SIP (2013), vol. 2, e4, page 1 of 20 © The Authors, 2013.

doi:10.1017/ATSIP.2013.5

OVERVIEW PAPER

Visual quality assessment: recent developments, coding applications and future trends

TSUNG-JUNG LIU[1], YU-CHIEH LIN[1], WEISI LIN[2], AND C.-C. JAY KUO[1]

Research on visual quality assessment has been active during the last decade. In this work, we provide an in-depth review of recent developments in the field. As compared with existing survey papers, our current work has several unique contributions. First, besides image quality databases and metrics, we put equal emphasis on video quality databases and metrics as this is a less investigated area. Second, we discuss the application of visual quality evaluation to perceptual coding as an example for applications. Third, we benchmark the performance of state-of-the-art visual quality metrics with experiments. Finally, future trends in visual quality assessment are discussed.

Keywords: Image quality assessment (IQA), video quality assessment (VQA), perceptual image coding, perceptual video coding

Received 17 February 2012; Revised 31 May 2013

I. INTRODUCTION

During recent years, digital images and videos have played more and more important roles in our work and life because of increasing availability and accessibility. Thanks to the rapid advancement of new technology, people can easily have an imaging device, such as a digital camera, camcorder, and cellular phone, to capture what they see and what happens in daily life. In addition, with the development of social network and mobile devices, photo and video sharing over the Internet becomes much more popular than before. Quality assessment and assurance for digital images and videos in an objective manner have become an increasingly useful and interesting topic in the research community.

In general, visual quality assessment can be divided into two categories. One is subjective visual quality assessment, and the other is objective visual quality assessment. As the name implies, the former is done by humans. It represents the most realistic opinion of humans toward an image or a video, and also the most reliable measure of visual quality among all available means (if the pool of subjects is sufficiently large and the nature of the circumstances allows such assessments).

For subjective evaluation of visual quality, the tests can be performed with the methods defined in [20, 23]: (a) pair comparison (PC); (b) absolute category rating (ACR); (c) degradation category rating (DCR) (also called double-stimulus impairment scale (DSIS)); (d) double-stimulus continuous quality scale (DSCQS); (e) single-stimulus continuous quality evaluation (SSCQE); (f) simultaneous double-stimulus for continuous evaluation (SDSCE). We have presented these methods in the Appendix for easy reference.

In general, methods (a)–(c) above can be used in multimedia applications. Television pictures can be evaluated with methods (c)–(f). In all these test methods, visual quality ratings evaluated by test subjects are then averaged to obtain the mean opinion score (MOS). In some cases, difference mean opinion score (DMOS) is used to represent the mean of differential subjective score instead of MOS.

However, the subjective method is time-consuming, and not applicable for real-time processing because the test has to be performed carefully in order to obtain meaningful results. Moreover, it is not feasible to have human intervention with in-loop and on-service processes (such as video encoding, transmission, etc.). Thus, most research has been focused on automatic assessment of quality for an image or a video.

This paper aims at an overview and discussion of the latest research in the area of objective quality evaluation of visual signal (both image and video). There have been a few good survey papers in this area before, such as [35, 56, 105]. Our current work has several new contributions. First, we put an equal emphasis on image and video quality assessment. Video quality assessment is a rapidly growing field and has progressed a lot in the last 3–4 years. The recent developments have not been well covered in the existing

[1]Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA. Phone: +1 213 740 4658.
[2]School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore

Corresponding author:
C.-C. Jay Kuo
Email: cckuo@sipi.usc.edu

survey papers. Here, we provide the most updated results in this field. Second, we have an in-depth discussion on the application of visual quality assessment to perceptual image/video coding, which is one of the most researched areas in applications. Third, we benchmark the performance of several state-of-the-art quality metrics for both images and videos with appropriate databases and experiments. Finally, future trends in visual quality assessment are discussed.

The rest of the paper is organized as follows. In Section II, the classification of objective quality assessment methods will be presented. Recent developments, applications, and publicly available databases in image quality assessment (IQA) will be examined in Section III, whereas those in video quality assessment (VQA) are to be introduced in Section IV. We follow the similar format of writing for images and videos respectively, for readers' easy reading, reference and comparison. Section V will present performance comparison for some recent popular visual quality metrics. Then, we will point out several possible future trends for visual quality assessment in Section VI. Finally, the conclusion will be drawn in Section VII.

II. CLASSIFICATION OF OBJECTIVE VISUAL QUALITY ASSESSMENT METHODS

There are several popular ways to classify the visual quality assessment methods [35, 56, 105]. In this section, we present two possibilities of classification to facilitate the presentation and understanding of the related problems, the existing solutions (taking into account the most recent developments), and future trends.

A) Classification based on the availability of reference

The classification depends on the availability of original (reference) image/video. If there is no reference signal available for the distorted (test) one to compare with, then a quality evaluation method is termed as a **no-reference (NR)** one [64]. The current NR methods [74, 90] do not perform well in general because they judge the quality solely based on the distorted medium and without any reference available.

If the information of the reference medium is partially available, e.g., in the form of a set of extracted features, then this is the so-called **reduced-reference (RR)** method [78]. Since the extracted partial reference information is much sparser than the whole reference, the RR approach can be used in a remote location (e.g., the relay site and receiving end of transmission) with reasonable bandwidth overheads to achieve better results than the NR method, or in a situation where the reference is available (such as a video encoder) to reduce the computational requirement (especially in repeated manipulation and optimization).

The last one is the **full-reference (FR)** method (e.g., [96]), as the opposite of the NR method. As the name suggests, an FR metric needs the complete reference medium to assess the distorted (test) medium. Since it has full information about the original medium, it is expected to have the best quality prediction performance. Most existing quality assessment schemes belong to this category, and can be usually used in image and video coding. We will discuss more in Sections III and IV.

B) Classification based upon methodology for assessment

The first type in this classification is **image/video fidelity metrics**, which operate based only on direct accumulation of errors and therefore are usually FR. Mean-squared error (MSE) and peak signal-to-noise ratio (PSNR) are two representatives in this category. Although being the simplest and still widely used, such a metric is often not a good reflection of perceived visual quality if the distortion is not additive.

The second type is **human visual system (HVS) model-based metrics**, which typically employ a frequency-based decomposition, and take into account various aspects of the HVS. This can include modeling of contrast and orientation sensitivity, spatial and temporal masking effects, frequency selectivity and color perception. Owing to the complexity of the HVS, these metrics can become very complex and computationally expensive. Examples of the work following this framework include the works in [32, 43, 62, 89], perceptual distortion metric (PDM) [104], the continuous VQM in [66], and the scalable wavelet-based video distortion index [65]. Recently, a new strategy to measure image quality, called most apparent distortion (MAD) [48], also belongs to this category.

Signal structure (information or other feature)-based metrics are the third type of metrics. Some of them quantify visual fidelity based on the assumption that a high-quality image or video is the one whose structural content, such as object boundaries or regions of high entropy, most closely matches that of the original image or video [84, 85, 96]. Other metrics of this type are based on the assumption that the HVS understands an image mainly through its low-level features. Hence, image degradations can be perceived by comparing the low-level features between the distorted and the reference images. The latest work is called feature-similarity (FSIM) index [108]. We will discuss in more detail on this type of metric in Section III.

The fourth type in the classification is **packet-analysis-based metrics**. This type of metric focuses on assessment of the impact caused by network impairments on visual quality. It is usually based on the parameters extracted from the transport stream to measure the quality loss. It also has the advantage of measuring the quality of several image/video streams in parallel. Lately, this type of metric has become more popular because of increasing video delivery service over networks, such as IPTV or Internet streaming. One example of such metrics is the V-Factor [105]. The details about this metric will be introduced in Section IV.

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

The last type of metric is the emerging **learning-oriented metrics**. Some recent works are [57–59, 63, 68, 70, 87]. Basically, it extracts specific features from the image or video, and then uses the machine learning approach to obtain a trained model. Finally, the trained model is used to predict the perceived quality of images/videos. The obtained experimental results are quite promising, especially for multi-metric fusion (MMF) approach [57, 59] that uses the major existing metrics as the components for the learnt model. The MMF is expected to outperform all the existing metrics as the fusion-based approach to allow the combination of merits from each metric.

III. RECENT DEVELOPMENTS IN IQA

A) Image quality databases

Databases with subjective data facilitate metric development and benchmarking, as the ground truth and source of inspiration. There are a number of publicly available image quality databases, including LIVE [9], TID2008 [15], CSIQ [2], IVC [7], IVC-LAR [8], Toyoma [16], WIQ [19], A57 [1], and MMSP 3D Image [12]. We will give a brief introduction for each database below.

LIVE Image Quality Database has 29 reference images (also called source reference circuits (SRC)) and 779 test images, including five distortion types – JPEG2000, JPEG, white noise in the RGB components, Gaussian blur, and transmission errors in the JPEG2000 bitstream using a fast-fading Rayleigh channel model. The subjective quality scores provided in this database are DMOS, ranging from 0 to 100.

Tampere Image Database 2008 (TID2008) has 25 reference images and 1700 distorted images, including 17 types of distortions and four different levels for each type of distortion. Hence, there are 68 test conditions (also called hypothetical reference circuits (HRC)). MOS is provided in this database, and the scores range from 0 to 9.

Categorical Image Quality (CSIQ) Database contains 30 reference images, and each image is distorted using six types of distortions – JPEG compression, JPEG2000 compression, global contrast decrements, additive Gaussian white noise, additive Gaussian pink noise, and Gaussian blurring – at 4–5 different levels, resulting in 866 distorted images. The score ratings (0–1) are reported in the form of DMOS.

IVC Database has 10 original images and 235 distorted images, including four types of distortions – JPEG, JPEG2000, locally adaptive resolution (LAR) coding, and blurring. The subjective quality scores provided in this database are MOS, ranging from 1 to 5.

IVC-LAR Database contains eight original images (four natural images and four art images) and 120 distorted images, including three distortion types – JPEG, JPEG2000, and LAR coding. The subjective quality scores provided in this database are MOS, ranging from 1 to 5.

Toyoma Database has 14 original images and 168 distorted images, including two types of distortions – JPEG and JPEG2000. The subjective scores in this database are MOS, ranging from 1 to 5.

Wireless Imaging Quality (WIQ) Database has seven reference images and 80 distorted images. The subjective quality scores used in this database are DMOS, ranging from 0 to 100.

A57 Database has three original images and 54 distorted images, including six distortion types – quantization of the LH subbands of a five-level DWT of the image using the 9/7 filters, additive Gaussian white noise, JPEG compression, JPEG2000 compression, JPEG2000 compression with Dynamic Contrast-Based Quantization (DCQ), and Gaussian blurring. The subjective quality scores used for this database are DOMS, ranging from 0 to 1.

MMSP 3D Image Quality Assessment Database contains stereoscopic images with a resolution of 1920 × 1080 pixels. Various indoor and outdoor scenes with a large variety of colors, textures, and depth structures have been captured. The database contains 10 scenes. Seventeen subjects participated in the test. For each of the scenes, six different stimuli have been considered corresponding to different camera distances (10, 20, 30, 40, 50, and 60 cm).

To make a clear comparison among these databases, we list important information for each database in Table 1.

B) Major IQA metrics

As mentioned earlier, the simplest and most widely used image quality metrics are MSE and PSNR because they are easy to calculate and are also mathematically convenient in the optimization sense. However, they often correlate poorly with subjective visual quality [95].

Hence, researchers have done a lot of work to include the characteristics of the HVS to improve the performance of quality prediction. The noise quality measure (NQM) [33], PSNR-HVS-M [79], and the visual signal-to-noise ratio (VSNR) [27] are several representatives in this category.

NQM (FR, HVS model-based metric), which is based on Peli's contrast pyramid [77], takes into account the following:

(1) variation in contrast sensitivity with distance, image dimensions, and spatial frequency;
(2) variation in the local luminance mean;
(3) contrast interaction between spatial frequencies; and
(4) contrast masking effects.

It has been demonstrated that the nonlinear NQM is a better measure of additive noise than PSNR and other linear quality measures [33].

PSNR-HVS-M (FR, HVS model-based metric) is a still image quality metric that takes into account contrast sensitivity function (CSF) and between-coefficient contrast masking of DCT basis functions. It has been shown that PSNR-HVS-M outperforms other well-known reference-based quality metrics and demonstrated high correlation with the results of subjective experiments [79].

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

Table 1. Comparison of image quality databases.

Database	Year	SRC (no. of reference images)	HRC (no. of test conditions)	Total no. of test images	Subjective Testing Method	Subjective score	Applications and merits
IVC	2005	10	25	235	DSIS	MOS (1–5)	For testing IQA metrics on images having compression distortions
LIVE	2006	29	27	779	ACR	DMOS (0–100)	For testing IQA metrics on images having compression distortions, transmission distortions, and acquisition distortions
A57	2007	3	18	54	–	DMOS (0–1)	For testing IQA metrics on images having compression distortions and acquisition distortions
Toyoma	2008	14	12	168	ACR	MOS (1–5)	For testing IQA metrics on images having compression distortions
TID2008	2008	25	68	1700	Proprietary	MOS (0–9)	For testing IQA metrics on images having compression distortions, transmission distortions and acquisition distortions
CSIQ	2009	30	29	866	Proprietary	DMOS (0–1)	For testing IQA metrics on images having compression distortions, transmission distortions and acquisition distortions
IVC-LAR	2009	8	15	120	DSIS	MOS (1–5)	For testing IQA metrics on images having compression distortions
WIQ	2009	7	–	80	DSCQS	DMOS (0–100)	For testing IQA metrics on images having transmission distortions
MMSP 3D Image	2009	9	6	54	SSCQE	MOS (0–100)	For testing images on 3D Quality of Experience (QoE)

(Notes: '-' Means no information available; 'proprietary' means the testing method is designed by the authors, not in [23] and [20].)

VSNR (FR, HVS model-based metric) is a metric computed by a two-stage approach [27]. In the first stage, contrast thresholds for detection of distortions in the presence of natural images are computed via wavelet-based models of visual masking and visual summation in order to determine whether distortions in the distorted image are visible. If the distortions are below the threshold of detection, the distorted image is claimed to be of perfect visual quality. If the distortions are higher than a threshold, a second stage is applied, which operates based on the visual property of perceived contrast and global precedence. These two properties are modeled as Euclidean distances in distortion-contrast space of a multi-scale wavelet decomposition, and the final VSNR is obtained by linearly summing these distances.

However, the HVS is a nonlinear and highly complicated system, and most models so far are only based on quasi-linear or linear operators. Hence, a different framework was introduced, based on the assumption that a measurement of structural information change should provide a good approximation to perceived image distortion. **Structural similarity (SSIM)** index (FR, signal structure-based metric) [96] is the most well-known one in this category.

Suppose two image signals $\mathbf{x}$ and $\mathbf{y}$, and let μ_x, μ_y, σ_x^2, σ_y^2, and σ_{xy} be the mean of $\mathbf{x}$, the mean of $\mathbf{y}$, the variance of $\mathbf{x}$, the variance of $\mathbf{y}$, and the covariance of $\mathbf{x}$ and $\mathbf{y}$ respectively. Wang *et al.* [96] define the luminance, contrast, and structure comparison measures as follows:

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}, \quad c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2},$$
$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{1}$$

where the constants C_1, C_2, and C_3 are included to avoid instabilities when $\mu_x^2 + \mu_y^2$, $\sigma_x^2 + \sigma_y^2$, and $\sigma_x\sigma_y$ are very close to zeros. Finally, they combine these three comparison measures and name the resulting similarity measure between image signals $\mathbf{x}$ and $\mathbf{y}$ as

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^\alpha \cdot [c(\mathbf{x},\mathbf{y})]^\beta \cdot [s(\mathbf{x},\mathbf{y})]^\gamma, \tag{2}$$

where $\alpha > 0$, $\beta > 0$, and $\gamma > 0$ are the parameters used to adjust the relative importance of these three components. In order to simplify the expression, set $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$. This results in a specific form of the SSIM index between image signals $\mathbf{x}$ and $\mathbf{y}$:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{3}$$

However, the standard SSIM defined above is only a single-scale method. To be able to consider image details at different resolutions (we do not know the right object sizes in general), **a multi-scale SSIM (MS-SSIM)** (FR, signal structure-based metric) [101] is adopted. Taking the reference and distorted image signals as the input, the system iteratively applies a low-pass filter and down-samples the filtered image by a factor of two. The original image is labeled as scale 1, and the highest scale as M, which is obtained after $M - 1$ iterations; at the j-th scale, the contrast comparison and the structure comparison are calculated and denoted as $c_j(\mathbf{x},\mathbf{y})$ and $s_j(\mathbf{x},\mathbf{y})$, respectively. The luminance comparison is computed only at scale M and denoted as $l_M(\mathbf{x},\mathbf{y})$. The overall SSIM evaluation is obtained by combining the

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

measurement at different scales using

$$\text{MS-SSIM}(\mathbf{x},\mathbf{y}) = [l_M(\mathbf{x},\mathbf{y})]^{\alpha_M} \prod_{j=1}^{M} [c_j(\mathbf{x},\mathbf{y})]^{\beta_j} [s_j(\mathbf{x},\mathbf{y})]^{\gamma_j}. \quad (4)$$

Similarly, exponents α_M, β_j and γ_j are used to adjust the relative importance of different components. As the simplest parameter selection, $\alpha_j = \beta_j = \gamma_j$ for all j's. In addition, normalization is performed for the cross-scale settings such that $\sum_{j=1}^{M} \gamma_j = 1$.

Since SSIM is sensitive to relative translations, rotations, and scalings of images [95], complex-wavelet SSIM (CW-SSIM) [100] has been developed. The CW-SSIM is locally computed from each subband, and then averaged over space and subbands, yielding an overall CW-SSIM index between the original and the distorted images. The CW-SSIM method is robust with respect to luminance changes, contrast changes, and translations [100].

Afterward, some researchers have tried to propose a new metric by modifying SSIM, such as three-component weighted SSIM (3-SSIM) [51], and information content weighted SSIM (IW-SSIM) [98]. They are all based on the similar strategy to assign different weightings to the SSIM scores.

Another metric based on the information theory to measure image fidelity is called **information fidelity criterion (IFC)** (FR, signal information-extracted metric) [85]. It was later extended to **visual information fidelity (VIF)** metric (FR, signal information-extracted metric) [84]. The VIF attempts to relate signal fidelity to the amount of information that is shared between two signals. The shared information is quantified using the concept of mutual information. The reference image is modeled by a wavelet domain Gaussian scale mixture (GSM), which has been shown to model the non-Gaussian marginal distributions of the wavelet coefficients of natural images effectively, and also capture the dependencies between the magnitudes of neighboring wavelet coefficients. Therefore, it brings good performance to the VIF index over a wide range of distortion types [86].

Reduced-reference image quality assessment (RRIQA) (RR, signal feature-extracted metric) is proposed in [52]. The authors use GSM statistical model of image wavelet coefficients to compute a divisive normalization transform (DNT) for images. Then, they evaluate the image quality based on the comparison between features extracted from the DNT of reference and distorted images. The proposed RR approach has improved performance and even works better than FR PSNR in LIVE Image Quality Database.

In [39], multi-scale geometric analysis (MGA) is used to decompose images and extract features to model the multi-channel structure of HVS. Moreover, several transforms (e.g., wavelet, curvelet, bandelet, and contourlet) are also utilized to capture different kinds of geometric information of images. CSF is used to weight the coefficients obtained by the MGA. Next, Just Noticeable Difference (JND) is applied to produce a noticeable variation. Finally, the quality of the distorted image is obtained by comparing the normalized histogram between the distorted image and reference one. In addition to the good consistency with human subjective evaluation, this **MGA-based IQA** (RR, signal feature-extracted metric) also has the advantage of using low data rate to represent features.

Ferzli *et al.* [36] proposed an objective image sharpness metric, called **Just Noticeable Blur Metric (JNBM)** (NR, HVS model-based metric). They claimed the just noticeable blur (JNB) is a function of local contrast and can be used to derive an edge-based sharpness metric with probability summation model over space. The experiment results showed this method can successfully predict the relative amount of sharpness/blurriness in images, even with different scenes.

In [30], the authors presented a method for IQA by combining the features obtained from the computation of mean and ratio of edge blurriness and noise (MREBN). The proposed metric **MREBN** (NR, signal feature-extracted metric) has high correlation with subjective quality scores. They also claimed the low computational load of the model because of linear combination of the features obtained.

In [48], Larson and Chandler suggested that a single strategy may not be sufficient to determine the image quality. They presented a quality assessment method, called **most apparent distortion (MAD)** (FR, HVS model-based metric), which can model two different strategies. First, they used local luminance and contrast masking to estimate detection-based perceived distortions in high-quality images. Then changes in the local statistics of spatial-frequency components are used to estimate the appearance-based perceived distortions in low quality images. In the end, the authors showed that combining these two strategies can predict subjective ratings of image quality well.

FSIM (FR, signal feature-extracted metric) [108] is a recently developed image quality metric, which compares the low-level feature sets between the reference image and the distorted image based on the fact that the HVS understands an image mainly according to its low-level features. Phase congruency (PC) is the primary feature to be used in computing FSIM. Gradient magnitude (GM) is the second feature to be added in FSIM metric because PC is contrast invariant and contrast information also affects the HVS' perception of image quality. Actually, in the FSIM index, similarity measures for PC and GM all follow the same formula as in the SSIM metric.

More recently, we proposed a **multi-metric fusion (MMF)** (FR, learning-oriented metrics) approach for visual quality assessment [57, 59]. This method is motivated by the observation that no single metric can give the best performance scores in all situations. To achieve MMF, a regression approach is adopted. First, we collected a large number of image samples, each of which has a score labeled by human observers and scores associated with different quality metrics. The new MMF score is set to be the nonlinear combination of scores obtained by multiple existing metrics (including SSIM [96], MS-SSIM [101], VSNR [27], IFC [85], VIF [84], PSNR, PSNR-HVS [34], NQM [33], FSIM [108], and MAD [48]) with suitable weights via a training process.

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

Table 2. Classification of IQA models based on reference availability and assessment methodology.

IQA model	Reference availability	Assessment methodology	Remarks (strength and weakness)
PSNR	FR	Image fidelity	• Simple • Low correlation
NQM	FR	HVS model	• Better measure of additive noise than PSNR • 80% correlation to visual results
PSNR-HVS-M	FR	HVS model	• Incorporate CSF model • 98% correlation with subjective scores
VSNR	FR	HVS model	• Low computational complexity and memory requirements • Accommodate different viewing conditions • 88.9% correlation with subjective scores in LIVE database
SSIM	FR	Signal structure	• Easy to implement • Good correlation with subjective scores
MS-SSIM	FR	Signal structure	• Incorporate image details at different resolutions • Better correlation with subjective scores than SSIM
IFC	FR	Signal structure	• Use mutual information to quantify signal fidelity • Better correlation with subjective scores than SSIM
VIF	FR	Signal structure	• Use mutual information to quantify signal fidelity • Better correlation with subjective scores than IFC
RRIQA	RR	Signal structure	• Better performance than PSNR
MGA-based IQA	RR	Signal structure	• Good consistency with subjective scores • Low data rate to represent features
JNBM	NR	HVS model	• Can predict the relative amount of sharpness/blurriness in images
MREBN	NR	Signal structure	• Good correlation with subjective scores • Low computation load
FSIM	FR	Signal structure	• Use low-level features • Very good correlation with subjective scores
MAD	FR	HVS model	• Combine two different strategies to predict visual quality • Good correlation with subjective scores
MMF	FR	Learning-oriented	• Use machine learning to automatically fuse the scores from multiple quality metrics • Very high correlation with subjective scores • Can incorporate new IQA metrics

We also term it as context-free MMF (CF-MMF) because it does not depend on image contexts. Furthermore, we divide image distortions into several groups and perform regression within each group, which is called context-dependent MMF (CD-MMF). One task in CD-MMF is to determine the context automatically, which is achieved by a machine learning approach. It is shown by experimental results that the proposed MMF metric outperforms all existing metrics by a significant margin.

Table 2 summarizes the IQA models that we have mentioned so far and the corresponding classifications based on reference availability and assessment methodology; we have also commented on the strength and weakness of the models under discussion in the table.

C) Application in perceptual image coding

IQA metrics are widely exploited for image coding. Different metrics, such as SSIM [28, 96] and VIF [84] are used to improve the perceptual performance of JPEG and JPEG2000 compression and provide feedback to rate-control algorithms. In other words, the concept of perceptual image coding is to assess the quality of the target image by using IQAs and then apply the index to improve coding efficiency. Each IQA reflects specific features. Thus, choosing the perceptual model is based on the need of specific application or codec. Coding distortion can be approximated from the extracted perceptual features and used to guide an image coder.

Yim and Bovik [107] analyzed the blockiness of compressed JPEG images. The proposed metric index focuses on discrete cosine transformed and quantized images. It has been shown that the blocking effect can be assessed by using the quality metric which detects differences of the neighborhoods of the target block. The blocking effect factor (BEF) is defined by the difference of the mean boundary pixel squared difference and the mean non-boundary pixel squared difference. The mean-squared error including the blocking effect (MSE-B) is calculated from the corresponded BEF and MSE and leads to peak signal-to-noise ratio including the blocking effect (PSNR-B). The PSNR-B can quantify the blocking effect in a boundary of macroblocks. Moreover, this can help to develop H.264/AVC de-blocking filters.

Hontzsch and Karam [41] presented a locally adaptive perceptual image coder, which optimizes the bit allocation of the targeted distortion type. The algorithm starts from extracting visual properties adaptively based on the local image features. It decomposes data into discrete cosine transform (DCT) coefficients, which are fed to the

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

perceptual model to generate perceptual properties. These properties are used to compute the local distortion adaptively and result in local distortion sensitivity profiles. The thresholds, which are derived from the profiles, reflect the characteristics of local image data. Two visual phenomena, contrast sensitivity dependent on background luminance and contrast masking, are modeled to generate the thresholds. For contrast sensitivity, the threshold is defined related to the luminance of the background to verify the sensitivity of the eye under the condition of the background. For contrast masking adjustment, contrast masking pertains to the visual change. The masker signal is in the form of the DCT subband coefficients of the input image comparing to the quantization error. Thus, the quantization step size is calculated from the threshold in order to achieve the target bitrate.

Rehman and Wang [80] addressed the practical use of SSIM. Instead of fully accessing the original image, reduced reference technique only uses partial information. The first step of the algorithm is the multi-scale multi-orientation DNT which extracts the neural features of the biological HVS. DNT coefficient distribution is parameterized and provides needed partial information of the reference image. This information can be used to define the distortion of the compressed image and reflect the SSIM value of the images. The proposed reduced reference version of SSIM shows linear relationship to the full reference version in specific circumstances. The application of the algorithm does not only measure the SSIM but also repair some distortions.

Besides VIF, other approaches are taken to JPEG2000. Tan *et al.* [88] proposed an image coder based on the just-noticeable distortion model which considers a variety of perceptual aspects. The algorithm is developed from a monochromatic vision model to a color image one. The monochromatic contrast gain control (CGC) model includes spatial masking, orientation masking and contrast sensitivity. The luminance and chromatic parts are modeled by the CGC. The distortion metric is designed to estimate perceptual error and applied to replace MSE which is used in the cost function in embedded block coding with optimal truncation (EBCOT). The 14 parameters in the metric are optimized with a two tiered approach. One calculates the parameter set recursively; the other fine-tunes the parameter set via algorithmic optimization.

SSIM is also exploited in JPEG2000. Richter *et al.* [81] proposed a JPEG encoder based on optimal Multi-scale SSIM (MS-SSIM) [101]. Efforts are made to modify MS-SSIM in order to be embedded to the encoder. The first step of the algorithm is trying to modify MS-SSIM to the logarithmic form. The contrast and the structure part of the index can be expressed by the reconstruction error; the luminance part is ignored due to its minor effect. The final term of the index can be computed by utilizing the results from EBCOT and wavelet decomposition process. Thus, the implementation integrates MS-SSIM into a JPEG2000 encoder.

IV. RECENT DEVELOPMENTS IN VQA

A) Video quality databases

To our knowledge, there are nine public video quality databases available, including VQEG FRTV-I [17], IRCCyN/IVC 1080i [5], IRCCyN/IVC SD RoI [6], EPFL-PoliMI [4], LIVE [10], LIVE Wireless [11], MMSP 3D Video [13], MMSP SVD [14], and VQEG HDTV [18]. We will briefly introduce them below.

VQEG FR-TV Phase I Database is the oldest public database on video quality applied to MPEG-2 and H.263 video with two formats: 525@60 Hz and 625@50 Hz in this database. The resolution for video sequence 525@60 Hz is 720 × 486 pixels and 720 × 576 pixels for 625@50 Hz. The video format is 4:2:2. The subjective quality scores provided are DMOS, ranging from 0 to 100.

IRCCyN/IVC 1080i Database contains 24 contents. For each content, there is one reference and seven different compression rates on H.264 video. The resolution is 1920 × 1080 pixels, the display mode is interleaving and the field display frequency is 50 Hz. The provided subjective quality scores are MOS, ranging from 1 to 5.

IRCCyN/IVC SD RoI Database contains six reference videos and 14 HRCs (i.e., 84 videos in total). The HRCs are H.264 coding with or without error transmission simulations. The contents of this database are SD videos. The resolution is 720 × 576 pixels, the display mode is interleaving, and the field display frequency is 50 Hz with MOS from 1 to 5.

EPFL-PoliMI Video Quality Assessment Database contains 12 reference videos (6 in CIF, and 6 in 4CIF), and 144 distorted videos, which are encoded with H.264/AVC and corrupted by simulating the packet loss due to transmission over an error-prone network. For CIF, the resolution is 352 × 288 pixels, and frame rate is 30 fps. For 4CIF, the resolution is 704 × 576 pixels, and frame rates are 30 fps and 25 fps. For each of the 12 original H.264/AVC videos, they have generated a number of corrupted ones by dropping packets according to a given error pattern. To simulate burst errors, patterns have been generated at six different packet-loss rates (PLR) and two channel realizations have been selected for each PLR.

LIVE Video Quality Database [83] includes 10 reference videos. All videos are 10 s long, except for Blue Sky. The Blue Sky sequence is 8.68 s long. The first seven sequences have a frame rate of 25 fps, while the remaining three (Mobile & Calendar, Park Run, and Shields) have a frame rate of 50 fps. There are 15 test sequences from each of the reference sequences using four different distortion processes – simulated transmission of H.264 compressed videos through error-prone wireless networks and through error-prone 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.

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

Table 3. Comparison of video quality databases.

Database	Year	SRC (no. of reference videos)	HRC (no. of test conditions)	Total no. of test videos	Subjective Testing Method	Subjective score	Applications and merits
VQEG FR-TV-I	2000	20	16	320	DSCQS	DMOS (0–100)	For testing VQA metrics on videos having compression distortions
IRCCyN/IVC 1080i	2008	24	7	192	ACR	MOS (1–5)	For testing VQA metrics on videos having compression distortions
IRCCyN/IVC SD RoI	2009	6	14	84	ACR	MOS (1–5)	For testing VQA metrics on videos having compression distortions and transmission distortions
EPFL-PoliMI	2009	16	9	165	ACR	MOS (0–5)	For testing VQA metrics on videos having compression distortions and transmission distortions
LIVE	2009	10	15	150	ACR	DMOS (0–100)	For testing VQA metrics on videos having compression distortions and transmission distortions
LIVE wireless	2009	10	16	160	SSCQE	DMOS (0–100)	For testing VQA metrics on videos having compression distortions and transmission distortions
MMSP 3D video	2010	6	5	30	SSCQE	MOS (0–100)	For testing videos on 3D quality of experience (QoE)
MMSP SVD	2010	3	24	72	PC	MOS (0–100)	For testing VQA metrics on videos having compression distortions and transmission distortions
VQEG HDTV	2010	45	15	675	ACR	MOS (0–5), DMOS (1–5)	For testing VQA metrics on videos having compression distortions and transmission distortions

LIVE Wireless Video Quality Assessment Database has 10 reference videos, and 160 distorted videos, which focus on H.264/AVC compressed video transmission over wireless networks. The video is YUV 4:2:0 formats with a resolution of 768 × 480 and a frame rate of 30 fps. Four bitrates and four packet-loss rates are performed. However, this database has been taken offline temporarily because it has limited video level contents and a tendency to cluster at 0.95–0.96 correlation for most objective metrics.

MMSP 3D Video Quality Assessment Database contains stereoscopic videos with a resolution of 1920 × 1080 pixels and a frame rate of 25 fps. Various indoor and outdoor scenes with a large variety of color, texture, motion, and depth structure have been captured. The database contains 6 scenes, and 20 subjects participated in the test. For each of the scenes, 5 different stimuli have been considered corresponding to different camera distances (10, 20, 30, 40, and 50 cm).

MMSP Scalable Video Database is related to two scalable video codecs (SVC and wavelet-based codec), three HD contents, and bit rates ranging between 300 kbps and 4 Mbps. There are three spatial resolutions (320 × 180, 640 × 360, and 1280 × 720), and four temporal resolutions (6.25 fps, 12.5 fps, 25 fps, and 50 fps). In total, 28 and 44 video sequences were considered for each codec, respectively. The video data are in the YUV 4:2:0 formats.

VQEG HDTV Database has four different video formats – 1080p at 25 and 29.97 fps, 1080i at 50 and 59.94 fps. The impairments are restricted to MPEG-2 and H.264, with both coding-only error and coding-plus-transmission error. The video sequences are released progressively via the Consumer Digital Video Library (CDVL) [3].

We summarize and compare these video quality databases in Table 3 for the convenience of readers.

B) Major VQA metrics

One obvious way to implement VQMs is to apply a still IQA metric on a frame-by-frame basis. The quality of each frame is evaluated independently, and the global quality of the video sequence can be obtained by a simple time average.

SSIM has been applied in VQA as reported in [99]. The quality of the distorted video is measured in three levels: the local region level, the frame level, and the sequence level. First, the SSIM indexing approach is applied to the Y, Cb, and Cr color components independently and combined into a local quality measure using a weighted summation. In the second level of quality evaluation, the local quality values are weighted to obtain a frame level quality index. Finally, in the third level, overall quality of the video sequence is given by weighted summation of the frame level quality index. This approach is often called **V-SSIM** (FR, signal structure-based metric), and has been demonstrated to perform better than KPN/Swisscom CT [91] (the best metric for the Video Quality Experts Group (VQEG) Phase I test dataset [17]) in [99].

Wang and Li [97] proposed **Speed-SSIM** (FR, signal structure based metric) that incorporated a model of the human visual speed perception by formulating the visual perception process in an information communication

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

framework. Consistent improvement over existing VQA algorithms has been observed in the validation with the VQEG Phase I test dataset [17].

Watson *et al.* [102] developed a VQM, which they call **digital video quality (DVQ)** (FR, HVS model-based metric). The DVQ accepts a pair of video sequences and computes a measure of the magnitude of the visible difference between them. The first step consists of various sampling, cropping, and color transformations that serve to restrict processing to a region of interest (ROI) and to express the sequence in a perceptual color space. This stage also deals with de-interlacing and de-gamma-correcting the input video. The sequence is then subjected to a blocking and a discrete cosine transform (DCT), and the results are transformed to local contrast. Then, the next steps are temporal, spatial filtering, and a contrast masking operation. Finally, the masked differences are pooled over spatial, temporal and chromatic dimensions to compute a quality measure.

Video Quality Metric (VQM) (RR, HVS model-based metric) [78] is developed by National Telecommunications and Information Administration (NTIA) to provide an objective measurement for perceived video quality. 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 efficient quality ratings. The General Model contains seven independent parameters. Four parameters (si_loss, hv_loss, hv_gain, and si_gain) are based on the features extracted from spatial gradients of Y luminance component, two parameters (chroma_spread, chroma_extreme) are based on the features extracted from the vector formed by the two chrominance components (Cb, Cr), and one parameter (ct_ati_gain) is based on the product of features that measure contrast and motion, both of which are extracted from Y luminance component. The VQM takes the original video and the processed video as inputs and is computed using the linear combination of these seven parameters. Owing to its good performance in the VQEG Phase II validation tests, the VQM method was adopted as a national standard by the American National Standards Institute (ANSI) and as International Telecommunications Union Recommendations [21, 22].

By analyzing subjective scores of various video sequences, Lee *et al.* [49] found out that the HVS is sensitive to degradation around edges. In other words, when edge areas of a video sequence are degraded, human evaluators tend to give low-quality scores to the video, even though the overall MSE is not large. Based on this observation, they proposed an objective video quality measurement method based on degradation around edges. In the proposed method, they first applied an edge detection algorithm to videos and located edge areas. Then, they measured degradation of those edge areas by computing MSEs and used it as a VQM after some post-processing. Experiments show that this proposed method **EPSNR** (FR, video fidelity metric) outperforms the conventional PSNR. This method was also evaluated by independent laboratory groups in the VQEG Phase II test. As a result, it was included in international recommendations for objective video quality measurement.

Kawayoke *et al.* [46] suggested a new objective VQA method, called **continuous video quality (CVQ)** (NR, learning-oriented metric). The metric can provide quality values at a rate of two scores per second according to the data obtained from subjective assessment tests under a SSCQE method. It is based on the concept that frame quality value needs to be adjusted by spatial and temporal information. As a result, the objective quality scores computed by this approach have a higher estimation accuracy than frame quality scores.

More recently, an approach integrates both spatial and temporal aspects of distortion assessment, known as **MOtion-based Video Integrity Evaluation (MOVIE)** index (FR, HVS model based metric) [82]. The MOVIE uses optical flow estimation to adaptively guide spatial–temporal filtering using three-dimensional (3D) Gabor filterbanks. The key differentiation of this method is that a subset of filters is selected adaptively at each location based on the direction and speed of motion, such that the major axis of the filter set is oriented along the direction of motion in the frequency domain. The video quality evaluation process is carried out with coefficients computed from these selected filters only. One component of the MOVIE framework, known as the Spatial MOVIE index, uses the output of the multi-scale decomposition of reference and test videos to measure spatial distortions in the video. The second component of the MOVIE index, known as the Temporal MOVIE index, captures temporal degradations in the video. The Temporal MOVIE index computes and uses motion information from the reference video, and evaluates the quality of the test video along the motion trajectories of the reference video. Finally, the Spatial MOVIE index and the Temporal MOVIE index are combined to obtain a single measure of video quality known as the MOVIE index. The performance of MOVIE on the VQEG FRTV Phase I dataset is summarized in [82].

In addition, **TetraVQM** (FR, HVS model-based metric) [25] has been proposed to utilize motion estimation within a VQA framework, where motion-compensated errors are computed between reference and distorted images. Based on the motion vectors and the motion prediction error, the appearance of new image areas and the display time of objects are evaluated. In addition, degradations on moving objects can be judged more exactly. In [72], Ninassi *et al.* tried to utilize models of visual attention (VA) and human eye movements to improve VQA performance. The temporal variations of the spatial distortions are evaluated both at eye fixation level and on the whole video sequence. These two kinds of temporal variations are assimilated into a short-term temporal pooling and a long-term temporal pooling, respectively.

V-Factor (NR, packet-analysis-based metric) [105] is a real-time, packet-based VQM, which works without the need of references. In [105], this metric is primarily used in MPEG-2 and H.264 video streamings over IP networks.

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

First, it inspects several parts of the video stream, including the transport stream (TS) headers, the packetized elementary stream (PES) headers, the video coding layer (VCL), and the decoded video signal. Then, it analyzes the bitstream to obtain static parameters, such as the frame rate and the image size. The dynamic parameters (e.g., variation of quantization steps) are also obtained along with the analysis. The final video quality is estimated based upon the content characteristics, compression methods, bandwidth constraints, delays, jitter, and packet loss. Among these six factors, the first three are affected by video impairments and the last three are caused by network impairments. In addition, this metric also analyzes real-time network impairments to calculate the packet loss probability ratio by using hidden Markov models. The final V-Factor value (i.e., the estimate of MOS) is obtained by using a codec-specific curve fit equation and inputs from the following three models: the bandwidth model, the VCL complexity model, and the loss model.

Li *et al.* [54] proposed to use temporal inconsistency measure (TIM) to describe visual disparity of the same object in consecutive distortion frames. First, they performed block-based motion estimation on the reference video to obtain the motion vectors. Then, the motion vectors can be used to create motion-compensated frames for reference and distorted videos, respectively. The difference between motion compensated and real frames of the reference video (DoR) is called inherent difference. Similarly, there is also a difference between motion compensated and real frames of the distorted video (DoD). However, DoD consists of two components, including inherent difference and temporal inconsistency. Hence, the TIM can be computed by subtracting DoR from DoD. In the end, they incorporated TIM into MSE, called **MSE_TIM** (FR, video fidelity metrics) and introduced a weighting parameter to adjust the importance between spatial impairment and TIM in quality prediction. The experiment results show that TIM improves the performance of MSE. Moreover, the performance becomes even better when using TIM alone.

In [24], the authors proposed a new VQM, named **spatial–temporal assessment of quality (STAQ)** (RR, HVS model-based metric). As the name suggests, it includes both spatial and temporal parts. In the first step, they used a temporal approach to find the matching regions in adjacent frames. One important change from existing motion estimation methods during this step is to use CW-SSIM instead of the mean absolute difference to compute the motion vectors. This will increase the precision of finding the matching regions. In the second step, a spatial method is used to compute the quality of the matching regions extracted via the temporal approach. The visual attention map (VAM) is used to weight each sub-block in the luminance channel based on the importance. In the final step, the video quality is estimated according to the values obtained from both the spatial and temporal domains, and quality of experience (QoE) is introduced as a function related to the motion activity density group of the video to control the pooling function. The results are quite promising in H.264 distorted video case, but are less competitive than MOVIE in either MPEG-2 or IP case.

There is also another approach integrating both spatial and temporal domains, called **spatiotemporal MAD (ST-MAD)** (FR, HVS model-based metric) [93], which is extended from the image quality metric MAD [48]. First, a spatiotemporal slice (STS) image is constructed from the time-based slices of the reference and distorted videos. The detailed procedure is as follows: a single column or row of the frame is extracted for each video frame, and these columns (or rows) are stacked from left to right (or top to bottom) to become a STS image. Then ST-MAD estimates motion-based distortions by using MAD's appearance-based model to STS images. Next, it gives larger weights to the fast-moving regions by applying optical-flow algorithm. Finally, it employs a combination rule to add spatial and temporal distortions together. Experimental results show that ST-MAD performs better than other state-of-the-art quality metrics in LIVE Video Quality Database, especially on H.264 and MPEG-2 distorted videos. However, MOVIE only outperforms ST-MAD for wireless distorted videos.

To summarize these VQA models, we present a simple comparison based on reference availability and assessment methodology in Table 4, as well as providing comments on strength and weakness of each metric.

C) Application in perceptual video coding

Since perceptual quality assessment is a hot topic in video coding, we use this as an example for applications. Currently, there are two main approaches of perceptual video coding. One is to use different IQA or VQA metrics to measure distortions and develop the perceptual rate-distortion model to achieve better performance in a perceptual sense. The other one is to utilize human visual features to develop a just noticeable distortion (JND) model for quantization step (QP) selection, or a visual attention (VA) model in order to find the ROI in the target video and optimize the bit allocation corresponding to ROI information. A JND model may be combined with a VA one for a more comprehensive evaluation (to become a foveated JND model).

For the former approach, not all applications are developed to the whole codec. Some efforts [31, 106] are made to tune the performance of encoding intra frames or made to optimize the coding efficiency of inter frames. The others target overall rate-distortion optimization of video coding. The algorithms are strongly bound to the codec type because the measurement of distortion is replaced in a perceptual fashion.

For the latter approach, the JND model is used to analyze the image features. Compared to the former method, it is more independent of the codec type.

Use of IQA or VQA metrics

Chen *et al.* [42, 75] proposed rate-distortion framework based on the SSIM index. In [42], the mode decision of H.264 intra-frame and inter-frame coding is optimized

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

Table 4. Classification of VQA models based on reference availability and assessment methodology.

VQA Model	Reference availability	Assessment methodology	Remarks (strength and weakness)
V-SSIM	FR	Signal structure	• Utilize different weighting strategy for the quality scores in three levels • Perform better than KPN/Swisscom CT in VQEG FR-TV-I database
Speed-SSIM	FR	Signal structure	• Incorporated a model of human visual speed perception • Consistent improvement in validation with the VQEG Phase I test dataset
DVQ	FR	HVS model	• Contrast masked differences are pooled over spatial temporal and chromatic dimensions to compute a quality measure
VQM	RR	HVS model	• Provide several quality models • Good performance in the VQEG Phase II validation tests, VQM was adopted as a national standard
EPSNR	FR	Video fidelity	• Video quality measurement based on degradation around edges • Outperform conventional PSNR
CVQ	NR	Learning-oriented	• Adjust frame quality value by spatial and temporal information • Have higher estimation accuracy than frame quality scores
MOVIE	FR	HVS model	• Use optical flow estimation to adaptively guide spatial–temporal filtering using 3D Gabor filterbanks • Perform the best in both LIVE and VQEG FR-TV-I databases
TetraVQM	FR	HVS model	• Based on the motion vectors and the motion prediction error, the appearance of new image areas and the display time of objects are evaluated • Degradations on moving objects are judged more exactly
V-Factor	NR	Packet analysis	• Real-time • Primarily used on MPEG-2 and H.264 video streaming
MSE_TIM	FR	Video fidelity	• Incorporate TIM into MSE and introduce a weighting parameter to adjust the importance between spatial impairment and TIM in quality prediction • Improves the performance of MSE
STAQ	RR	HVS model	• QoE is introduced as a function related to motion activity density group of the video to control the pooling function • The results are quite promising for H.264 distorted videos
ST-MAD	FR	HVS model	• A spatiotemporal slice (STS) image is constructed from the time-based slices of the reference and distorted videos • Give larger weights to fast-moving regions • Perform better than other state-of-the-art quality metrics in LIVE Video Quality Database, especially on H.264 and MPEG-2 distorted videos

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

perceptually by using SSIM index. The SSIM index is applied to replace the SSD to measure the difference between the reference block and the reconstructed block. Since it is hard to determine rate-distortion optimization by the SSIM index, the proposed approach to rate-distortion modeling provides a way to determine the Lagrange multiplier which is related to SSIM in the cost function. The rate-distortion curve fitting is defined by two parameters α and β which can be computed from two data points of the key frame. By using the data, the rate-distortion curves of subsequent frames can be estimated. For the given rate-distortion curve, the Lagrange multiplier can be calculated by the gradient or slope of the curve. In [75], the perceptual encoding scheme is based on the rate control algorithm in [42] and extended to bit allocation. The proposed rate-control scheme separates the coding methods of key frames and other frames. The algorithm adopts extra quantization parameters for key frames to update the rate-distortion model. More precisely, the Lagrange multiplier is selected adaptively according to the input data from key frames.

The perceptual cost function determines the target bit budget in the frame level and the QP sizes. By combining [42] and [75], the proposed technique is thoroughly implemented to improve perceptual rate control optimization of H.264/AVC.

In [94], a model related to the reduced reference SSIM is developed to improve rate-distortion optimization. Instead of DNT, the proposed algorithm extracts the frame features from discrete cosine transform (DCT). With less computing complexity than DNT, the DCT coefficients provide required partial information of the reference image and lead to the estimated reduced reference SSIM index, which is an important parameter of the proposed rate-distortion model. The SSIM index is generated by the local SSIM index via sliding windows. The SSIM is provided by overlapped blocks, but the macroblocks are processed individually in the encoder. Also, the boundaries of the macroblocks are not continuous. To solve these issues, the macroblocks are extended to 22×22 and a sliding 4×4 window is applied to get the SSIM index. The reference-reduced SSIM index is derived from the DCT coefficients. At first, the DCT coefficients of 4×4 non-overlap blocks are calculated and then grouped into 16 subbands. The reduced reference distortion can be defined from the DCT subbands and MSE to the reference frame. Since the measured distortion is linearly equivalent to the SSIM index, the reduced reference SSIM index can be written in the form of the distortion. The proposed algorithm tends to update the parameters of the model in frame level and adjust the Lagrange multiplier in macro-block level.

The SSIM index is introduced to video coding to model the perceived distortion. Since SSIM is not a traditional block-based distortion measurement, current video compression standard can be optimized perceptually by introducing SSIM as a distortion measurement. In [42, 75], the RD curve is parameterized to fit the SSIM RD curve; the complexity of SSIM can be reduced and a more practical method is proposed in [94].

Use of JND and VA models

Besides SSIM, JND is also applied to video coding algorithms. The JND is measured based on sensitivity of the HVS. With the JND, priority bit-allocation can be determined. In [29], a foveated JND model is proposed to measure distortion. This model combines the spatial JND model and the temporal JND model. For spatial JND, the measurement is based on the luminance of the background. If the luminance of the background is not high enough for human observers to recognize the targeted objects, then a larger QP is used to encode the frame. The threshold of background luminance is not only defined by spatial features but also considered temporal features. In the temporal model, change of luminance across frames is the key point. In the proposed model, inter-frame luminance change is considered as larger visibility threshold and separated in two cases, which are high-to-low and low-to-high. The former change results in more significant VA. The foveated JND is integrated to the H.264/AVC encoder. The QP is adjusted by weighting the macroblocks. If the macroblocks are perceived in higher priority, they can tolerate less distortion and preserve more bit budgets.

Itti *et al.* [55] developed a VA model to detect the ROI in the video. The model is based on human visual characteristics including color information, contrast, shape, motion, etc. The model prediction generates the saliency map which is used in the bit allocation strategy. To improve the saliency map, frame to frame information is considered to update the salient locations of the objects. The relationships of the object across frames are determined by the four criteria: the Euclidean distance between the location in different frames, the Euclidean distance between feature vectors corresponding to the locations, a penalty term of the differences between frames to depress permuting pairings, and a tracking priority according to the intensity of the saliency to encourage track of the salient objects. With the criteria, the proposed algorithm can identify the salient objects and track their locations in the map. Combing in the information, the more significant object is assigned to higher priority for bit allocation.

More consideration of temporal and textural features

Motion and texture are significant features to the HVS for videos. Video coding by considering texture and motion can achieve good performance in a perceptual way. The approach in [26] is based on texture and motion modeling. The texture model employed in the algorithm is to separate perceptually relevant and non-relevant regions. The relevant region needs more bits to encode. The temporal (motion) model tries to improve consistency in textural regions across frames. Texture analysis provides information of textural regions to the encoder; the texture synthesis is applied to the decoder to reconstruct the scene. In texture analysis, frames are divided into groups with the same textures and the boundaries of the regions are detected. The features extracted in this stage include gray-level co-occurrence matrix, angular second moment, dissimilarity,

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

correlation, entropy, sum of squares, and coefficients of Gabor filters. The employed segmentation techniques are split-and-merge method and K-means clustering. In order to track the region from frame to frame, motion vectors are bound to the textural regions. The temporal model is parameterized by the motion vectors to obtain the location of the regions in the consequent frames. In the encoder side, only key frames and non-synthesizable parts are coded by H.264/AVC. At the decoder, texture synthesis is designed to construct the other parts. With the temporal information, textures of the synthesizable frames are derived from the key frames and segmentation information is also passed from the encoder via the channel as side information to reconstruct the frame at the decoder.

To guarantee temporal consistency of texture-based video coding, a different approach was taken in [73]. The framework is established on cube-based texture growing method [71]. The proposed algorithm utilizes side information, which is a coded bitstream with a larger QP of the source video for two advantages. One is that the side information can be generated by any coding tool hence it can be associated to any video coding system. The other one is that the amount of side information can be adjusted by the QP with the result that the algorithm is flexible. To achieve the goal, an area-adaptive side information selection scheme that can decide the proper amount of side information is devised. The scheme determines rate-distortion optimization of the output coded data and side information bitrate. The results show that the gap between the analyzed and synthesized texture regions can be fulfilled and the perceptual quality of the regions is similar. In [26], the algorithm can significantly help to save more bits used in the side information. For intra coding, the proposed algorithm in [73] reconstructs the texture by the texture seed from a low-quality video, so the side information can be reduced by controlling the mechanism.

Naccari and Pereira [67] designed a complete perceptual video coding algorithm covering decoding, encoding, and testing tools. The JND model generates a threshold for each DCT subband coefficient. The adopted JND model contains spatial masking and temporal masking components. The spatial masking model is related to three properties: frequency band masking, luminance variations masking, and image pattern masking. Frequency band masking reflects the visual sensitivity of the noise introduced in DCT coefficients. Luminance variations masking reflects the change of the luminance part in different image regions. The JND threshold of image pattern masking varies with the threshold of frequency band masking and luminance variations masking.

The temporal masking model uses an existing model [103] because of its performance compared to other solutions. The model is established by using motion vector information. To apply this model, the issues of B-frame and intra frame are considered. Two motion vectors are used in the B-frame, and only the past vector is adopted in the model. For intra, skip motion vector is introduced to the JND computation. In decoder side, the JND model is employed to estimate average block luminance, integer DCT coefficients, and JND thresholds. In encoder side, the model is integrated into quantization, motion estimation, and rate-distortion optimization. The QP for each DCT band of a given macroblock is adjusted by the respective JND threshold. The motion estimation andthe rate-distortion optimization processes are weighted by the JND thresholds. The weighting process tends to weight the estimation error to provide the error in a perceptual fashion. Perceptual distortion is employed to motion estimation and rate-distortion optimization. Thus, the cost function of rate-distortion optimization is converted to perceptual cost function and the Lagrange multiplier is also changed in the flavor. The proposed testing procedure is to assess rate-distortion performance. The algorithm is to compare the performance of a codec and another one based on a quality metric.

Other attempts

Besides visual quality metrics and perceptual models, audio information can be used to improve coding efficiency. In practical cases, audio is bound to video, hence the audio is also perceived by human observers synchronously. Lee *et al.* [50] proposed the video coding algorithm combined with audio information. The proposed scheme utilized the relation of the sound source and corresponding spatial location to gain the efficient coding with the scene that contains multiple moving objects. The work is to find the sound source and its region. Based on the assumption that human observers tends to recognize the sound object as the ROI, the corresponding region is encoded with more bits. The implementation encoded the ROI blocks with smaller QP relative to the non-ROI ones.

V. PERFORMANCE COMPARISON

We use the following three indexes to measure metric performance [91, 92]. The first index is the Pearson linear correlation coefficient (PLCC) between objective/subjective scores after non-linear regression analysis. It provides an evaluation of prediction accuracy. The second index is the Spearman rank order correlation coefficient (SROCC) between the objective/subjective scores. It is considered as a measure of prediction monotonicity. The third index is the root-mean-squared error (RMSE). Before computing the first and second indexes, we need to use the logistic function and the procedure outlined in [91] to fit the objective model scores to the MOS (or DMOS) in order to account for quality rating compression at the extremes of the test range and prevent the overfitting problem. The monotonic logistic function used to fit the objective prediction scores to the subjective quality scores [91] is:

$$f(x) = \frac{\beta_1 - \beta_2}{1 + \exp^{-(x-\beta_3)/|\beta_4|}} + \beta_2, \tag{5}$$

where x is the objective prediction score, $f(x)$ is the fitted objective score, and the parameters β_j $(j = 1, 2, 3, 4)$ are

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

Table 5. Performance comparison among IQA models in CSIQ database.

Measure			
IQA model	PLCC	SROCC	RMSE
MS-SSIM	0.8666	0.8774	0.1310
SSIM	0.8594	0.8755	0.1342
VIF	0.9253	0.9194	0.0996
VSNR	0.8005	0.8108	0.1573
NQM	0.7422	0.7411	0.1759
PSNR-HVS	0.8231	0.8294	0.1491
IFC	0.8358	0.7671	0.1441
PSNR	0.8001	0.8057	0.1576
FSIM	0.9095	0.9242	0.1091
MAD	0.9502	0.9466	0.0818
IW-SSIM	0.9025	0.9212	0.1131
CF-MMF	**0.9797**	**0.9755**	**0.0527**
CD-MMF	**0.9675**	**0.9668**	**0.0664**

Table 6. Performance comparison among IQA models in database.

Measure			
IQA model	PLCC	SROCC	RMSE
MS-SSIM	0.9402	0.9521	9.3038
SSIM	0.9384	0.9479	9.4439
VIF	0.9597	0.9636	7.6737
VSNR	0.9235	0.9279	10.4816
NQM	0.9128	0.9093	11.1570
PSNR-HVS	0.9134	0.9186	11.1228
IFC	0.9261	0.9259	10.3052
PSNR	0.8701	0.8756	13.4685
FSIM	0.9540	0.9634	8.1938
MAD	0.9672	0.9669	6.9419
IW-SSIM	0.9425	0.9567	9.1301
CF-MMF	**0.9734**	**0.9732**	**6.2612**
CD-MMF	**0.9802**	**0.9805**	**5.4134**

chosen to minimize the least squares error between the subjective score and the fitted objective score. Initial estimates of the parameters were chosen based on the recommendation in [91]. For an ideal match between the objective prediction scores and the subjective quality scores, PLCC = 1, SROCC = 1, and RMSE = 0.

A) Image quality metric benchmarking

To examine the performance of existing popular image quality metrics in this work, we choose CSIQ, LIVE, and TID2008 to test image quality metrics since they include the largest number of distorted images and also span more distortion types; these three databases cover most image distortion types that other publicly available image quality databases can provide. The performance results are listed in Tables 5–7 with the three indexes given above. The two best performing metrics are highlighted in bold. Clearly, MMF (both CF-MMF and CD-MMF) [57, 59] have the highest PLCCs, SROCCs, and the smallest RMSEs among the 13 image quality metrics under comparison.

Table 7. Performance comparison among IQA models in TID2008 database.

Measure			
IQA model	PLCC	SROCC	RMSE
MS-SSIM	0.8389	0.8528	0.7303
SSIM	0.7715	0.7749	0.8537
VIF	0.8055	0.7496	0.7953
VSNR	0.6820	0.7046	0.9815
NQM	0.6103	0.6243	1.0631
PSNR-HVS	0.5977	0.5943	1.0759
IFC	0.7186	0.5707	0.9332
PSNR	0.5355	0.5245	1.1333
FSIM	0.8710	0.8805	0.6592
MAD	0.8306	0.8340	0.7474
IW-SSIM	0.8488	0.8559	0.7094
CF-MMF	**0.9525**	**0.9487**	**0.4087**
CD-MMF	**0.9538**	**0.9463**	**0.4032**

Table 8. Performance comparison of VQA models in database.

Measure			
VQA model	PLCC	SROCC	RMSE
PSNR	0.5465	0.5205	9.1929
VSNR	0.6880	0.6714	7.9666
SSIM	0.5413	0.5233	9.2301
V-SSIM	0.6058	0.5924	8.7337
VQM	0.7695	0.7529	7.0111
Q_{SVR} [70]	0.7924	0.7820	6.6908
MOVIE	**0.8116**	**0.7890**	**6.4130**
ST-MAD [93]	**0.8299**	**0.8242**	—

Table 9. Performance comparison of VQA models in EPFL-POLIMI database [76].

Measure		
VQA model	PLCC	SROCC
PSNR	0.7951	0.7983
VSNR	0.8955	0.8958
SSIM	0.8341	0.8357
VQM	0.8433	0.8375
MOVIE	**0.9302**	**0.9203**

B) Video quality metric benchmarking

For the comparison of the state-of-the-art VQMs, LIVE Video Quality Database, and EPFL-PoliMI Video Quality Assessment Database are adopted.

Although most people use VQEG-FRTV Phase I Database (built in 2000) to test their video metric performance previously [82, 99], we use LIVE Video Quality Database (released in 2009) as our test database because it is new and contains distortion types in more processes, such as H.264 compression, simulated transmission of H.264 packetized streams through error-prone wireless networks and error-prone IP networks, and MPEG-2 compression. The comparison results are summarized in Table 8. Here, the

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

image quality metrics (i.e., PSNR, VSNR, and SSIM) are used on a frame-by-frame basis for the video sequence, and then time-averaging the frame scores to obtain the video quality score.

In Table 8, the results of ST-MAD are extracted from [93]. From Table 8, we can see that ST-MAD and MOVIE are the best metrics (which are both highlighted in bold) for LIVE Video Quality Database; VQM ranks the third. It means that MOVIE and ST-MAD correlate better with subjective results than other approaches under comparison. The reason why ST-MAD and MOVIE perform well is that they both consider the spatial and temporal features. In general, consideration of temporal information as well as interaction of spatial and temporal features [69] can improve the video quality prediction performance.

In addition, we also summarize the performance results in Table 9 from [76] to see if the existing quality metrics can predict the quality well for videos distorted with different PLRs. We can observe that MOVIE still works the best compared to other metrics in Table 9 with packet loss.

VI. DISCUSSION ON FUTURE TRENDS

Although many visual quality assessment metrics have been developed for both image and video during the past decade, there are still great technological challenges ahead and much space for improvement, toward effective, reliable, efficient, and widely accepted replacement for MSE/PSNR, for both standalone and embedded applications. We will discuss the possible directions in this section.

A) PSNR or SSIM-modified metrics

PSNR has always been criticized for poor correlation with human subjective evaluations. However, according to our observations [57, 59], PSNR sometimes still can work very well on some specific distortion types, such as additive and quantization noise. Hence, a lot of metrics have been developed or derived from PSNR, such as PSNR-HVS [34], EPSNR [49], and SPHVSM [45]. They either incorporate some related HVS characteristics into PSNR or include some experimental observations to modify PSNR to improve the correlation. Promising results can be achieved in this way of modification. Among the quality metrics we just mentioned above, only the EPSNR is developed to use on VQA.

As a single metric, the SSIM is considered the well-performed metric among all visual quality evaluation metrics, in terms of consistency. Thus, researchers in the field have managed to transform it by changing its pooling method or using other image features. Several examples of the former are V-SSIM [99], Speed-SSIM [97], 3-SSIM [51], and IW-SSIM [98], while FSIM index [108] is an example of the latter. They are all proven quite useful in improving the quality prediction performance, especially FSIM, which shows superior performance in several image quality databases, including TID2008, CSIQ, LIVE, and IVC.

Building new metrics based upon more mature metrics (like PSNR and SSIM) is expected to continue, especially in new application scenarios (e.g., for 3D scenes, mobile media, medical imaging, image/video retargeting, computer graphics, and so on).

B) Multiple strategies or MMF approaches

MAD [48] and MMF [57, 59] are representatives for multiple strategies and MMF, respectively. Especially for the latter one, appropriate fusion of existing metrics opens the chances to build on the strength of each participating metric and the resultant framework can be even used when new, good metrics emerge. More careful and in-depth investigation is needed for this topic.

Most recently, a block-based MMF (BMMF) [44] approach is proposed on coping with IQA. The authors first decomposed images into smaller block size. Then they classify the blocks into three types (smooth, edge, and texture). And they also divided all the images into five different distortion groups, like in [57, 59]. Finally, only one appropriate quality metric is selected for each block based on the distortion groups and the block types. Fusion through all the blocks leads to the final quality score for each image. Performing MMF this way helps to reduce the high complexity caused by using multiple metrics.

C) Migration from IQA to VQA

Up to now, more research has been performed for IQA. As mentioned before, video quality evaluation can be done by using image quality metrics on a frame-by-frame basis, and then averaging to obtain a final video quality score. However, this only works well when video contents do not have large motion in temporal domain. When there exists a large motion, we need to find the temporal structure and temporal features.

The most common method is to use motion estimation to find out the motion vectors and measure the variations in the temporal domain. One simple realization of this idea is in [60]. The authors extended one existing IQA metric to a VQM by considering temporal information and converted it into a compensation factor to correct the video quality score obtained in the spatial domain. There are also other VQMs that utilize motion estimation to detect temporal variations, such as Speed-SSIM [97], MOVIE [82], TetraVQM [25], MSE_TIM [54], STAQ [24], and ST-MAD [93]. All of the above approaches improve the correlation between predictions and subjective quality scores more or less. This demonstrates that the temporal variation is indeed an important factor that we need to consider for VQA.

Another feasible method is to extend original image quality metric into a VQM by considering three additional processing steps: temporal channel decomposition, temporal masking, and temporal pooling. One example of this is recently proposed in [53]. Their resultant VQM shows a quite good performance in matching subjective scores for LIVE Video Quality Database.

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

Similarly, we can also use the MMF strategy on VQA, via fusing the scores obtained from all available VQMs. A possible problem of this approach is the high complexity because multiple metrics and video data are involved. One solution to realize efficient MMF for video is to pick up the best features used in all metrics, including both spatial and temporal features, instead of using all participating metrics as they are. Moreover, this solution gives a chance to eliminate the repetition in feature detection among different metrics, and proper machine learning techniques will be customized for this purpose. In addition, VA modeling [61] may play a more active role in VQA than IQA.

D) Audiovisual quality assessment for 4G networks

During recent years, the term quality of experience (QoE) has been used and defined as the users' perceived quality of service (QoS). More often than not in multimedia applications, quality assessment has to be performed with audio and video (images) being presented together. It is an important but less investigated research topic, in spite of some early work in this area [37, 38, 40].

It has been proposed that a better QoE can be achieved when the QoS is considered both in the network and application layers as a whole [47]. In the application layer, QoS is affected by the factors such as resolution, frame rate, sampling rate, number of channels, color, video codec type, audio codec type, and layering strategy. The network layer introduces impairment parameters such as packet loss, jitter, network delay, burstiness, decreased throughput, etc. These are all the key factors that affect the overall audiovisual QoE. Hence, the investigation into the quality assessment methods for both audio and video is also important and meaningful because video chats and video conferences over 4G networks may be frequently used by the general public in the near future. We believe this is a significant extension of the current research work and very meaningful in total multimedia experience evaluation.

Currently, there is no public database for joint audiovisual quality and experience evaluation. The establishment of such databases will facilitate research and promote advancement in this field.

E) Perceptual image/video coding

The accuracy of IQA is becoming better and better. The performance of perceptual image coding could be further improved under some specific conditions. Perceptual considerations can help the performance to be enhanced compared to the traditional image coding. As the introduced applications above, IQA metrics have been associated to video coding for some time. More and more related research is in progress.

In general, VQA-related video compression is less investigated. Seshadrinathan and Bovik [82] addressed motion-based video integrity evaluation (MOVIE) index to evaluate video quality. The MOVIE index based on Gabor decomposition is calculated from two components, which are Spatial MOVIE map and Temporal MOVIE map. The spatial part is established as a combination of SSIM and VIF; the temporal part is brought by using motion information. The performance of MOVIE shows the potential to be employed to video coding. Nevertheless, it is challenging to be handled in video coding because it needs to parse the whole video to give the index. Hence, modifying VQA to low complexity and real-time processing would be a possible goal to integrate VQA to video coding. These are issues to apply VQA to perceptual video coding.

F) No-Reference (NR) quality metrics

As we know, the NR method does not perform as well as the FR one in general because it judges the quality solely based on the distorted medium and without any reference available. However, it can be used in wider scope of applications because of its suitability in both situations with and without reference information. Moreover, the computational requirement is usually less because there is no need to process the reference. In addition to the traditional NR cases (like the relay site and receiving end of transmission), there are emerging NR applications (e.g., super-resolution construction, image, and video retargeting/adaption, and computer graphics/animation). That is the reason why several NR quality metrics have been proposed recently, including MREBN [30] and JNBM [36] in images, and CVQ [46] and V-Factor [105] in videos. We believe that there will be more quality metrics developing along this direction.

VII. CONCLUSION

In this paper, we have first reviewed the existing visual quality assessment methods and their classifications in a comprehensive perspective. Then, we introduced recent developments in IQA, including the popular public image quality databases that play important roles in facilitating relevant research activities in this field and several well-performed image quality metrics. In a similar format, we also discussed recent developments for VQA in general, the publicly available video quality databases and several state-of-the-art VQA metrics. In addition, we have presented and discussed several possible directions for future visual signal quality assessment, i.e., PSNR or SSIM-modified metrics, multiple strategy and MMF approaches, migration of IQA to VQA, joint audiovisual assessment, perceptual image/video coding, and NR quality assessment, with reasoning based upon our experience and understanding of the related research. In the end, we have compared the major existing IQA and VQA metrics, and given some discussion, by using the most comprehensive image and video quality databases, respectively.

One important class of applications of visual quality assessment is perceptual image and video coding. The perceptually driven coding methods have demonstrated their merits, compared to the traditional MSE-based coding

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

techniques. Such research takes a different path (i.e., removing perceptual signal redundancy apart from the statistical one) to further improve coding performance and makes it more use-oriented because humans are the ultimate appreciators of almost all processed visual signals. Existing and interesting methods include: utilizing a perceptual quality index to measure distortion; utilizing JND and VA models in coding; integrating motion or texture information to improve coding efficiency in a perceptual sense. We believe that there are still a lot of possibilities for perceptual coding and beyond, which wait to be discovered.

REFERENCES

[1] A57 Database. [Online]. Available: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html.

[2] Categorical Image Quality (CSIQ) Database. [Online]. Available: http://vision.okstate.edu/csiq.

[3] Digital Video Library. [Online]. Available: http://www.cdvl.org/.

[4] EPFL-PoliMI Video Quality Assessment Database. [Online]. Available: http://vqa.como.polimi.it/.

[5] IRCCyN/IVC 1080i Database. [Online]. Available: http://www.ec-nantes.fr/spip.php?article541.

[6] IRCCyN/IVC SD RoI Database. [Online]. Available: http://www.irccyn.ec-nantes.fr/spip.php?article551.

[7] IVC Image Quality Database. [Online]. Available: http://www2.irccyn.ec-nantes.fr/ivcdb.

[8] IVC-LAR Database. [Online]. Available: http://www.irccyn.ec-nantes.fr/~autrusse/Databases/LAR.

[9] LIVE Image Quality Assessment Database. [Online]. Available: http://live.ece.utexas.edu/research/quality/subjective.htm.

[10] LIVE Video Quality Database. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_video.html.

[11] LIVE Wireless Video Quality Assessment Database. [Online]. Available: http://live.ece.utexas.edu/research/quality/live_wireless_video.html.

[12] MMSP 3D Image Quality Assessment Database. [Online]. Available: http://mmspg.epfl.ch/cms/page-58394.html.

[13] MMSP 3D Video Quality Assessment Database. [Online]. Available: http://mmspg.epfl.ch/3dvqa.

[14] MMSP Scalable Video Database. [Online]. Available: http://mmspg.epfl.ch/svd.

[15] Tampere Image Database. [Online]. Available: http://www.ponomarenko.info/tid2008.htm.

[16] Toyoma Database. [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html.

[17] VQEG FRTV Phase I Database, 2000. [Online]. Available: ftp://ftp.crc.ca/crc/vqeg/TestSequences/.

[18] VQEG HDTV Database. [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/hdtv/.

[19] Wireless Imaging Quality (WIQ) Database. [Online]. Available: http://www.bth.se/tek/rcg.nsf/pages/wiq-db.

[20] Methodology for the Subjective Assessment of the Quality of Television Pictures. *Recommendation ITU-R BT.500-11* (2002).

[21] Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference. *Recommendation ITU-T J.144* (Feb. 2004).

[22] Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference. *Recommendation ITU-R BT.1683* (Jan. 2004).

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

[24] Amirshahi, S.A.; Larabi, M.: Spatial–temporal video quality metric based on an estimation of QoE. in *Quality of Multimedia Experience (QoMEX), 2011 Third International Workshop on* (2011), IEEE, pp. 84–89.

[25] Barkowsky, M.; Bialkowski, J.; Eskofier, B.; Bitto, R.; Kaup, A.: Temporal trajectory aware video quality measure. *Selected Topics in Signal Processing, IEEE J.*, **3**(2) (2009), 266–279.

[26] Bosch, M.; Zhu, F.; Delp, E.J.: Segmentation-based video compression using texture and motion models. *Selected Topics in Signal Processing, IEEE J.*, **5**(7) (2011), 1366–1377.

[27] Chandler, D.M.; Hemami, S.S.: VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *Image Processing, IEEE Trans.*, **16**(9) (2007), 2284–2298.

[28] Channappayya, S.S.; Bovik, A.C.; Heath, R.W.: Rate bounds on ssim index of quantized images. *Image Processing, IEEE Trans.*, **17**(9) (2008), 1624–1639.

[29] Chen, Z.; Guillemot, C.: Perceptually-friendly H.264/avc video coding based on foveated just-noticeable-distortion model. *Circuits and Systems for Video Technology, IEEE Trans.*, **20**(6) (2010), 806–819.

[30] Choi, M.G.; Jung, J.H.; Jeon, J.W.: No-reference image quality assessment using blur and noise, in proceedings of world academy of science, engineering and technology 50, 2009.

[31] Cui, Z.; Zhu, X.; Subjective quality optimized intra mode selection for H.264i frame coding based on ssim. in *Image and Graphics (ICIG), 2011 Sixth International Conference on* (2011), IEEE, pp. 157–162.

[32] Daly, S.J.: Visible differences predictor: an algorithm for the assessment of image fidelity. In *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology* (1992), International Society for Optics and Photonics, pp. 2–15.

[33] Damera-Venkata, N.; Kite, T.D.; Geisler, W.S.; Evans, B.L.; Bovik, A.C.: Image quality assessment based on a degradation model. *Image Processing, IEEE Trans.*, **9**(4) (2000), 636–650.

[34] Egiazarian, K.; Astola, J.; Ponomarenko, N.; Lukin, V.; Battisti, F.; Carli, M.: New full-reference quality metrics based on hvs. in *CD-ROM Proc. Second Int. Workshop Video Processing and Quality Metrics* (2006).

[35] Engelke, U.; Zepernick, H.-J.: Perceptual-based quality metrics for image and video services: A survey. in *Next Generation Internet Networks, 3rd EuroNGI Conf. on* (2007), IEEE, pp. 190–197.

[36] Ferzli, R.; Karam, L.J.: A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). *Image Processing, IEEE Trans.*, **18**(4) (2009), 717–728.

[37] Frater, M.R.; Arnold, J.F.; Vahedian, A.: Impact of audio on subjective assessment of video quality in videoconferencing applications. *Circuits and Systems for Video Technology, IEEE Trans.*, **11**(9) (2001), 1059–1062.

[38] Furini, M.; and Ghini, V.: A video frame dropping mechanism based on audio perception. in *Global Telecommunications Conference Workshops, 2004. GlobeCom Workshops 2004. IEEE* (2004), IEEE, pp. 211–216.

[39] Gao, X.; Lu, W.; Tao, D.; Li, X.: Image quality assessment based on multiscale geometric analysis. *Image Processing, IEEE Trans.*, **18**(7) (2009), 1409–1423.

[40] Ghinea, G.; Thomas, J.P.: Quality of perception: user quality of service in multimedia presentations. *Multimedia, IEEE Trans.*, **7**(4) (2005), 786–789.

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

[41] Hontzsch, I.; Karam, L.J.: Adaptive image coding with perceptual distortion control. *Image Processing, IEEE Trans.*, **11**(3) (2002), 213–222.

[42] Huang, Y.-H.; Ou, T.-S.; Su, P.-Y.; Chen, H.H.: Perceptual rate-distortion optimization using structural similarity index as quality metric. *Circuits and Systems for Video Technology, IEEE Trans.*, **20**(11) (2010), 1614–1624.

[43] Jayant, N.; Johnston, J.; Safranek, R.: Signal compression based on models of human perception. *Proceedings of the IEEE*, **81**(10) (1993), 1385–1422.

[44] Jin, L.; Egiazarian, K.; Kuo, C.-C.J.: Perceptual image quality assessment using block-based multi-metric fusion (BMMF). in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE Int. Conf.* (2012), IEEE, pp. 1145–1148.

[45] Jin, L.; Ponomarenko, N.; Egiazarian, K.: Novel image quality metric based on similarity. in *Signals, Circuits and Systems (ISSCS), 2011 10th Int. Symp.* (2011), IEEE, pp. 1–4.

[46] Kawayoke, Y.; Horita, Y.: NR objective continuous video quality assessment model based on frame quality measure. in *Image Processing, 2008. ICIP 2008. 15th IEEE Int. Conf.* (2008), IEEE, pp. 385–388.

[47] Khan, A.; Li, Z.; Sun, L.; Ifeachor, E.: Audiovisual quality assessment for 3G networks in support of e-healthcare services. in *Proc. 3rd Int. Conf. Comput. Intell. Med. Healthcare* (2007), Citeseer.

[48] Larson, E.C.; Chandler, D.M.; Most apparent distortion: full-reference image quality assessment and the role of strategy. *J. Electron. Imaging*, **19**(1) (2010), 011006–011006.

[49] Lee, C.; Cho, S.; Choe, J.; Jeong, T.; Ahn, W.; Lee, E.: Objective video quality assessment. *Opt. Eng.*, **45**(1) (2006), 017004–017004.

[50] Lee, J.-S.; Ebrahimi, T.: Efficient video coding in H.264/avc by using audio-visual information. in *Multimedia Signal Processing, 2009. MMSP'09. IEEE Int. Workshop on* (2009), IEEE, pp. 1–6.

[51] Li, C.; Bovik, A.C.: Three-component weighted structural similarity index. in *IS&T/SPIE Electronic Imaging* (2009), Int. Soc. Opt. Photon., pp. 72420Q–72420Q.

[52] Li, Q.; Wang, Z.: Reduced-reference image quality assessment using divisive normalization-based image representation. *Selected Topics in Signal Processing, IEEE J.*, **3**(2) (2009), 202–211.

[53] Li, S.; Ma, L.; Ngan, K.N.: Video quality assessment by decoupling additive impairments and detail losses. in *Quality of Multimedia Experience (QoMEX), 2011 Third Int. Workshop on* (2011), IEEE, pp. 90–95.

[54] Li, S.; Ma, L.; Zhang, F.; Ngan, K.N.: Temporal inconsistency measure for video quality assessment. in *Picture Coding Symposium (PCS), 2010* (2010), IEEE, pp. 590–593.

[55] Li, Z.; Qin, S.; Itti, L.: Visual attention guided bit allocation in video compression. *Image Vision Comput.* **29**(1) (2011), 1–14.

[56] Lin, W.; Kuo, C.-C.J.: Perceptual visual quality metrics: A survey. *J. Visual Commun. Image Represent.*, **22**(4) (2011), 297–312.

[57] Liu, T.-J.; Lin, W.; Kuo, C.-C.J.: A multi-metric fusion approach to visual quality assessment. in *Quality of Multimedia Experience (QoMEX), 2011 Third Int. Workshop on* (2011), IEEE, pp. 72–77.

[58] Liu, T.-J.; Lin, W.; Kuo, C.-C.J.: A fusion approach to video quality assessment based on temporal decomposition. in *Signal & Information Processing Association Annual Summit and Conference (APSIPA ASC), 2012 Asia-Pacific* (2012), IEEE, pp. 1–5.

[59] Liu, T.-J.; Lin, W.; Kuo, C.-C.J.: Image quality assessment using multi-method fusion. *Image Processing, IEEE Trans.*, **22**(5) (2013), 1793–1807.

[60] Liu, T.-J.; Liu, K.-H.; Liu, H.-H.: Temporal information assisted video quality metric for multimedia. in *Multimedia and Expo (ICME), 2010 IEEE Int. Conf.* (2010), IEEE, pp. 697–701.

[61] Lu, Z.; Lin, W.; Yang, X.; Ong, E.; Yao, S.: Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation. *Image Processing, IEEE Trans.*, **14**(11) (2005), 1928–1942.

[62] Lubin, J.: A visual discrimination model for imaging system design and evaluation. *Vision Models Target Detect. Recogn.*, **2** (1995), 245–357.

[63] Luo, H.: A training-based no-reference image quality assessment algorithm. in *Image Processing, 2004. ICIP'04. 2004 Int. Conf.* (2004), vol. 5, IEEE, pp. 2973–2976.

[64] Marziliano, P.; Dufaux, F.; Winkler, S.; Ebrahimi, T.: A no-reference perceptual blur metric. in *Image Processing. 2002. Proceedings. 2002 Int. Conf.* (2002), vol. 3, IEEE, pp. III–57.

[65] Masry, M.; Hemami, S.S.; Sermadevi, Y.: A scalable wavelet-based video distortion metric and applications. *Circuits and Systems for Video Technology, IEEE Trans.*, **16**(2) (2006), 260–273.

[66] Masry, M.A.; Hemami, S.S.: A metric for continuous quality evaluation of compressed video with severe distortions. *Signal process.: Image commun.*, **19**(2) (2004), 133–146.

[67] Naccari, M.; Pereira, F.: Advanced H.264/avc-based perceptual video coding: Architecture, tools, and assessment. *Circuits and Systems for Video Technology, IEEE Trans.*, **21**(6) (2011), 766–782.

[68] Narwaria, M.; Lin, W.: Objective image quality assessment based on support vector regression. *Neural Networks, IEEE Trans.*, **21**(3) (2010), 515–519.

[69] Narwaria, M.; Lin, W.: Machine learning based modeling of spatial and temporal factors for video quality assessment. in *Image Processing (ICIP), 2011 18th IEEE Int. Conf.* (2011), IEEE, pp. 2513–2516.

[70] Narwaria, M.; Lin, W.: Video quality assessment using temporal quality variations and machine learning. in *Multimedia and Expo (ICME), 2011 IEEE Int. Conf.* (2011), IEEE, pp. 1–6.

[71] Ndjiki-Nya, P.; Stuber, C.; Wiegand, T.: Texture synthesis method for generic video sequences. in *Image Processing, 2007. ICIP 2007. IEEE Int. Conf.* (2007), vol. 3, IEEE, pp. III–397.

[72] Ninassi, A.; Le Meur, O.; Le Callet, P.; Barba, D.: Considering temporal variations of spatial visual distortions in video quality assessment. *Selected Topics in Signal Processing, IEEE J.*, **3**(2) (2009), 253–265.

[73] Oh, B.T.; Su, Y.; Segall, C.; Kuo, C.-C.: Synthesis-based texture video coding with side information. *Circuits and Systems for Video Technology, IEEE Trans.*, **21**(5) (2011), 647–659.

[74] Ong, E.; Lin, W.; Lu, Z.; Yao, S.; Yang, X.; Jiang, L.: No-reference JPEG-2000 image quality metric. in *Multimedia and Expo, 2003. ICME'03. Proc. 2003 Int. Conf.* (2003), vol. 1, IEEE, pp. I–545.

[75] Ou, T.-S.; Huang, Y.-H.; Chen, H.H.: SSIM-based perceptual rate control for video coding. *Circuits and Systems for Video Technology, IEEE Trans.*, **21**(5) (2011), 682–691.

[76] Park, J.; Seshadrinathan, K.; Lee, S.; Bovik, A.C.: Video quality pooling adaptive to perceptual distortion severity, IEEE Trans. on image processsing, **22**(2) (2013), 610–620.

[77] Peli, E.: Contrast in complex images. *JOSA A*, **7**(10) (1990), 2032–2040.

[78] Pinson, M.H.; Wolf, S.: A new standardized method for objectively measuring video quality. *Broadcasting, IEEE Trans.*, **50**(3) (2004), 312–322.

[79] Ponomarenko, N.; Silvestri, F.; Egiazarian, K.; Carli, M.; Astola, J.; Lukin, V.: On between-coefficient contrast masking of dct basis

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

functions. in *Proc. Third Int. Workshop on Video Processing and Quality Metrics* (2007), vol. 4.

[80] Rehman, A.; Wang, Z.: Reduced-reference ssim estimation. In *Image Processing (ICIP), 2010 17th IEEE Int. Conf.* (2010), IEEE, pp. 289–292.

[81] Richter, T.; Kim, K.J.: A ms-ssim optimal jpeg 2000 encoder. in *Data Compression Conference, 2009. DCC'09.* (2009), IEEE, pp. 401–410.

[82] Seshadrinathan, K.; Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Trans.*, **19**(2) (2010), 335–350.

[83] Seshadrinathan, K.; Soundararajan, R.; Bovik, A.C.; and Cormack, L.K.: Study of subjective and objective quality assessment of video. *Image Processing, IEEE Trans.*, **19**(6) (2010), 1427–1441.

[84] Sheikh, H.R.; Bovik, A.C.: Image information and visual quality. *Image Processing, IEEE Trans.*, **15**(2) (2006), 430–444.

[85] Sheikh, H.R.; Bovik, A.C.; De Veciana, G.: An information fidelity criterion for image quality assessment using natural scene statistics. *Image Processing, IEEE Trans.*, **14**(12) (2005), 2117–2128.

[86] Sheikh, H.R.; Sabir, M.F.; Bovik, A.C.: A statistical evaluation of recent full reference image quality assessment algorithms. *Image Processing, IEEE Trans.*, **15**(11) (2006), 3440–3451.

[87] Suresh, S.; Babu, V.; Sundararajan, N.: Image quality measurement using sparse extreme learning machine classifier. in *Control, Automation, Robotics and Vision, 2006. ICARCV'06. 9th Int. Conf.* (2006), IEEE, pp. 1–6.

[88] Tan, D.; Tan, C.; Wu, H.: Perceptual color image coding with JPEG2000. *Image Processing, IEEE Trans.*, **19**(2) (2010), 374–383.

[89] Teo, P.C.; Heeger, D.J.: Perceptual image distortion. in *Image Processing, 1994. Proceedings. ICIP-94., IEEE Int. Conf.* (1994), vol. 2, IEEE, pp. 982–986.

[90] Tong, H.; Li, M.; Zhang, H.-J.; Zhang, C.: No-reference quality assessment for JPEG2000 compressed images. in *Image Processing, 2004. ICIP'04. 2004 Int. Conf.* (2004), vol. 5, IEEE, pp. 3539–3542.

[91] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase I. *Mar. 2000. [Online]. Available*: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI.

[92] VQEG. Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II. *Aug. 2003. [Online]. Available*: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII.

[93] Vu, P.V.; Vu, C.T.; Chandler, D.M.: A spatiotemporal most-apparent-distortion model for video quality assessment. In *Image Processing (ICIP), 2011 18th IEEE Int. Conf.* (2011), IEEE, pp. 2505–2508.

[94] Wang, S.; Rehman, A.; Wang, Z.; Ma, S.; Gao, W.: SSIM-motivated rate-distortion optimization for video coding. *Circuits and Systems for Video Technology, IEEE Trans.*, **22**(4) (2012), 516–529.

[95] Wang, Z.; Bovik, A.C.: Mean squared error: love it or leave it? a new look at signal fidelity measures. *Signal Process. Mag., IEEE*, **26**(1) (2009), 98–117.

[96] Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *Image Process. IEEE Trans.*, **13**(4) (2004), 600–612.

[97] Wang, Z.; Li, Q.: Video quality assessment using a statistical model of human visual speed perception. *JOSA A*, **24**(12) (2007), B61–B69.

[98] Wang, Z.; Li, Q.: Information content weighting for perceptual image quality assessment. *Image Process. IEEE Trans.*, **20**(5) (2011), 1185–1198.

[99] Wang, Z.; Lu, L.; Bovik, A.C.: Video quality assessment based on structural distortion measurement. *Signal Process.: Image Commun.*, **19**(2) (2004), 121–132.

[100] Wang, Z.; Simoncelli, E.P.: Translation insensitive image similarity in complex wavelet domain. in *In Acoustics, Speech, and Signal Processing, 2005. Proceedings (ICASSP05). IEEE Int. Conf.* (2005), Citeseer.

[101] Wang, Z.; Simoncelli, E.P.; Bovik, A.C.: Multiscale structural similarity for image quality assessment. in *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conf.* (2003), vol. 2, IEEE, pp. 1398–1402.

[102] Watson, A.B.; Hu, J.; McGowan, J.F.: Digital video quality metric based on human vision. *J. Elect. imaging*, **10**(1) (2001), 20–29.

[103] Wei, Z.; Ngan, K.N.: Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain. *Circuits and Systems for Video Technology, IEEE Trans.*, **19**(3) (2009), 337–346.

[104] Winkler, S.: *Digital video quality: vision models and metrics*. Wiley, 2005.

[105] Winkler, S.; Mohandas, P.: The evolution of video quality measurement: from psnr to hybrid metrics. *Broadcasting, IEEE Trans.*, **54**(3) (2008), 660–668.

[106] Yang, C.-L.; Leung, R.-K.; Po, L.-M.; Mai, Z.-Y.: An SSIM-optimal H.264/avc inter frame encoder. in *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE Int. Conf.* (2009), vol. 4, IEEE, pp. 291–295.

[107] Yim, C.; Bovik, A.C.: Quality assessment of deblocked images. *Image Processing, IEEE Trans.*, **20**(1) (2011), 88–98.

[108] Zhang, L.; Zhang, L.; Mou, X.; Zhang, D.: FSIM: a feature similarity index for image quality assessment. *Image Processing, IEEE Trans.*, **20**(8) (2011), 2378–2386.

APPENDIX. STANDARD SUBJECTIVE TESTING METHODS [20, 23]

a) Pair Comparison (PC)

The method of PCs implies that the test sequences are presented in pairs, consisting of the same sequence being presented first through one system under test and then through another system.

b) Absolute Category Rating (ACR)

The ACR method is a category judgment where the test sequences are presented one at a time and are rated independently on a discrete five-level scale from "bad" to "excellent". This method is also called Single Stimulus Method.

c) Degradation Category Rating (DCR) (also called the Double-Stimulus Impairment Scale (DSIS))

The reference picture (sequence) and the test picture (sequence) are presented only once or twice. The reference is always shown before the test sequence, and neither is repeated. Subjects rate the

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press

amount of impairment in the test sequence on a discrete five-level scale from "very annoying" to "imperceptible".

d) Double-Stimulus Continuous Quality Scale (DSCQS)

The reference and test sequences are presented twice in alternating fashion, in the order of the two chosen randomly for each trial. Subjects are not informed which one is the reference and which one is the test sequence. They rate each of the two separately on a continuous quality scale ranging from "bad" to "excellent". Analysis is based on the difference in rating for each pair, which is calculated from an equivalent numerical scale from 0 to 100.

e) Single-Stimulus Continuous Quality Evaluation (SSCQE)

Instead of seeing separate short sequence pairs, subjects watch a program of 20–30 minutes duration which has been processed by the system under test. The reference is not shown. The subjects continuously rate the perceived quality on the continuous scale from "bad" to "excellent" using a slider.

f) Simultaneous Double-Stimulus for Continuous Evaluation (SDSCE)

The subjects watch two sequences at the same time. One is the reference sequence, and the other one is the test sequence. If the format of the sequences is the standard image format (SIF) or smaller, the two sequences can be displayed side by side on the same monitor; otherwise two aligned monitors should be used. Subjects are requested to check the differences between the two sequences and to judge the fidelity of the video by moving the slider. When the fidelity is perfect, the slider should be at the top of the scale range (coded 100); when the fidelity is the worst, the slider should be at the bottom of the scale (coded 0). Subjects are aware of which one is the reference and they are requested to express their opinion while they view the sequences throughout the whole duration.

Tsung-Jung Liu received the B.S. degree in Electrical Engineering from National Tsing Hua University, Hsinchu, Taiwan, in 1998, and the M.S. degree in Communication Engineering from National Taiwan University, Taipei, Taiwan, in 2001, respectively. Now, he is pursuing the Ph.D. degree in Electrical Engineering at University of Southern California (USC), Los Angeles. He is currently a Research Assistant with Signal and Image Processing Institute, Ming Hsieh Department of Electrical Engineering, USC. He was also a Visiting Student with the School of Computer Engineering, Nanyang Technological University, Singapore from June 2011 to July 2011. His research interests include machine learning, perceptual image/video processing, and visual quality assessment.

Yu-Chieh Lin is a Ph.D. student in majoring Electrical Engineering at the University of Southern California. He received B.A. and M.S. degrees from the National Chiao Tung University, Hsinchu, Taiwan in 2004 and 2006. From 2007 to 2011, he was a Senior Engineer of Corel, working on optimizing video codecs and building new features with AMD, Intel, and Nvidia. His accolades include the recipient of Government scholarship to study abroad in 2011, the winning award in National Micro-Computer Applied System Design Production Contest in 2005, and the merit award of 4th MXIC Golden Silicon Awards in 2005. His current research interests are image/video quality assessment and video coding theory.

Weisi Lin received his Ph.D. from King's College London. He was the Lab Head and Acting Department Manager for Media Processing, in Institute for Infocomm Research, Singapore. Currently, he is an Associate Professor in Computer Engineering, Nanyang Technological University, Singapore. His research areas include image processing, perceptual multimedia modeling and evaluation, and video compression. He published 240+ refereed papers in international journals and conferences. He is on the editorial boards of IEEE Trans. on Multimedia, IEEE SIGNAL PROCESSING LETTERS and *Journal of Visual Communication and Image Representation*. He chairs the IEEE MMTC IG on Quality-of-Experience. He has been elected as an APSIPA Distinguished Lecturer (2012/3). He is the Lead Technical-Program Chair for Pacific-Rim Conference on Multimedia (PCM) 2012, and a Technical-Program Chair for *IEEE International Conference on Multimedia and Expo* (ICME) 2013. He is a fellow of Institution of Engineering Technology, and an Honorary Fellow, Singapore Institute of Engineering Technologists.

C.-C. Jay Kuo received the B.S. degree from the National Taiwan University, Taipei, Taiwan, in 1980, and the M.S. and Ph.D. degrees from the Massachusetts Institute of Technology, Cambridge, USA, in 1985 and 1987, respectively. He is Director of the Media Communications Laboratory and a Professor of electrical engineering, computer science, and mathematics at the University of Southern California, Los Angeles, USA, and the President of the Asia-Pacific Signal and Information Processing Association (APSIPA). His research interests include digital image/video analysis and multimedia data compression. He is a co-author of about 210 journal papers, 850 conference papers, 50 patents and 12 books. He has guided 115 students to their Ph.D. degrees and supervised 23 postdoctoral research fellows. He is Editor-in-Chief for the *IEEE Transactions on Information Forensics and Security* and Editor Emeritus for the *Journal of Visual Communication and Image Representation*. Dr. Kuo is a Fellow of AAAS, IEEE and SPIE.

https://doi.org/10.1017/ATSIP.2013.5 Published online by Cambridge University Press