction of the picture quality. The model attempted to predict the subjective quality of moving monochrome television pictures containing arbitrary impairments. Out of the three classes of distortion measures used, namely raw, filtered, and masked, the filtered error measure provided the best quality prediction.

The MPQM by van den Branden Lambrecht and Verscheure [63] simulates the spatio-temporal model of the human visual system with a filter bank approach. The perceptual decomposition of the filter accounted for the key aspects of contrast sensitivity and masking. Since the eye's sensitivity varies as a function of spatial frequency, orientation, and temporal frequency, and the perception of a stimulus is a function of its background, the authors jointly modeled the contrast sensitivity function and the masking function to explain visual detection. The metric also accounted for the normalization of cortical receptive field responses and intra-channel masking. Pooling of the prediction data from the original and coded sequences in the multi-channel model justifies higher levels of perception. The authors present a global quality measure and also metrics for the performance of basic features, such as uniform areas, contours, and textures in a video sequence. The metrics were tested for applications of high bitrate broadcasting using the MPEG-2 coder and low bit rate communication using H.263. The sequences used are *Mobile, Calendar, Flower Garden, and Basket Ball* for the MPEG-2 coder and *Carphone* and *LTS* Sequence for H.263. Conducting encoding experiments, the metric's saturation effect is compared with PSNR and found to be in correlation with aspects of human vision.

The Digital Video Quality (DVQ) model described by Watson *et al.* [64] incorporates the discrete cosine transform to gauge the objective video quality. The model considers aspects of luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation for quality evaluation. After calibration and pre-processing of both the original and processed video sequences, a block DCT is applied, using a block size of 8×8 pixels. The ratio of DCT amplitude to DC component for the corresponding block is computed to estimate the local contrast. Using a suitable recursive discrete second-order filter, temporal filtering is conducted to compute temporal contrast sensitivity. From the local contrast information, just-noticeable differences (JNDs) are estimated for both sequences. The difference between the DCT coefficients of the original and test sequences is computed over local regions and converted into JND units by dividing it by the local JNDs. Also, using the original sequence, after JND conversion, a first order low-pass IIR filter is applied to estimate the degree of temporal masking. Finally, using the Minkowski metric, the JND-weighted differences are first pooled over each video frame and then over all the sequence of video frames in order

to estimate the visual quality of the video sequence. Sixty-five test sequences (five original and 60 processed) of ITU-601 PAL Format (576×720, interlaced, 4:2:2 sampling) were used for testing the metric. The quality ratings obtained were found to have RMS error of 14.61 when compared with scores from the double stimulus continuous quality scale (DSCQS) subjective test. However, it was observed that, the metric was not a good fit for sequences at very low bit rates.

Subsequently, as an extension of Watson's DVQ [64], Xiao [65] proposed a modification which made use of the fact that the human eyes' sensitivity to spatio-temporal patterns decreases with high spatial and temporal frequencies. The method is similar to Watson's model, except that the local contrast achieved with the DC components is further converted to just noticeable differences using a spatial contrast sensitivity (SCS) matrix for static frames and a matrix (for e.g., the SCS matrix raised to a power) which accounts for the temporal property for dynamic frames. This DCT based video quality metric also called VQM (not to be confused with NTIA VQM [58]) was defined in terms of a weighted mean distortion $\bar{D}$ and a maximum distortion $D_{\max}$ as follows:

$$\mathrm{VQM} = [\bar{D} + 0.005 \cdot D_{\max}], \tag{18}$$

where the mean and maximum distortions were obtained based on the absolute differences between the original and compressed video sequences. The metric's performance was compared to the Root Mean Squared Error (RMSE) with tests involving addition of spatial frequency noise to images and block-based distortions. It performs better than RMSE in terms of correlation with subjective scores.

Lee and Kwon [66] proposed an objective video quality model based on the wavelet transform. The model uses a multi-level wavelet transform to compute the spatial frequencies based on the resulting subbands. For each subband of the frame, the difference squared error between the original and processed wavelet coefficients is computed and summed, resulting in an error vector for each frame. These error vectors only capture the spatial frequency degradation. For capturing the temporal degradation, a modified 3-D wavelet transform is applied on the 2-D array formed by arranging the error vectors for each frame as a column. Finally, an average of the resulting vectors is computed to account for both the spatial and temporal degradation. From the generated difference vectors, the quality rating is derived as a weighted sum of the vector elements. The weights are derived using a training data set, based on maximizing the degree of correlation between the given subjective scores and the predicted objective scores. The validation tests were performed on two video formats (525/50 Hz and 625/60 Hz), both of 8 seconds duration, with coding methods H.263 and MPEG-2 for test sequences. The testbench comprised of 10 input video sequences and 16 hypothetical reference circuits for each. It was found that the quality ratings showed a high correlation of 0.94 with the DMOS subjective quality prediction scores.

More recently, a full reference video quality metric called MOtion-based Video Integrity Evaluation (MOVIE) index was proposed by Seshadrinathan and Bovik [67]. The MOVIE model, which is not standardized, strives to capture the characteristics of the middle temporal (MT) visual area of the visual cortex in the human brain for video quality analysis. Neuroscience studies indicate that the visual area MT is critical for the perception of video quality [87]. The response characteristics of the visual area MT are modeled using separable Gabor filter banks. The model described two indices, namely a Spatial MOVIE index that primarily captures spatial distortions and a Temporal MOVIE index that captures temporal distortions. After applying the Gabor filter banks on both the reference and distorted video sequences, the spatial distortion is captured as a function of difference squared between Gabor coefficients. The error measure is normalized by a masking coefficient, which is defined as a function of the local energy content. For capturing low frequency distortions, a Gaussian filter operating at DC is used and the error measure is computed similar to the one for the Gabor coefficients. Both, the Gabor and Gaussian errors are pooled together to give the spatial error measure for a given pixel. The motion information from optical flow fields of the reference video along with the spatio-temporal Gabor decompositions help in measuring the temporal distortions at each pixel. The frame-level spatial distortion is measured as the ratio of standard deviation to mean of the spatial error over all pixels. Similarly, the frame-level temporal distortion is pooled using the temporal error of all pixels. The spatial error indices are averaged across all frames to provide the Spatial MOVIE index. Similarly, the average of all frame-level temporal error indices is computed, the square-root of which gives the Temporal MOVIE index. The final MOVIE index for the video sequence is computed as the product of these two indices. The performance of the model on the VQEG FRTV Phase 1 dataset was $\mathrm{PCC} = 0.821$, $\mathrm{SROCC} = 0.833$, and $\mathrm{OR} = 0.644$ [67].

2) Pixel Domain: The HVS feature of sensitivity to edges and local changes in luminance is exploited by Hekstra *et al.* [68] to propose the objective video quality model called Perceptual Video Quality Metric (PVQM; also known as the Swisscom/KPN metric). The model uses a linear combination of three distortion indicators, namely *edginess, temporal decorrelation*, and *color error* to measure the perceptual quality. The edginess is computed using a local gradient filter for the luminance signal of both the reference and processed video signal. The normalized change in edge information is computed to account for loss or introduction of sharpness. Hekstra *et al.* claim that the perceived spatial distortion is more profound for frames with low motion content, than for frames with high motion content. The edge error is compensated with the temporal decorrelation factor to account for the perceived spatial distortion. The temporal variability indicator is computed by subtracting the correlation between the current and previous frame from the one for the reference video luminance frames. The processed video signal is not considered in computing the temporal variability as it might be influenced by errors. The normalized color error is computed based on the maximum color saturation of the original and processed video signal. Finally, the video quality rating is obtained as a weighted linear combination of these indicators. The PVQM performance results were based on tests over 26,000 subjective scores generated on 20 different video sequences and processed by

16 different video systems. The results of PVQM were based on training on a medium to high quality video database, that comprised various digital codec distortions, such as H.263 with and without frame repeat, MPEG2, ETSI codecs as well as analog PAL, VHS, and Betacam distortions. The Pearson correlation between subjective quality score (DMOS) and objective quality score produced by the PVQM was observed to be 0.934. In the validations done by VQEG in their Phase 1 study on the objective models of video quality assessments, PVQM was observed to show the highest correlation between subjective and objective quality scores [79].

Lu *et al.* [69] proposed saliency-weighted reduced reference and no reference metrics to measure visual distortions based on visual attention, eye fixation/movement, and the path of vision/retina which are considered the three aspects of perception. For this purpose, Lu *et al.* [69] estimate a perceptual quality significance map (PSQM) to model visual attention and eye fixation/movement, while existing visual quality metrics are adopted to simulate the retina. Thus, the metric by Lu *et al.* [69] mainly integrates the derived PSQM with existing reduced reference and no-reference metrics. Three steps are used to estimate PQSM, namely feature extraction, stimuli integration, and post processing. The feature extraction step is used to extract visual attention related features from an input video sequence and map these into a visual stimuli map. The extracted visual features include relative and absolute motion, luminance, contrast, texture and skin/face features. The stimulus integration step is used to integrate the various visual stimuli into one PQSM by the means of a nonlinear additivity model. Postprocessing is used to better model the eye fixation and movement by representing the saliency locations as localized regions rather than isolated points. The PQSM-based metrics are tested for VQEG data sets using the Spearman and Pearson Correlation coefficients. The obtained results show that integrating the PQSM with existing visual quality metrics can result in an approximately 10% increase in the PCC and SROCC.

In the video quality model proposed by Ong *et al.* [70], [71], the perceptual quality is measured as a function of distortion-invisibility, blockiness, and content fidelity factor. The visibility threshold gives a measure of the maximum amount of distortion that a particular pixel can undergo and still be imperceptible by the human vision. The distortion-invisibility feature is measured as a function of luminance masking, spatial-textural masking, and temporal masking at a particular pixel. The luminance masking factor is deduced based on HVS characteristics to accept distortion when background luminance is above or below a threshold value. Based on the strength of gradients around a pixel in four different directions, the spatial-textural masking factor is deduced. The temporal masking factor is derived as a function of motion content, based on the ability of the HVS to tolerate distortions at a particular pixel location due to large motion. The blockiness is measured as a function of the MSE of 4×4 blocks between the original and distorted video frames. Finally, the content fidelity factor provides a measure of content richness, based on the tendency of the HVS to provide higher subjective scores for vivid content. The content fidelity factor is computed based on the frequencies of pixel values appearing in the original and distorted video frame. The video quality for a given frame is expressed as a product of distortion-invisibility measure, blockiness, and color fidelity. The final video score for the sequence is obtained by computing a weighted sum of scores considering each color component. The test was done using ninety test video sequences that were generated from twelve different CIF and QCIF original video sequences (*Container*, *Coast Guard*, *Japan League*, *Foreman*, *News*, and *Tempete*). The MPEG-4 codec with bit-rates from 24 kbps to 384 kbps, and frame rates from 7.5 Hz to 30 Hz was used. The scores from Double-Stimulus Impairment Scale variant II (DSIS-II) subjective tests performed with 20 subjects were used to assess the performance of the model. When subjective scores were compared with the objective model scores, the Pearson correlation coefficient and Spearman rank-order correlation values were found to lie within a confidence interval of 95%.

Based on the earlier works of Ong *et al.* [70] [71], Nya *et al.* [72] proposed an improved full-reference video quality model. One of the suggested modifications include using a Sobel filter to approximate the gradient of local luminance compared to the complex equations used in [70] and [71]. The block fidelity measure proposed by Ong *et al.* [70], [71] inherently measured blurring artifacts. Also, the contrast loss detection property used in [70], [71] was observed to ignore major structural information if macroblock grid matching is not performed. Nya *et al.* [72] modified the feature point selection method used in [70], [71], where a macroblock of size $k \times k$ was assumed, and incorporated a binary mask that defined regions of interest. As a result, the model was found to account for both tiling effects and distortions effecting block boundaries. The performance assessment was done using MPEG data sets (that included QCIF 10 Hz and 15 Hz, 10 s, 32 kbps and 64 kbps) which were used to benchmark the performance of MPEG-2 and H.26L. Also, five video sequences (QVGA 12.5 Hz, 10 s, variable bit-rate) provided by the Fraunhofer Heinrich-Hertz Institute (HHI) were used. The clip contents consisted of news, sports, monochrome, cartoon, and color movies. The obtained objective quality scores were compared with existing objective video quality metrics, including the NTIA Video Quality General Model [82], and the earlier model proposed by Ong *et al.* [70], [71] in terms of correlation with available DMOS subjective scores. For both the MPEG and HHI videos, the Pearson correlation coefficient was observed to be almost the same as for the NTIA Video Quality General Model [82], but higher than the ones obtained for the Ong *et al.* [70], [71] model and the PSNR. Furthermore, the Spearmen correlation coefficient was observed to be higher for the proposed model compared to the others.

The VSNR metric presented by Chandler and Hemami [73], is essentially a full-reference still-image quality metric but has also shown a promising performance in assessing video quality when applied on a frame-by-frame basis and then averaged. The metric aimed at minimizing the suprathreshold problem in the HVS modeling. The model uses visual masking and visual summation concepts to identify the perceptually detectable distortions. In the case that the distortions are above the threshold of detection, a second stage is applied which operates on properties of perceived contrast and global precedence. These properties

TABLE III
COMPARISON OF HD OBJECTIVE VIDEO QUALITY MODELS

Model	Approach	Test Details	Subj. Model	Performance
Wolf et al. [75]	edge impairment filter	Twelve 30-second video seq. in 1920 × 1080 format. Five DivX Pro encoders, WM9, three MPEG-2 codecs, coded at 2–19 Mbps	SSCQE	PCC = 0.84, RMSE = 9.7
Sugimoto et al. [76]	blockiness, blur measure, edge quality	242 sequences using 12 coding setups consisting of x264 software encoder for H.264 and SONY BDKP-14 E2001 hardware encoder for MPEG-2, coded at 2.0–20 Mbps	ACR with hidden reference (HR)	correlation coefficient of 0.91
Okamoto et al. [77]	PSNR, block distortion and motion blur	HDTV videos encoded using H.264 encoder and MEncoder as decoder	ACR-HR	correlation coefficient of 0.94, PCC = 0.76, RMSE = 0.74
SwissQual VQuad-HD [78]	similarity and difference measures, jerkiness and blockiness measures, fitting to perceptual scale	VQEG HDTV Phase I dataset	ACR-HR, DMOS	PCC = 0.87, RMSE = 0.56

are modeled as Euclidean distances of distortion and contrast and the metric is defined as a simple sum of the distances.

Opticom (www.opticom.de), a firm specializing in developing perceptual voice, audio, and video quality testing products, introduced a proprietary full-reference objective video quality metric called Perceptual Evaluation of Video Quality (PEVQ) [74] based on the PVQM model discussed earlier. The quality evaluation consists of five main stages. The first stage pre-processes both the original and distorted video signals by extracting the region of interest (ROI). The ROI is derived by cropping the actual frame, with a cropping size defined by the video format. These ROI-derived frames are used in subsequent stages. Stage two spatially and temporally aligns the pre-processed video signals. Stages three and four compute four spatial distortion measures, namely (*edginess in luminance, edginess in chrominance, and two temporal variability indicators*), as well as a temporal distortion measure. In particular, a gradient filter is applied on both the luminance and chrominance part of the video signals to obtain the edge information. From the edge information for each frame, the normalized change in edginess for the distorted video signal with respect to the original video signal is computed and averaged over all frames to obtain the edginess in luminance and chrominance. The temporal variability of a frame is defined as the difference of *(i)* the absolute difference between the current and previous frame of the original signal, and *(ii)* the absolute difference between the current and previous frame of the distorted signal. The negative part of the temporal variability measures the new spatial information introduced in the signal, and the positive part of the temporal variability measures the effect of spatial information lost in the signal. The temporal distortion is computed from the amount of frame freezing as well as frame delay or loss information. Stage five uses a sigmoid approach to map the distortions to the DMOS video quality measure, with the mappings defined based on the input video format (QCIF, CIF, or VGA). PEVQ was one of the two best performing methods in the VQEG Multimedia Quality Assessment, Phase I [84], and included as normative model in ITU-T Recommendation J.247 [74].

The other one of the two best performing methods in [84] is a proprietary full-reference metric developed by Psytechnics (www.psytechnics.com). The Psytechnics method consists of three main stages. First, the video registration phase matches frames in the distorted video to frames in the reference video. Second, the perceptual features of the comparison between distorted and reference frame are extracted through several analysis methods, including spatial frequency analysis, edge distortion analysis, blur analysis, block distortion analysis, as well as spatial and temporal distortion analysis. Third, the individual perceptual feature measures are linearly combined with weights determined through an extensive training set to obtain an overall quality prediction DMOS. The model performed well in the VQEG tests, as summarized in Table II, and was included as a normative model in ITU-T Recommendation J.247 [74].

V. OBJECTIVE VIDEO QUALITY MEASUREMENT METHODS FOR HD VIDEO

HDTV systems need higher resolution display screens compared to SDTV systems. For HDTV systems, though the viewing distance will be closer in terms of picture height, the spatial resolution is higher. As a result, approximately the same number of pixels per degree of viewing angle exist for both the HDTV and SDTV systems [75]. However, HDTV has a higher horizontal viewing angle (approximately 30 degrees) when compared to SDTV (12 degrees), which might influence the quality decisions. Also, because of the larger screen size, the eye has to roam around the pictures to track specific objects, and quality degradations that are detected outside this region of immediate attention, will be less perceived when compared to SDTV systems.

Recently, novel models have been proposed for evaluating the perceptual video quality of HD videos. Also, the VQEG has developed validation tests for objective quality metrics applicable to HD video [88]. We review the new HD video quality evaluation methods in this section and summarize the methods in Table III.

Wolf and Pinson [75], performed a study of the performance of the NTIA General Model (discussed in Section IV-B-2) for HDTV video sequences, and measured the degree of accuracy by comparing it with the results of the SSCQE subjective quality rating approach. Twelve video sequences (of both uncompressed and mildly compressed origin, compression ratios ranging from 4:1 to 10:1), each of 30-second duration and shot in 1080i format (1920 × 1080) were considered. To assess the

VQM performance under different conditions, sixteen HDTV video systems were used. Five different encoders (DivX Pro, WM9, 3MBTM MPEG-2, TMPGEnc PlusTM 2.58.44.152 MPEG-2 and MainConceptTM MPEG-2 With Adobe Premiere ProTM version 1.5) were used to generate bit-streams ranging from 2 Mbps to 19 Mbps). The tests indicated that the General VQM Model rating highly correlated with the subjective ratings obtained from the SSCQE. Calibration was used only for the sequences for which transmission errors were introduced in the processed sequences. It was observed that video sequences without errors did not introduce any anomaly in the VQM rating when used without calibration. The Pearson correlation coefficient among the two methods was found to be 0.84 and the Root Mean Square (RMS) error between the best fit line and subjective data scale was found to be 9.7 (on a scale of 0 to 10).

Sugimoto *et al.* [76] proposed a model for evaluating the perceived video quality of HD video considering distortions such as blockiness, the MSE variance in the sequence, temporal PSNR degradation, average power of inter-frame difference in the sequence, average MSE of the blocks having high variance, degradation of lower frequency components, and degradation of higher frequency components. The blockiness feature is derived by using the average of the DC difference between the current 8×8 block, and four adjacent blocks (formed by left, top left, top, and top right blocks). From the MSE error between the original and processed video frames, the MSE variance is computed to assess the coding quality. The temporal PSNR degradation factor for a given frame is measured by subtracting the PSNR of the current frame from the average PSNR of the previous and next frame. Also, the average power of inter-frame differences in the sequence is considered to characterize temporal distortions. From the variance information of average MSE of blocks, the loss of high frequency information (blurring) is assessed. Then, to account for the degradation of low frequency components, the MSE between the original and processed video sequences is considered after initially applying a low-pass filter. For edge quality assessment, a feature extraction procedure similar to the one used for the lower frequency components is followed, but with the lowpass filter replaced with a Laplacian filter. Finally, the video quality is estimated using a weighted sum of all the extracted features. The performance evaluation experiment consisted of 242 sequences, generated using 12 coding setups that included the x264 software encoder for H.264 and the SONY BDKP-E2001 hardware encoder for MPEG-2, coding at 2.0–20 Mbps. The results showed that the model presents a high correlation coefficient of 0.91 when compared with the ACR-HR (absolute category rating with hidden reference) subjective quality model test that is recommended in ITU-T P.910.

Based on their earlier work for PC and mobile services [59], [74], Okamoto *et al.* [77] proposed a full-reference perceptual video quality model for HDTV using fuzzy measures. In the earlier work, the quality was measured as a linear combination of spatial and temporal distortions, based on features such as PSNR, block distortion, and motion blur measures. When this earlier method was applied to HDTV video, it was observed that the characteristic of video quality predicted was non-linear, with different trends in low quality and high quality regions, though a correlation coefficient of 0.87 was achieved. To account for this non-linearity, instead of an additive measure, a fuzzy measure using Choquet integrals is used to measure the video quality. Using the fuzzy measure, the resulting metric was observed to achieve a correlation coefficient of 0.94 with the absolute category rating with hidden reference ACR-HR subjective method for HDTV videos encoded using the H.264 encoder and the MEncoder as decoder. A version of this method, which was developed at NTT, achieved $\mathrm{PCC} = 0.76$ and $\mathrm{RMSE} = 0.74$ for the aggregated VQEG HDTV Phase I dataset [88].

The company SwissQual (www.swissqual.com) has developed a proprietary full-reference HD video quality assessment method called VQuad-HD [78]. VQuad-HD consists of four main components, namely *(i)* analysis of the distribution of local pixel similarities and differences, *(ii)* blockiness analysis, *(iii)* jerkiness analysis, and *(iv)* aggregation of similarity, difference, blockiness, and jerkiness characteristics. VQuad-HD initially lowpass filters and downsamples the original and processed frames from the 1080×1920 pixel HD resolution to the 540×960, 270×480, and 96×128 resolutions. The reference and processed frames at resolution 96×128 are temporally aligned, followed by a spatial alignment. For the resulting aligned frame-pairs, VQuad-HD computes local similarity and difference pixel value measures for local regions of size 13×13 pixels in the 270×480 frames. The form of the distribution of these local similarity and difference measures is characterized through averages computed over prescibed quantiles of their distribution. The blockiness analysis is conducted at the 540×960 resolution to focus on visible edges. Horizontal and vertical edges are identified and averages of subsamples of the horizontal and vertical edges are compared to detect strong block structures. The jerkiness analysis considers the joint impact of display times of frames (which capture temporal impairments, such as pauses or reduced frame rates) and the motion intensity in successive frames. Generally, for a fixed temporal impairment, the jerkiness increases with increasing motion intensity. The VQuad-HD jerkiness measure therefore averages the product of display time and motion intensity, which are both transformed with an S-shaped function that compresses small (imperceptible) values and scales up large (perceptually significant) values [89]. In the aggregation process, similar S-shaped functions with parameters determined through fitting large sets of sample data are used to transform the similarity, difference, blockiness, and jerkiness characteristics to perceptual scales. Furthermore, a time transform is used to reduce the effect of a second degradation occurring soon after a first degradation. VQuad-HD was the best performing full-reference method in the VQEG HD tests achieving $\mathrm{PCC} = 0.87$ and $\mathrm{RMSE} = 0.56$ for the aggregated VQEG HDTV Phase I dataset [88] and as a result is the normative full-reference model in ITU-T Recommendation J.341 [78].

The second best performing full-reference method in the VQEG HD tests is a proprietary method by Tektronix (www.tek.com), which achieved $\mathrm{PCC} = 0.82$ and $\mathrm{RMSE} = 0.65$ for the aggregated VQEG HDTV Phase I dataset. The method incorporates adaptive components for spatial alignment and human visual perception and cognition modeling [90]. The VQEG HD tests also considered a version

TABLE IV
COMPARISON OF PERFORMANCE OF VIDEO QUALITY ASSESSMENT METRICS ON LIVE VIDEO QUALITY DATABASE

Class	Metric	PCC	SROCC	OR	RMSE
Traditional Point-based Metric	PSNR	0.5489	0.5233	0.0200	9.1755
Natural Visual Statistics - Image Quality Metric	SSIM [49]	0.5423	0.5251	0.0333	9.2228
Natural Visual Statistics - Image Quality Metric	MS-SSIM [51]	0.7387	0.7321	0.0067	7.3982
Natural Visual Statistics - Video Quality Metric	VSSIM [50]	0.6058	0.5924	0.0200	8.7337
Natural Visual Statistics - Image Quality Metric	VIF [53]	0.5701	0.5565	0.0400	9.0185
Natural Visual Features - Video Quality Metric	VQM [58]	0.72361	0.7026	0.0133	7.5767
Perceptual, Frequency Domain - Video Quality Metric	MOVIE [67]	0.8116	0.7890	-	-
Perceptual, Pixel Domain - Image Quality Metric	VSNR [73]	0.6884	0.6725	0.0000	7.9616

of PEVQ by Opticomm ($\mathrm{PCC} = 0.63, \mathrm{RMSE} = 0.88$) and a proprietary full-reference model developed by Yonsei University, Korea ($\mathrm{PCC} = 0.76, \mathrm{RMSE} = 0.74$). The Yonsei model relies on edge detection, followed by feature extraction from the edge areas. The degradation of edge areas is measured as edge PSNR, which is refined by additional features [88].

VI. PERFORMANCE COMPARISONS

To examine the performance of a representative set of the surveyed video quality metrics, we quantitatively evaluate state-of-the-art objective quality assessment methods from our classification categories. Specifically, we compare the methods listed in Table IV. It should be noted that the still-image quality metrics shown in Table IV are used to assess the visual video quality by applying these metrics on each video frame separately and then averaging the resulting frame scores.

Currently, the publicly available video databases include the VQEG FRTV Phase I database [91] and the LIVE Video Quality Database [92]. The VQEG FRTV Phase I database was built in 2000. There have been significant advances in video processing technology since then. The LIVE Video Quality Database was recently released in 2009, and includes videos distorted by H.264 compression, as well as videos resulting from simulated transmission of H.264 packetized streams through error prone communication channels. Consequently, we use the more recent LIVE video database.

The LIVE Video Quality Database includes 10 reference videos. The first seven sequences have a frame rate of 25 frames per second (fps), while the remaining three (*Mobile and Calendar, Park Run*, and *Shields*) have a frame rate of 50 fps. In addition, for each reference video, there are 15 corresponding test sequences that were generated using four different distortion processes, namely simulated transmission of H.264 compressed bit streams through error-prone wireless networks and IP networks, H.264 compression, and MPEG-2 compression. All video files have planar YUV 4:2:0 formats and do not contain any headers. The spatial resolution of all videos is 768 × 432 pixels. We include all 150 test sequences in our evaluation. We independently conducted the evaluations of all metrics shown in Table IV, except for MOVIE for which we include the results from [93]. The ASU Image and Video QUality Evaluation SofTware (IVQUEST) [94], [95] was used to test and compare the performance of these metrics using the LIVE Video Quality Database (except for VSSIM, which we implemented and tested as a standalone function as we did not yet integrate it in the current IVQUEST Software Package Release 1.0).

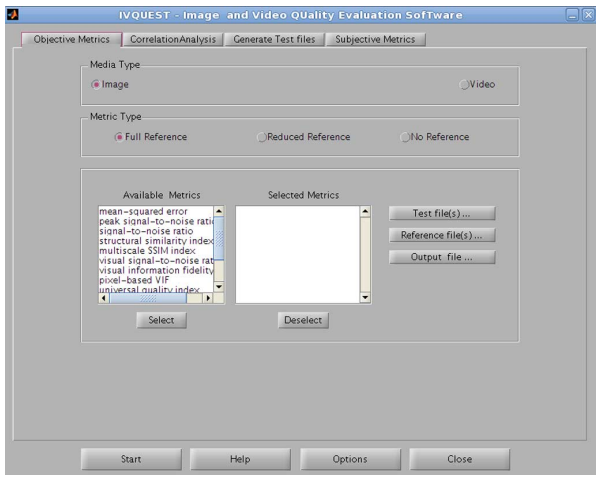


Fig. 4. IVQUEST objective metric view.

The IVQUEST software takes as input the 150 test video sequences from the LIVE Video Quality Database in addition to their corresponding subjective DMOS scores. It enables the user to select the objective quality metrics to be applied to the selected input video sequences. The software can then compute, in a batch processing mode, the results for the selected objective metrics using the input videos. The software can also perform nonlinear regression and correlation analysis on the obtained objective metric results, as recommended in [48], using the input DMOS scores in order to evaluate the performance of the chosen objective quality metrics. The IVQUEST software supports several performance evaluation tools, including the Pearson correlation coefficient (PCC), the Spearman rank order correlation coefficient (SROCC), root-mean-squared error (RMSE), mean absolute error (MAE), and outlier ratio (OR). The PCC and SROCC were computed after performing nonlinear regression on the objective metrics' scores using a four-parameter logistic function as recommended in [48]. In addition, linear rescaling was applied to the SSIM [49], MS-SSIM [51], VSSIM [50], and VIF [53] metrics to facilitate numerical convergence of the nonlinear regression. Figs. 4 and 5 show, respectively, the objective metric selection view and the correlation analysis view of the IVQUEST software. The obtained PCC, SROCC, OR, and RMSE performance results are shown in Table IV.

From Table IV, we observe that the MS-SSIM, VQM, and MOVIE metrics result in the highest PCC and SROCC values as compared to the other metrics, which indicates higher correlation with subjective scores. In addition, the MS-SSIM and VQM metrics have the smallest OR and RMSE values as compared to

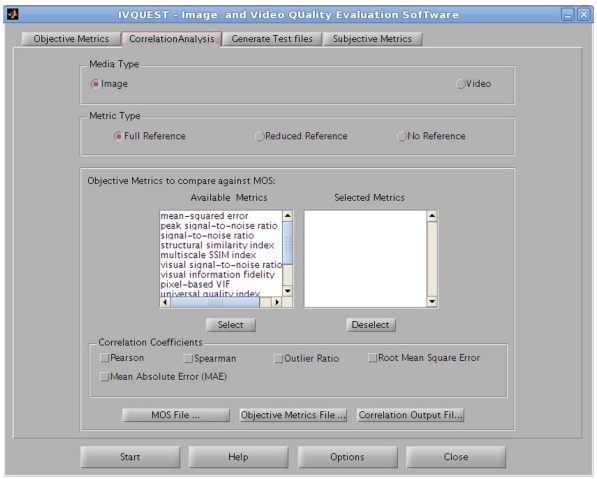


Fig. 5. IVQUEST correlation analysis view.

the other metrics. Therefore, MS-SSIM, VQM, and MOVIE are the best performing image/video quality assessment methods among these six objective quality metrics for the LIVE Video Quality Database. We note that the VSSIM has a significantly higher performance when applied to the VQEG Phase 1 video data set [91], but has poor performance using the more recent LIVE Video Quality Database.

In addition, from Table IV, it can be observed that the full reference still-image quality metric MS-SSIM [51] achieves a performance that is comparable to the state-of-the-art full-reference video quality metrics, such as VQM [58] and MOVIE [67], while outperforming the others, such as VSSIM [50]. Consequently, improved spatio-temporal modeling is needed for video quality assessment as current video quality metrics do not offer improved performance as compared to some existing still-image quality metrics that are applied to video.

VII. Summary and Outlook

Given the growing interest in delivery of multimedia services over wired and wireless networks, perceptual quality measurement has become a very active area of research. With the advent of highly efficient image and video codecs, there is a strong need for metrics being able to measure and quantify transmission and coding quality as perceived by the end-user. In this paper, we have introduced a classification of objective video quality metrics based on their underlying methodologies and approaches for measuring video quality. Within the framework of our classification, we have conducted a comprehensive survey of the proposed full-reference and reduced reference objective video quality metrics. The metrics reviewed in this paper represent important steps towards comprehensive full and reduced reference video quality metrics. We conducted independent performance comparisons and have shown results of popular objective video quality assessment methods with sequences from the LIVE video database.

There are many challenges remaining to be resolved in the field of full-reference and reduced-reference objective video quality assessment methods. There is a wide scope for the development of improved reliable video quality metrics that achieve high performance using a variety of video databases and video content. Developing hybrid methods that combine methods from two or more of our classification categories (e.g., combine statistical and feature based methods), may provide improved results and can be used in developing new metrics in the future. Moreover, extensive comparative analysis experiments will continue to be important for validating the performance of the developed metrics. A reliable perceptual video quality metric will eventually help in benchmarking various video processing techniques. This will require coordinated research efforts in the areas of human vision, color science, and video processing and focused research on quality evaluation of recent image and video codecs, such as H.264. In addition, a more sequenced verification process should be followed as specified in [96] to show meaningful results and to have a common basis for the comparison of various techniques.

Considering the broader field of objective video quality assessment methods, there are many open challenges for full/reduced-reference and no-reference methods. For instance, the existing methods consider any changes from the original sequence as reducing video quality and are thus not suitable for evaluating postprocessing feature improvement mechanisms. Similarly, the effects of scaling the video in the temporal, spatial, or SNR dimension in conjunction with display on a wide range of devices call for new video quality assessment methods. Moreover, the emerging three-dimensional (3D) video will require the design and evaluation of an entirely new class of objective video quality assessment methods. Furthermore, the notion of video quality is currently being broadened to the notion of Quality of Experience (QoE), which encompasses the complete context of the video consumption experience. Objective assessment of the QoE will require a broadening of the video quality assessment methods to capture related parameters influencing the viewer experience.

In order to facilitate the performance evaluation of newly developed quality metrics it is very important that databases with test materials are publicly available. There is currently a shortage of such databases for both image quality evaluation and video quality evaluation. However, this issue is more problematic for video since these require large storage and bandwidth. This issue is even more pronounced for 3D video. Large diverse databases that are shared among researchers would greatly help in conducting sound performance evaluations.

Acknowledgment

The authors thank Tsung-Jung Liu for assisting with computing the VSSIM metric and Milind Gide for assisting with the evaluations with the IVQUEST software. They are grateful to Jens Berger and Silvio Borer of SwissQual for providing insights into the VQuad-HD model. They are grateful to the three anonymous reviewers whose thoughtful comments have helped to significantly improve this article.

References

[1] G. Van der Auwera, P. David, and M. Reisslein, "Traffic characteristics of H.264/AVC variable bit rate video," *IEEE Commun. Mag.*, vol. 46, no. 11, pp. 164–174, Nov. 2008.

[2] M. Isnardi, "Historical overview of video compression in consumer electronic devices," in *Proc. Int. Conf. Consum. Electron.(ICCE)*, 2007, pp. 1–2.

[3] M. Wien, H. Schwarz, and T. Oelbaum, "Performance analysis of SVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1194–1203, Sep. 2007.

[4] B. Ciubotaru and G.-M. Muntean, "SASHA—A quality-oriented handover algorithm for multimedia content delivery to mobile users," *IEEE Trans. Broadcast.*, vol. 55, no. 2, pp. 437–450, Jun. 2009.

[5] J. Monteiro, C. Calafate, and M. Nunes, "Evaluation of the H.264 scalable video coding in error prone IP networks," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 652–659, Sep. 2008.

[6] M. Pinson, S. Wolf, and G. Cermak, "HDTV subjective quality of H.264 vs. MPEG-2, with and without packet loss," *IEEE Trans. Broadcast.*, vol. 56, no. 1, pp. 86–91, Mar. 2010.

[7] F. Speranza, A. Vincent, and R. Renaud, "Bit-rate efficiency of H.264 encoders measured with subjective assessment techniques," *IEEE Trans. Broadcast.*, vol. 55, no. 4, pp. 776–780, Dec. 2009.

[8] T. Wiegand, L. Noblet, and F. Rovati, "Scalable video coding for IPTV services," *IEEE Trans. Broadcast.*, vol. 55, no. 2, pp. 527–538, Jun. 2009.

[9] Cisco, Inc., "Visual networking index: Global mobile data traffic forecast update, 2009–2014," Feb. 2010.

[10] L. Karam, T. Ebrahimi, S. Hemami, T. Pappas, R. Safranek, Z. Wang, and A. Watson, "Introduction to the special issue on visual media quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 189–192, Mar./Apr. 2009.

[11] G. Van der Auwera and M. Reisslein, "Implications of smoothing on statistical multiplexing of H.264/AVC and SVC video streams," *IEEE Trans. Broadcast.*, vol. 55, no. 3, pp. 541–558, Sep. 2009.

[12] N. Staelens, S. Moens, W. Van den Broeck, I. Marien, B. Vermeulen, P. Lambert, R. Van de Walle, and P. Demeester, "Assessing quality of experience of IPTV and video on demand services in real-life environments," *IEEE Trans. Broadcast.*, vol. 56, no. 4, pp. 458–466, Sep. 2010.

[13] L. Guo and Y. Meng, "What is wrong and right with MSE?," in *Proc. 8th Int. Conf. Signal Image Process.*, 2006, pp. 212–215.

[14] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, ITU-R Recommendation BT.500-11.

[15] Subjective video quality assessment [Online]. Available: http://www.acceptv.com

[16] *Subjective Video Quality Assessment Methods for Multimedia Applications*, ITU-T Recommendation-P.910, Sep. 1999.

[17] K. Brunnstrom, D. Hands, F. Speranza, and A. Webster, "VQEG validation and ITU standardization of objective perceptual video quality metrics," *IEEE Signal Process. Mag.*, vol. 26, no. 3, pp. 96–101, May 2009.

[18] A. Takahashi, D. Hands, and V. Barriac, "Standardization activities in the ITU for a QoE assessment of IPTV," *IEEE Commun. Mag.*, vol. 46, no. 2, pp. 78–84, Feb. 2008.

[19] *User Requirements for Objective Perceptual Video Quality Measurements in Digital Cable Television*, ITU-T Recommendation J.1443m, May 2000.

[20] B. Ciubotaru, G.-M. Muntean, and G. Ghinea, "Objective assessment of region of interest-aware adaptive multimedia streaming quality," *IEEE Trans. Broadcast.*, vol. 55, no. 2, pp. 202–212, Jun. 2009.

[21] S. Winkler, A. Sharma, and D. McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications," in *Proc. of Int. Symp. Wireless Personal Multimedia Commun.*, 2001, pp. 553–556.

[22] M. Siller and J. Woods, "QoE in multimedia services transmission," in *Proc. 7th World Multiconf. Systemics, Cybernetics Inf.*, 2003, vol. 7, pp. 74–76.

[23] M. Venkataraman, S. Sengupta, M. Chatterjee, and R. Neogi, "Towards a video QoE definition in converged networks," in *Proc. Int. Conf. Digital Telecommun.*, 2007, pp. 92–97.

[24] K. Yamagishi and T. Hayashi, "Parametric packet-layer model for monitoring video quality of IPTV services," in *Proc. Int. Conf. Commun.*, 2008, pp. 1026–1030.

[25] J. Kim, H. Lee, M. Lee, H. Lee, W. Lyu, and G. Choi, "The QoE evaluation method through the QoS-QoE correlation model," in *Proc. 4th Int. Conf. Networked Comput. Adv. Inf. Manage. (NCM)*, 2008, vol. 2, pp. 719–725.

[26] P. Simoens, S. Latre, B. De Vleeschauwer, W. Van de Meerssche, F. De Turck, B. Dhoedt, P. Demeester, S. Van Den Berghe, and E. Gilon, "Design of an autonomic QoE reasoner for improving access network performance," in *Proc. Int. Conf. Autonomic Autonomous Syst.*, 2008, pp. 233–240.

[27] M. Garcia and A. Raake, "Impairment-factor-based audio-visual quality model for IPTV," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2009.

[28] G. W. Cermak, "Subjective video quality as a function of bit rate, frame rate, packet loss, and codec," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2009.

[29] P. Calyam, P. Chandrasekaran, G. Trueb, N. Howes, D. Yu, Y. Liu, L. Xiong, R. Ramnath, and D. Yang, "Impact of router queuing disciplines on multimedia QoE in IPTV deployments," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2009.

[30] U. Engelke and H.-J. Zepernick, "Perceptual-based quality metrics for image and video services: A survey," in *Next Gen. Internet Netw. (NGI)—3rd EuroNGI Conf. Next Gen. Internet Networks: Design Eng. Heterogeneity* , 2007, pp. 190–197.

[31] H. Cheng and J. Lubin, "Reference-free objective quality metrics for MPEG-coded video," in *Proc. SPIE—Int. Soc. Opt. Eng.*, Mar. 2005, vol. 5666, no. 1, pp. 160–167.

[32] Y. Kawayoke and Y. Horita, "NR objective continuous video quality assessment model based on frame quality measure," in *Proc. Int. Conf. Image Process.*, 2008, pp. 385–388.

[33] M. A. Saad and A. C. Bovik, "Natural motion statistics for no-reference video quality assessment," in *Proc. Int. Workshop Quality Multimedia Experience (QoMEX)*, Jul. 2009.

[34] Y. Tian and M. Zhu, "Analysis and modelling of no-reference video quality assessment," in *Proc. Int. Conf. Computer and Automation Eng.*, 2009, pp. 108–112.

[35] T. Oelbaum, C. Keimel, and K. Diepold, "Rule-based no-reference video quality evaluation using additionally coded videos," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 294–303, Apr. 2009.

[36] C. Keimel, T. Oelbaum, and K. Diepold, "No-reference video quality evaluation for high-definition video," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2009, pp. 1145–1148.

[37] S. Hemami and A. Reibman, "No-reference image and video quality estimation: Applications and human-motivated design," *Signal Process.: Image Commun.*, vol. 25, no. 7, pp. 469–481, Aug. 2010.

[38] S. Olsson, M. Stroppiana, and J. Baina, "Objective methods for assessment of video quality : State of the art," *IEEE Trans. Broadcast.*, vol. 43, no. 4, pp. 487–495, Dec 1997.

[39] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. Hoboken: Wiley, 2005.

[40] H. Wu and K. R. Rao, *Digital Video Image Quality and Perceptual Coding*. Boca Raton: CRC Press, 2005.

[41] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Process.*, vol. 78, no. 2, pp. 231–252, 1999.

[42] S. Rihs, "The influence of audio on perceived picture quality and subjective audio-video delay tolerance," in *MOSAIC Handbook* 1996, pp. 183–187.

[43] T. Virtanen, J. Radun, P. Lindroos, S. Suomi, T. Saamanen, T. Vuori, M. Vaahteranoksa, and G. Nyman, "Forming valid scales for subjective video quality measurement based on a hybrid qualitative/quantitative methodology," in *Proc. SPIE, Vol. 6808, Image Quality Syst. Perform. V*, Jan. 2008.

[44] P. Corriveau, C. Gojmerac, B. Hughes, and L. Stelmach, "All subjective scales are not created equal: The effects of context on different scales," *Signal Process.*, vol. 77, no. 1, pp. 1–9, Aug. 1999.

[45] H. R. Wu, Z. Yu, and B. Qiu, "Multiple reference impairment scale subjective assessment method for digital video," in *Proc. Int. Conf. Digital Signal Process. (DSP)*, 2002, pp. 185–189.

[46] Final report from the video quality experts group on the validation of objective models of video quality assessment, Phase II 2003Video Quality, Experts Group (VQEG).

[47] A. Stuart, K. Ord, and S. Arnold, *Kendall's Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*, 6th ed. Hoboken: Wiley, 2009.

[48] "Final report from the video quality experts group on the validation of objective quality metrics for video quality assessment," Study Group 9, 2000ITU-T, Jun. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

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

[50] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

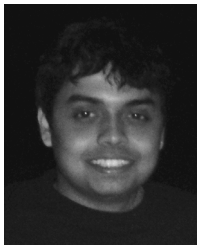
[51] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Conf. Rec. 37th Asilomar Conf. Signals, Syst. Comput.*, 2003, vol. 2, pp. 1398–1402.

[52] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opti. Soc. America A (Optics, Image Sci., Vision)*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[53] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[54] L. Lu, Z. Wang, A. Bovik, and J. Kouloheris, "Full-reference video quality assessment considering structural distortion and no-reference quality evaluation of MPEG video," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2002, vol. 1, pp. 61–64.
[55] A. Shnayderman, A. Gusev, and A. Eskicioglu, "Multidimensional image quality measure using singular value decomposition," in *Proc. SPIE— Int. Soc. Opt. Eng.*, 2003, vol. 5294, no. 1, pp. 82–92.
[56] P. Tao and A. M. Eskicioglu, "Video quality assessment using M-SVD," in *Proc. Int. Soc. Opt. Eng. (SPIE)*, 2007, vol. 6494.
[57] A. Pessoa, A. Falcao, R. Nishihara, A. Silva, and R. Lotufo, "Video quality assessment using objective parameters based on image segmentation," *Soc. Motion Pictures Television Eng. (SMPTE) J.*, vol. 108, no. 12, pp. 865–872, Dec. 1999.
[58] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.
[59] J. Okamoto, T. Hayashi, A. Takahashi, and T. Kurita, "Proposal for an objective video quality assessment method that takes temporal and spatial information into consideration," *Electron. Commun. Japan, Part 1 (Commun.)*, vol. 89, no. 12, pp. 97–108, 2006.
[60] S.-O. Lee and D.-G. Sim, "New full-reference visual quality assessment based on human visual perception," in *Proc. Int. Conf. Consum. Electron. (ICCE)*, 2008, pp. 75–76.
[61] A. Bhat, I. Richardson, and S. Kannangara, "A new perceptual quality metric for compressed video," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2009, pp. 933–936.
[62] F. Lukas and Z. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. Commun.*, vol. 30, no. 7, pp. 1679–1692, Jul. 1982.
[63] C. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of the human visual system," in *Proc. Int. Soc. Opt. Eng. (SPIE)*, 1996, vol. 2668, pp. 450–461.
[64] A. Watson, J. Hu, and J. McGowan, "Digital video quality metric based on human vision," *J. Electron. Imaging*, vol. 10, no. 1, pp. 20–29, Jan. 2001.
[65] F. Xiao, "DCT-based video quality evaluation," Winter 2000.
[66] C. Lee and O. Kwon, "Objective measurements of video quality using the wavelet transform," *Optical Eng.*, vol. 42, no. 1, pp. 265–272, Jan. 2003.
[67] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.
[68] A. Hekstra, J. Beerends, D. Ledermann, F. de Caluwe, S. Kohler, R. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss, "PVQM—A perceptual video quality measure," *Signal Process. Image Commun.*, vol. 17, no. 10, pp. 781–798, Nov. 2002.
[69] Z. Lu, W. Lin, E. Ong, X. Yang, and S. Yao, "PQSM-based RR and NR video quality metrics," in *Proc. Int. Soc. Opt. Eng. (SPIE)*, 2003, vol. 5150, pp. 633–640.
[70] E. Ong, X. Yang, W. Lin, Z. Lu, and S. Yao, "Video quality metric for low bitrate compressed videos," in *Proc. Int. Conf. Image Process.*, 2004, vol. 5, pp. 3531–3534.
[71] E. Ong, W. Lin, Z. Lu, and S. Yao, "Colour perceptual video quality metric," in *Proc. Int. Conf. Image Process.*, 2006, pp. 1172–1175.
[72] P. Ndjiki-Nya, M. Barrado, and T. Wiegand, "Efficient full-reference assessment of image and video quality," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2007, pp. 125–128.
[73] D. Chandler and S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.
[74] *Objective perceptual multimedia video quality measurement in the presence of a full reference*, ITU-T Recommendation J.247, Aug 2008.
[75] M. Pinson and S. Wolf, "Application of the NTIA General Video Quality Metric VQM to HDTV quality monitoring," in *3rd Int. Workshop Video Process. Quality Metrics Consum. Electron. (VPQM-07)*, Jan. 2007 [Online]. Available: http://www.its.bldrdoc.gov/n3/video/documents.htm
[76] O. Sugimoto, S. Naito, S. Sakazawa, and A. Koike, "Objective perceptual picture quality measurement method for high-definition video based on full reference framework," in *Proc. Int. Soc. Opt. Eng. (SPIE)*, 2009, vol. 7242, p. 72421A, (9 pp.).
[77] J. Okamoto, K. Watanabei, A. Hondaii, M. Uchidaiii, and S. Hangaiiv, "HDTV objective video quality assessment method applying fuzzy measure," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2009.
[78] *Objective Perceptual Multimedia Video Quality Measurement of HDTV for Digital Cable Television in the Presence of a Full Reference*, ITU-T Recommendation J.341, Jan. 2011.
[79] "Final report from VQEG on the validation of objective models of video quality assessment," ITU-T Study Group 12 Temporary Document 8 (WP2/12), May 2000.
[80] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[81] A. A. Stocker and E. P. Simoncelli, "Noise characteristics and prior expectations in human visual speed perception," *Nature Neuroscience*, vol. 9, pp. 578–585, 2006.
[82] "Video quality measurement techniques," NTIA report 02-392, 2002 [Online]. Available: http://www.its.bldrdoc.gov/pub/ntia-rpt/02-392/vqm_techniques_v2.pdf
[83] *American National Standard for Telecommunications—Digital Transport of One-Way Video Signals—Parameters for Objective Performance Analysis*, ANSI T1.801.03-1996, 1996.
[84] "Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, Phase I," Video Quality Experts Group (VQEG), 2008.
[85] T. Oelbaum, K. Diepold, and W. Zia, "A generic method to increase the prediction accuracy of visual quality metrics," in *Picture Coding Symp. (PCS)*, 2007.
[86] O. Faugeras, "Digital color image processing within the framework of a human vision model," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 4, pp. 380–393, 1979.
[87] R. T. Born and D. C. Bradley, "Structure and function of visual area MT," *Annu. Rev. Neurosci.*, vol. 28, no. 1, pp. 157–189, 2005.
[88] "Report on the validation of video quality models for high definition video content," Video Quality Experts Group (VQEG), 2010.
[89] S. Borer, "A model of jerkiness for temporal impairments in video transmission," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jun. 2010, pp. 218–223.
[90] K. Ferguson, "An adaptable human vision model for subjective video quality rating prediction among CIF, SD, HD, and e-cinema," in *Proc. 3rd Int. Workshop Video Process. Quality Metrics for Consum. Electron. (VPQM)*, 2007 [Online]. Available: http://enpub.fulton.asu.edu/resp/vpqm2007/
[91] "VQEG FRTV phase 1 database," 2000 [Online]. Available: ftp://ftp.crc.ca/crc/vqeg/TestSequences/
[92] "LIVE video quality database," 2009 [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html
[93] K. Seshadrinathan, R. Soundararajan, A. Bovik, and L. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jun. 2010.
[94] A. Murthy and L. Karam, "IVQUEST- Image and video quality evaluation software," [Online]. Available: http://ivulab.asu.edu/Quality/IVQUEST
[95] A. Murthy and L. Karam, "A MATLAB based framework for image and video quality evaluation," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jun. 2010, pp. 242–247.
[96] C. Keimel, T. Oelbaum, and K. Diepold, "Improving the verification process of video quality metrics," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2009.

Shyamprasad Chikkerur received the B.E. degree in electrical engineering from Visveswaraiah Technological University, India, and M.S. degree in electrical engineering from Arizona State University, Tempe.

He worked as a Software Engineer for Mindtree Consulting Ltd, India. Currently, he is working as Design Engineer for Picture Quality Group at Trident Microsystems Inc., USA. His areas of interest include video processing, video codecs, video quality assessment, and architectures for video processing systems.



Vijay Sundaram (M'10) received the B.Tech. degree in Electronics and Instrumentation Engineering from the National Institute of Technology Tiruchirapalli, India, in 2008, and his M.S. degree in Electrical Engineering from Arizona State University, Tempe, in 2010. His research interests are broadly in video/image processing, coding and quality, specializing in fast algorithms and high throughput architecture designs for video coding applications. He is currently a 3D/Video Design Engineer with the Visual and Parallel Computing Group at Intel Corporation, California, working on video enhancement algorithms for next generation accelerated CPU graphics.

Martin Reisslein received the Dipl.-Ing. (FH) degree from the Fachhochschule Dieburg, Germany, in 1994, and the M.S.E. degree from the University of Pennsylvania, Philadelphia, in 1996. Both in electrical engineering. He received the Ph.D. in systems engineering from the University of Pennsylvania in 1998. During the academic year 1994—1995 he visited the University of Pennsylvania as a Fulbright scholar.

He is an Associate Professor in the School of Electrical, Computer, and Energy Engineering at Arizona State University (ASU), Tempe. From July 1998 through October 2000 he was a scientist with the German National Research Center for Information Technology (GMD FOKUS), Berlin and lecturer at the Technical University Berlin. He currently serves as Associate Editor for the IEEE/ACM Transactions Networking and for Computer Networks. He maintains an extensive library of video traces for network performance evaluation, including frame size traces of MPEG-4 and H.264 encoded video, at http://trace.eas.asu.edu. His research interests are in the areas of multimedia networking, optical access networks, and engineering education.

Lina J. Karam received the Bachelor of Engineering degree in computer and communications engineering from the American University in Beirut in 1989 and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, Atlanta, in 1992 and 1995, respectively.

She is currently a Professor in the School of Electrical, Computer and Energy Engineering (ECEE) at Arizona State University, Tempe, where she directs the Image, Video, and Usability (IVU) and the Real-Time Embedded Signal Processing (RESP) Laboratories. She worked at Schlumberger Well Services (Austin, Texas) on problems related to data modeling and visualization, and in the Signal Processing Department of AT&T Bell Labs (Murray Hill, New Jersey) on problems in video coding during 1992 and 1994, respectively. Prof. Karam is the recipient of an NSF CAREER Award.

Dr. Karam served as the Chair of the IEEE Communications and Signal Processing Chapters in Phoenix in 1997 and 1998. She also served as an Associate Editor of the IEEE Transactions Image Processing from 1999 to 2003 and of the IEEE Signal Processing Letters from 2004 to 2006, as a member of the IEEE Signal Processing Society's Conference Board from 2003 to 2005, and as a member of the IEEE Signal Processing Society's Technical Direction Board from 2008 to 2009. Prof. Karam served as the lead guest editor of the IEEE Journal on Selected Topics in Signal Processing, Special Issue on Visual Media Quality Assessment and as a Technical Program Chair of the 2009 IEEE International Conference on Image Processing. She co-founded the International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM) and the International Workshop on Quality of Multimedia Experience (QoMEX). She currently serves on the editorial boards of the IEEE Trans. Image Processing and the Foundations and Trends in Signal Processing journals. She is the General Chair of the 2011 IEEE Signal Processing Society's DSP and SPE Workshops and of the 2016 IEEE International Conference on Image Processing (IEEE ICIP). She is an elected member of the IEEE Circuits and Systems Society's DSP Technical Committee, the IEEE Signal Processing Society's IVMSP Technical Committee, and the IEEE Signal Processing Society's Education Technical Committee.