Signal Processing: *Image Communication* 25 (2010) 469–481

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

No-reference image and video quality estimation: Applications and human-motivated design

Sheila S. Hemami [a], Amy R. Reibman [b,*]

[a] *Cornell University, USA*
[b] *AT&T Labs-Research, USA*

ARTICLE INFO

Keywords:
No-reference
Video quality
Quality metrics
Quality estimator
Applications of quality metrics
Blind quality assessment

ABSTRACT

This paper reviews the basic background knowledge necessary to design effective no-reference (NR) quality estimators (QEs) for images and video. We describe a three-stage framework for NR QE that encompasses the range of potential use scenarios for the NR QE and allows knowledge of the human visual system to be incorporated throughout. We survey the measurement stage of the framework, considering methods that rely on bitstream, pixels, or both. By exploring both the accuracy requirements of potential uses as well as evaluation criteria to stress-test a QE, we set the stage for our community to make substantial future improvements to the challenging problem of NR quality estimation.

1. Introduction

Quality estimators (QE) for images and video have been the topics of numerous recent and not-so-recent surveys [1–4]. This paper approaches the topic by providing a survey of what the authors consider to be basic background knowledge for the design of an effective QE. While a plethora of terms have been used in conjunction with *quality* to describe quality estimation and quality estimators (e.g., analysis, assessment, evaluation, measurement, metric, among others), we selected the verb *estimate* to reflect the statistical nature of the ground-truth subjective scores which a quality estimator strives to predict.

The goal of a QE is to characterize the quality of a test image or video, $v = v_{test}$, which is typically the output of a system. If $Q_{subj}(\cdot)$ is the measured, perceived quality as estimated using an appropriate subjective experiment, then an ideal objective QE produces objective scores $Q_{obj}(\cdot)$ which perfectly predict subjective scores $Q_{subj}(\cdot)$ for all inputs. This is clearly challenging, as it requires not only that the QE be accurate for a wide range of input content and processing types, but also that the QE take into account a variety of environmental viewing conditions and a variety of viewers with disparate experience, expectations, and involvement.

To tackle this challenge, QE designers have taken a number of approaches to restrict the problem. One approach, taken by *full-reference* (FR) quality estimation, measures the quality of the test image or video v_{test} relative to that of a reference v_{ref}. A distorted image or video v_{test} is written as the sum of an original v_{ref} plus distortions d

$$v_{test} = v_{ref} + d. \tag{1}$$

FR QEs have the original signal v_{ref} given as *a priori* information, so they are able to compute d exactly. However, a *distortion* as defined above is not necessarily visible. Therefore, the mere presence of d does not imply that subjective quality is degraded. FR QEs that use models of the human visual system (HVS) attempt to partition the distortions into those that are visible and those that are nonvisible as

$$v_{test} = v_{ref} + (d_{nonvisible} + d_{visible}), \tag{2}$$

* Corresponding author.
E-mail address: amy@research.att.com (A.R. Reibman).

0923-5965/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2010.05.009

using experimental evidence of HVS sensory mechanisms. In FR QEs, the original signal is considered to be a mask, and the goal is to determine how the mask affects the distortions that are introduced by the processing chain.

In practice, however, FR algorithms are only applicable when both v_{test} and v_{ref} can be made available at the same physical location. In addition, one fundamental assumption of most FR QEs is that the original v_{ref} has maximum quality. This assumption may be violated when v_{ref} represents the image or video prior to an enhancement algorithm (e.g., edge enhancement), or even prior to applying high-rate quantization, which may have a denoising effect.

In contrast, in *no-reference* (NR) quality estimation, neither v_{ref} nor d is available. As such, NR QEs are the most broadly applicable type of QE, and are the focus of this paper. Without either v_{ref} or d, NR QEs must distinguish the visible distortions from the rest of the signal:

$$v_{test} = (v_{ref} + d_{nonvisible}) + d_{visible}. \tag{3}$$

Thus, designers of NR QEs face additional challenges beyond those mentioned above. Using limited input information, NR QE must be able to distinguish signal from visible distortion, when varied processing (e.g., encoding, transmission) introduces different *artifacts* (e.g., blocking or blurring), into a wide range of source content. Further, they must achieve this despite the fact that many desired signals may “look” very similar to typical artifacts.

In this paper, we delineate three approaches that have been effectively used in the design of NR QEs to address these challenges. First, a NR QE can restrict the domain of the problem based on the desired use of the QE. Instead of striving for perfect accuracy, QEs can be designed for the more realistic performance goal: to achieve the *required accuracy* for its *application* over the *set of input content* and *artifacts* for which it was designed. Second, using knowledge of the expected processing and the expected signals, sophisticated signal and artifact models can be developed to improve NR QE design. Third, NR QEs gather as many sources of information as possible in addition to v_{test}, including assumptions about the processing and information about the bitstream.

In addition, it is paramount for NR QE to incorporate as much information about the human visual system (HVS) as possible. A complete model of the human is not possible; for example, feelings or emotions evoked by content are extremely difficult to predict and quantify, and aesthetics are also highly observer-dependent [5–8]. However, substantial modeling of the human observer has been performed by various communities, including psychology, vision, and photography, and this large body of work should inform QE design. While FR QEs have focused on including low-level psychophysics, NR quality estimation provides the opportunity to include work rooted in the photographic community, which is inherently no-reference.

We begin this paper by describing a three-stage framework for quality estimation in Section 2, which includes measurement, pooling, and mapping to quality. Because NR QEs need only be as accurate as the application for which they are designed, we next describe in Section 3 a range of applications of NR QE, including algorithm optimization, benchmarking, and outage detection. Each application has different requirements for the set of input content $\mathcal{V}$ and the set of artifacts $\mathcal{A}$ over which it must be accurate. Section 4 briefly describes the variety of artifacts that may be introduced by different processing. Next, since models of human vision and perception should be integrated at all three proposed stages of a NR QE, in Section 5 we discuss three levels of human-based models, including *subjective* quality evaluation for the final crucial stage. However, since most of the attention to date has focused on the first measurement stage, we provide in Section 6 a brief survey of current approaches to measurements for NR quality estimation. We discuss appropriate performance evaluation for QEs in Section 7, and provide concluding thoughts in Section 8.

2. A framework for no-reference quality estimation

A generic no reference quality estimator consists of three steps: *measure*, *pool*, and *map to quality*. In this framework, the input and the corresponding quality estimate can correspond to an entire video, several frames, a single image, or a segment of an image. Input data to the system consists of one or more sources of actual or estimated information, depending on the quality estimation application. The input includes

- the pixels corresponding to v_{test};
- the bitstream corresponding to v_{test}, including packet headers and data encoding parameters (e.g., quantizer step sizes, picture types);
- assumptions about the statistics of the original signal v_{ref} or the class of signals $\mathcal{V}$; and
- assumptions about the distortions and/or artifacts in v_{test}.

Note that v_{test} itself is not necessary; some techniques perform NR quality estimation without considering the actual pixel data.

Furthermore, since NR QEs attempt to estimate perceived quality as measured by an appropriate subjective experiment, models of and information about human perception, preferences, and ground-truth quality scores should be incorporated into each of the three components. Each of these issues will be discussed in greater detail in Section 5; here, we simply indicate which issues are appropriate for each component.

Measuring computes physical quantities (we will refer to them as *features*) using the input. Selection of specific quantities to compute can be guided by the previously mentioned assumptions as well as the ultimate application of the quality estimator. Both perception models and preference models can be incorporated into the measurement. An example of the former is identifying the presence of known artifacts and estimation of the visibility threshold for those artifacts, while an example of the latter is edge sharpness. Multiple measurements can be computed.

Pooling combines the possibly linearized measurements over an appropriate subset of space and/or time for the QE. For example, in an individual image, spatially local pooling over frequency and orientation results in a spatial map of responses. Subsequent pooling over space produces a single response estimate. In video, pooling can be performed over combinations of spatial frequency, orientation, space, and time. As described in Section 3, the subset of space and/or time will be defined by the application of the QE.

The pooling step includes an optional linearization process that consists of a nonlinear mapping for each measurement to rescale or renormalize the values to an appropriate scale (e.g., the same dynamic range or just-noticeable-differences (JNDs)) prior to pooling.

The pooling mechanism should be motivated by the measurements and known properties of human observers. Interactions among artifacts and between artifacts and the image itself should be accounted for in pooling, using characterizations of *masking* in human vision (cf. Section 5).

Linear pooling can be appropriate when the individual measurements satisfy the assumptions required by linear regression. Minkowski summation is more general and is motivated by additivity in low-level vision; its use requires selection of an appropriate exponent which should be based on measurements. Temporal pooling combines multiple frames into a score for each relevant time scale. The pooling may be linear, Minkowski summation, or a maximum-type operator, with the goal of incorporating both temporal masking and temporal summation. Statistical learning techniques can also been applied to combine measurements, but these require appropriate analysis to provide insight into whether all measurements are truly contributing and are in fact behaving as desired.

The last component, *mapping to quality*, applies a nonlinearity to map the output of the pooling component into an estimate of perceived quality. If the output of the pooling element is already linear, this stage may not be necessary. The exact form of the nonlinearity should be dictated by a best-fit of the output of the pooling component to ground-truth subjective data. As such, assumptions regarding the "best" and "worst" expected qualities in a system are implicitly included in the QE through the training data. It is extremely important that the training data be appropriate for the measurements and for the application (cf. Section 7 on evaluation of QEs).

This final component is required because even when the previous two components are typically sufficiently accurate to provide monotonicity and approximate rank-order preservation, they may still not accurately map to true quality scores (e.g., one end of the scale is often inappropriately compressed or expanded).

3. Use scenarios for quality estimators

Before content is finally displayed to a viewer or customer, it may undergo a wide variety of different types of processing, by different algorithms in different subsystems, many introducing distinct artifacts. In addition, it may pass from one owner to another; different companies may be responsible for different stages of the processing or different stages of the delivery. Quality estimation is appropriate whenever content is passed from one owner or entity, one piece of hardware, or one algorithm to another. However, a QE that is useful for one application may not be appropriate for another. In this section, we discuss and differentiate a variety of applications.

Each application is characterized by several aspects: the set $\mathcal{V}$ of sources (i.e., undistorted images or videos), the set $\mathcal{A}$ of *artifacts* for which the QE must be accurate, the degree of accuracy required, and (for video) the time scale at which quality values are required.

Algorithm optimization in processing employs QEs in a closed-loop during compression or other processing algorithms, to maximize quality of the output. Such QEs can be either FR (e.g., in compression, in which the original is obviously available) or NR (e.g., for enhancement at the decoder, or for transcoding applications in the network).

In such "in the loop" applications, the set $\mathcal{A}$ is limited to known artifacts resulting from the processing algorithm. The QE must correctly reflect increases or decreases in these known artifacts for a single source at a time, and hence the set $\mathcal{V}$ contains only one undistorted source. If the QE incorrectly characterizes quality increases or decreases, then incorrect decisions may be made during optimization that will result in worse visual quality than without optimization.

A QE for algorithm optimization may or may not contain a real-time requirement, depending on the application. Real-time video encoding clearly imposes both causality as well as computational restrictions on a QE, while image or off-line video processing does not.

Product benchmarking allows purchasers and marketers to compare the performance among different products or components. Product marketers use benchmarking with the goal of demonstrating superiority of their product over others. Benchmarking is typically applied to an individual algorithm (e.g., encoding) or a small bundle of algorithms (e.g., decoding and error concealment), although it could plausibly also be applied to a business owner's subsystem (to compare, for example, a cable TV offering to a DSL TV offering). A desirable benchmarking statement compares two products p_1 and p_2, producing processed content $p_1(v)$ and $p_2(v)$, respectively, and may take either the form

$$Q_{subj}(p_1(v)) > T \ for \ \alpha\% \ of \ v \in \mathcal{V}, \tag{4}$$

or

$$Q_{subj}(p_1(v)) > Q_{subj}(p_2(v)) + \delta \ for \ \beta\% \ of \ v \in \mathcal{V}, \tag{5}$$

where T is a quality threshold. If p_1 is the marketer's product and p_2 belongs to a competitor, a marketer may carefully select the set of sources $\mathcal{V}$ for which α, β, and δ are as large as possible. Selection could be content-based (e.g., sports and action) or even specific source-based (e.g., *Monsters Inc.* and *Raiders of the Lost Ark*). For statements of the form of (4) and (5), only one QE score per input is required regardless of sequence duration. An appropriate QE will be accurate for heterogeneous artifacts so that it can compare systems, for example, both with and without deblocking.

System provisioning occurs prior to deployment, and involves the design of an end-to-end system to achieve a target quality. A typical problem statement seeks a set of system parameters $\{\phi\}$ (e.g., bit-rate, maximum packet loss rate, server capacity, or spatial resolution or temporal resolution) for a system s, according to

$$\text{Determine}\{\phi\} \text{ for which } Q_{subj}(s(v)) > T \text{ for } \alpha\% \text{ of } v \in \mathcal{V}, \tag{6}$$

where for video, $Q_{subj}(\cdot)$ operates on multiple time scales. Here, $\mathcal{V}$ is representative of all content that will be handled by the system. To maximize system robustness, both the minimum quality over short time intervals and the average quality over longer time intervals are of interest.

While (6) looks fairly similar to (4), several important aspects distinguish the applications. First, a QE for system provisioning needs only be accurate near the system design point threshold T. Second, a QE for system provisioning must be accurate across processing that results in different spatial (and for video) and/or temporal resolutions such that when resources are constrained, bandwidth-reducing decisions are made that retain the best perceived quality. Third, for system provisioning, the set $\mathcal{V}$ is large and must be inclusive of all possible types of content likely in the system, while product marketing focuses on choosing a subset of sources tailored to the product's strengths.

Content acquisition and delivery are important for their use in *service level agreements* (SLAs), which are contracts typically between business entities. For example, for video delivery to the home, it is common to have either implicit or explicit contracts between consumers, service providers, content providers, and network providers. SLAs and other legal contracts constrain the quality of both the incoming and outgoing material.

A QE for *content acquisition* and *content delivery* determines if either the incoming material (whether from a camera, e.g., [9] or at the input of a large-scale content delivery network) or the outgoing material (i.e., to another network provider or to the viewer's end-system) has sufficient quality. For acquisition and delivery of images, a QE must

$$\text{Alarm when } Q_{subj}(v) < T \text{ for more than } \alpha\% \text{ of images.} \tag{7}$$

Video acquisition and delivery require ongoing monitoring, as does *outage detection*, which considers substantially larger degradations including the loss of video entirely (termed *blackout*). For these applications, a video QE must

$$\text{Alarm when } Q_{subj}(v) < T \text{ more than } N \text{ times in } t \text{ seconds.} \tag{8}$$

QEs for outage detection and content acquisition must be accurate for *any* content and *any* type of artifact, including those that represent acquisition failures. For example, the same end-user perception of outage occurs either when a video server fails or when, despite correct coding and transmission, video is received that was captured without removing the lens cap. QEs for these applications must minimize the instances of false alarm and of missed detection.

Transient quality failures can be seen neither by periodic assessment (i.e., checking once an hour) nor by time-averaged assessment (i.e., quality averaged over an hour). As such, the time scale at which video QE is performed for these applications includes seconds, minutes, hours, and days. Accurate operation over such a large range of time scales is challenging—while accuracy over second or minutes can be evaluated easily in a development environment, evaluating accuracy over hours and days may require more sophisticated approaches.

Troubleshooting occurs after outage detection, to pinpoint the cause of the problem, so that it can be fixed. The objective, identifying why $Q_{subj}(v) < T$, can be addressed using a set of artifact detectors. Troubleshooting requires artifact detectors that operate independently; a ringing artifact should not influence the output of a noise detector (e.g., [10]).

Summary: QEs have a wide range of applications in both processing and transmission, differing in set $\mathcal{V}$ of sources, the set $\mathcal{A}$ of artifacts, the degree of accuracy required, and the time scale of operation. Designing a QE for a particular application should consider these requirements. Next, we present specific artifacts and describe how they are introduced in the processing chain.

4. Artifacts in the processing chain

The processing chain encompasses acquisition, compression, transmission or storage, decoding, and display, and artifacts can be introduced at various stages. At any point in the chain, the content can be repurposed, which can entail re-acquisition, re-compression, or additional transmission. In addition, at any point an enhancement algorithm can be applied.

Acquisition and display are inherently without reference. Compression of originals has a reference; transcoding does not. Transmission or storage can result in lost or errant packets or bits, which induce decoding errors later in the chain; quality estimation at the decoder is also most commonly without reference.

The interested reader is referred to the surveys of artifacts in [11–13] for additional information and visual examples.

4.1. Image and video acquisition and display

The two ends of the processing chain are inherently without reference. An important aspect of both of these operations is the treatment of color data, both its appropriate interpretation during acquisition and its subsequent appropriate rendering during display. As such, many aspects of color image workflow are no-reference quality estimation problems, and these are approached with human-centric goals—to represent colors as they would have been perceived by a human observer, and to then display colors in a perceptually pleasing manner.

During acquisition, artifacts may be introduced due to the optical lens, the density and accuracy of the sensing elements, or the digitization process. *Blurring* can be introduced by defocus (focal blur) or due to camera or object motion with too slow a shutter (motion blur). *Noise* can be introduced in the sensing elements. Insufficient dynamic range in A/D conversion can lead to *contouring*. Insufficiently dense sampling results in *aliasing*, which has a variety of artifacts including jagginess, geometric distortions, and inhomogeneity of contrast [14] as well as color artifacts. Failures in camera-based algorithms including white point selection and balancing and exposure adjustment, among others, can result in images which inaccurately represent colors and scene brightness as seen by a human observer.

Artifacts caused by the display are difficult to measure but can be estimated if appropriate display parameters are known. Such artifacts include LCD motion blur, overscan, and potential interlacing artifacts when interlaced video is displayed on progressive monitors. Inaccurate display characterization can cause problems in both tone mapping and gamut mapping, in which brightness and colors of the content are mapped to those of the display.

4.2. Encoding and decoding

Compression with block-based coders (JPEG, MPEG-2, H.261, H.263, H.264) can introduce a number of artifacts. *Blocking* appears at deterministic locations and is caused by heavy quantization of the transform coefficients. *Mosquito noise* is temporal shimmering caused by time-varying blockiness. False edges, also called motion-compensated edge artifacts (*MCEA*), are the result of blocking artifacts that move away from the block boundaries due to motion compensation process [15]. Flatness [16] is a lack of resolution in fine detail.

Wavelet-based coders (including JPEG-2000) introduce a different set of artifacts, including blurring, ringing, and aliasing. Both wavelet and block-based coders may skip frames during compression, making video appear jerky.

Perfectly received data undergoes no decoder-induced artifacts. If bits or packets are damaged or lost in transmission or storage, artifacts introduced at the decoder are very dependent on the encoding strategy, the decoder design, and error concealment strategies. For JPEG and JPEG-2000 images, decoding artifacts can include DC shift caused by DPCM decoding errors, horizontal or vertical shifts of image data within an image, and loss of detail.

For motion-compensated video, artifacts can include motion jerkiness resulting from dropped frames, individual frames exhibiting concealment distortions, concealment distortions which propagate over time, "missing" blocks displayed as solid colors, and propagation of such missing blocks over time. Hardware faults may also occur in video decoders, introducing a range of artifacts considered in [17,18].

4.3. Repurposing and enhancement

The most frequent type of repurposing is displaying at low resolution (for example, on a mobile device) content that was acquired at higher resolution. Scalable coding implicitly establishes a framework for repurposing; selective discard of scalably coded bitstreams during transmission is simply repurposing.

Processing for repurposing includes spatial resampling, temporal resampling and frame-rate conversion, re-compression or transcoding. Thus, repurposing may introduce some artifacts already discussed. Additional examples include interlace artifacts in video frames treated as still images and a variety of artifacts that occur when converting video from high-definition (HD) to standard-definition (SD) or vice versa (incorrect aspect ratios, frame truncation, or color artifacts).

Digital video that has been reacquired from analog video or film can exhibit some unique artifacts. Demodulated analog NTSC or PAL video may have "rainbow" effects [19] where color artifacts appear in regions of high luminance spatial frequency, or luminance artifacts appear where colors are saturated. Analog multipath transmission can result in ghosting. Film degradation introduces a wide variety of artifacts [20] including *flicker*, a fluctuation of picture brightness.

Operator or equipment error during repurposing can cause additional artifacts after decoding, such as the two fields of a frame being presented in the wrong order. Visually, the entire frame appears to have interlacing artifacts.

Enhancement may also introduce artifacts that were not present previously. Sharpening can cause ringing; deblocking and denoising can cause blurriness; de-interlacing can cause ghosting or motion artifacts.

5. Modeling humans in quality estimator design

Because NR quality estimation attempts to estimate perceived quality by a human observer, models of and information about human perception, preferences, and ground-truth quality scores should all be incorporated into a NR QE. This data includes

- low-level psychophysical models, which can be used to estimate $d_{visible}$;
- measured sensitivities to particular artifacts, which can also be used to estimate $d_{visible}$;
- known preferences for "perceptually pleasing content," including that on colorfulness, sharpness, degrees of blurring, addition of noise; and
- ground-truth subjective quality scores associated with a database of training content.

This section first reviews fundamentals of low-level vision as they are applicable to quality estimation. Here, psychophysical experiments measure responses of the human visual system to simple stimuli such as sinusoids or spatially correlated bandlimited noise. The results provide a characterization of vision which can be applied in a “bottom-up” manner to complex stimuli such as processed images and video. Low-level vision is most commonly used in FR QEs, but is also clearly relevant to NR QEs.

Several alternative “top-down” approaches are next reviewed. Responses are measured to stimuli that consist of natural images or video processed to include one or more synthetic or actual artifacts. We include the study of preferences in this approach, which is rooted in the photographic community and is inherently no-reference.

This section concludes with a brief discussion on ground-truth data which is used for both training and validation of any QE.

5.1. Low-level vision

Low-level vision is generally thought to perform a multichannel decomposition, where bandlimited channels process spatial frequency, temporal frequency, and color. With respect to quality estimation, not only are the responses of each channel relevant, but so are intra- and inter-channel interactions. The former permit an estimation of the HVS’s response to bandlimited, simple stimuli, while the latter (more commonly known as *masking*) permit an estimation of its response to compound stimuli such as images and video. Masking is a general term that refers to the perceptual phenomenon in which the presence of masking signal (the *masker*) reduces a subject’s ability to detect a given signal (the *target*). With respect to quality estimation, an estimate of masking is essential to separate distortions into visible distortions and those distortions that are masked (i.e., hidden) by the source.

While a thorough review of low-level vision is beyond the scope of this paper, this section provides a brief overview. Readers are encouraged to further explore not only the references in this section, but also several FR video quality estimators which provide different design decisions in implementing a HVS model [21–24].

5.1.1. Contrast

While digital pixels are stored as bits, luminance (measured in candelas/meter^2) represents the light entering the eye, and contrast contributes to the *perceived* luminance. Because perceived quality is experienced by a viewer, any quality estimator must include the display device in computing how the stored data is displayed to the viewer.

Contrast is qualitatively defined as *luminance change* divided by *mean background luminance* and can be computed globally or locally on a natural image. Many definitions of contrast exist (e.g., the Weber fraction, Michelson [25], bandlimited contrast [26], local bandlimited contrast [27], RMS [28]; see [27] for a review), leaving flexibility to select the appropriate definition based on the needs of a particular application.

5.1.2. Spatial vision

The human contrast sensitivity function (CSF) is a well-accepted description of spatial frequency perception; the HVS has band-pass characteristics. The *multi-channel model* postulates that the CSF represents the aggregate response of frequency- and orientation-tuned individual *channels* having increasing bandwidth with increasing frequency [29].

Current explanations of spatial masking can be divided into four paradigms: (1) *Noise masking* [30]; (2) *contrast masking* [31–34]; (3) *entropy masking* [35]; and (4) *structural masking* [36].

5.1.3. Temporal vision

While spatiotemporal [37,38] or spatiovelocity [39] responses have been measured, HVS-based QEs typically apply a separate temporal frequency response. The HVS can be modeled as having both transient (i.e., bandpass) and sustained (i.e., lowpass) temporal response mechanisms [40–42]. For examples on how spatial and temporal frequency responses can be combined in QEs, the reader is directed to four examples [21–24].

Temporal summation and temporal masking have also been measured and modeled [43,44].

5.1.4. Color vision

The eye contains three cone types with different spectral sensitivities, colloquially known as *red*, *green*, and *blue*. Opponent color theory [45] suggests and psychological experiments have demonstrated that the visual system has three color channels which are roughly independently processed at a low level (e.g., [46,47]). These channels represent achromatic vision and two chroma channels: red-green, and blue-yellow.

While CSFs have been measured for the red-green and blue-yellow channels [48,49], color channel sensitivity has been substantially less studied than luma channel sensitivity. The chrominance CSFs differ from that of luminance in that they are low-pass rather than band-pass, and they fall off sooner than the luminance CSF. Temporal responses for red-green and blue-yellow have also been measured [50].

Interactions between color and luma channels are also not well understood, but color provides substantially more masking of luminance than the reverse (see [51, Chapter 7]).

5.1.5. Pooling

Estimates of responses in the spatial, temporal, and possibly color channels must be combined, or *pooled*, to provide an aggregate response estimate. A Minkowski sum is most commonly used in modeling low-level vision.

5.2. Top-down perception

Due to the challenges associated with applying a low-level characterization of vision to complex stimuli, especially when no reference is available, other approaches have been explored in which the images and video themselves are used as the stimuli. One benefit of this approach is avoidance of explicit masking (and sometimes pooling) models; unfortunately, the knowledge gained is limited to the specific preference or artifact under test and can also be limited to the specific source content used.

5.2.1. Preferences

Preferences refer to characteristics of an image or video that can be computed by some measure and quantified as statistical functions of large groups of observers. They have a historical basis in photography, and are therefore inherently no-reference. Much work in color representation and reproduction is based on human preferences, including color scaling and color naturalness [52,53]; edge sharpness [54], color saturation, flesh tone preferences, and use of dynamic range [55].

5.2.2. Quantifying responses to specific artifacts and impairments

A second "top-down" approach directly measures observer responses to the artifacts likely to be encountered in a system or application, including all artifacts mentioned in Section 4. Single artifacts may be generated synthetically [56] in an attempt to understand their impact in isolation. However, measuring human responses to individual artifacts does not provide information on cross-impairment masking or on how observers perceive two or more simultaneously presented artifacts.

While some experiments have inserted multiple synthetic artifacts and quantified the simultaneous response [57], it is more common for experiments to employ stimuli produced by systems, e.g., compressed MPEG video having undergone packet losses. Many processing techniques create impairments that are strongly correlated with the source content v_{ref}. Often, such experiments are also designed to measure the impact of the content on the response, and hence are essentially quantifying masking. In contrast to masking experiments for low-level vision, however, these results are less generalizable.

As examples, responses have been measured to freeze frames [58–60]; synthetic blockiness, blurriness, noisiness and ringing [56,57], MPEG-2 compression impairments localized in both space and time [61], packet loss [62] and its visibility [63,64].

Another approach to understand human perception for a type of processing (e.g., image scaling or JPEG compression) that creates multiple artifacts with complicated interrelationships is to use naive viewers to label images in terms of their perceived *quality* and to have expert viewers label *artifact strength* [14]. A regression analysis then determines the impact of the latter on the former. These results are also difficult to generalize beyond the particular experimental setup.

Lastly, some experiments vary parameter settings for the processor, and evaluate the subjective response. Many studies take this approach, including [65] for compression at different bit-rates and [66,67] for packet loss impact at different packet loss rates. It is difficult to generalize the results of these subjective tests to other processing, to other parameter settings, and most importantly, to different sources in $\mathcal{V}$.

5.2.3. Multidimensional scaling (MDS)

Multidimensional Scaling [68,69] is a statistical technique for quantifying responses to multiple preferences and/or artifacts. With an input matrix of distances between stimuli (e.g., resulting from a perceptual experiment), it attempts to find a coordinate system in N-dimensional space which preserves the distances between the stimuli (N is user-defined). However, the dimensions themselves may not be perceptually meaningful. Examples include [70–73].

5.3. Subjective estimation of quality—ground truth data

An ideal quality estimator predicts quality estimates as measured by an appropriate experiment with human subjects. As such, a quality estimator must be designed and evaluated using ground-truth subjective data gathered from observers. The subjective experiment is also critical for defining the scope of a estimator, and in particular for understanding both appropriate and potentially inappropriate uses of an estimator.

The performance of an estimator is limited by the nature of the data to which it has been tuned. While many estimators provide rank-ordered assessments that match those of human observers on images that contain differing amounts of a single artifact (e.g., JPEG compression) they are not as successful at rank-ordering degraded images from the same original that have different artifacts (e.g., comparing JPEG distortions with white noise). One reason for this weakness is a lack of actual observer scores for such comparisons and hence a lack of accurate training data. Use of a protocol such as SAMVIQ [74], in which observers simultaneously view and score all distorted versions of a source, avoids this problem. Comparisons between different distortions can also be made using the quality ruler protocol [75].

A full discussion of the design of subjective tests for gathering ground-truth subjective data for quality estimation is beyond the scope of this paper. The reader is referred to the VQEG committee documents which provide excellent "case study" discussions of subjective test design for three VQEG test phases [76–78], as well as various international standards (e.g., [79,80]) and other references [81,82].

We list below various issues for test design which should be carefully considered prior to beginning testing:

- selection of a testing protocol, including use of category ("excellent," "good," etc.) or continuous (non-quantized) ratings, and whether evaluations are done with respect to a reference image/video (e.g., double stimulus protocols) or singly;
- collection of data at relevant (and potentially multiple) time scales for video;
- the required number of observers, and a methodology for determining the validity of particular observers;
- choice of subject matter, including use of gray-scale or color images;
- environmental viewing conditions, including background and room lighting, display calibration, and viewing distance;

- observer instructions and clarity of task wording; and
- human subjects approval of the protocol by an appropriate body at the researcher's institution.

6. Overview of existing measurements for NR quality estimators

A large body of work addresses various aspects of NR QE design. Much of this work, however, does not in fact estimate *quality*. Rather, it stops at the measurement step, having computed a single feature, without any inclusion of human observer data. Other works compute a single feature and then map to quality, thus limiting the pooling stage to spatial or temporal averaging. Nevertheless, when considering this body of work, substantial progress has been made toward developing solutions to the first step of the framework for NR QE described in Section 2. Therefore, in this section we briefly review some of the literature studying this measurement step.

6.1. Direct estimation of mean-squared error

We begin with a class of NR QEs whose measurement stage attempts to separate v_{test} into v_{ref} and distortions, d, statistically. While many of these types of NR QEs to date only consider this first stage, we also describe two examples where HVS models or subjective data are incorporated into subsequent processing.

The first methods in this class of NR QE were designed to predict the Mean-Squared-Error (MSE) caused by block-based *compression* like MPEG-2 [83–87], JPEG [87,88], or H.264 [89,90,87,91,92]. With the exception of [84], which uses the decoded pixels v_{test}, these techniques use information only from the received bitstream. The basic approach is to model the DCT coefficients using a Laplacian distribution, and estimate the Laplacian parameter for each of the 8 × 8 coefficients. However, this has been extended to generalized Gaussian [90] and Cauchy distributions [91,92] as well.

These techniques have the common drawback that they only obtain an MSE estimate for each 8 × 8 block; they are unable to predict neighboring pixel differences and hence be extended to estimate artifacts like blockiness. In addition, the accuracy of the estimated MSE for each of these methods is lower when the bit-rate is smaller, due to the presence of more coefficients quantized to zero. To improve the accuracy for low bit-rates, [91,88] rely on training data to obtain improved estimates of the coefficient distributions.

There are also several attempts to design NR QE to predict the Mean-Squared-Error (MSE) caused by *packet loss* errors. Bitstream-only approaches are designed in [93,94] for motion-compensated video compression with packet loss. Content-specific information is extracted from the parsed bitstream, to estimate local means, variances, and correlations. Together with extracted motion vectors, these are combined using a Gauss–Markov model to estimate initial MSE. Motion-compensated error propagation is incorporated into the overall estimate of MSE. In [95], both v_{test} and information extracted from its bitstream are combined to estimate MSE due to packet loss for H.264. The initial error is estimated by separately considering the impact of missing motion vectors and missing prediction errors. Finally, [96] estimates the MSE due to packet losses in motion-JPEG2000.

Noise estimation approaches estimate MSE using two basic methods [97]. The first is to smooth v_{test} and define any difference between v_{test} and its smoothed version to be noise [97,98]. The second is to identify smooth areas in v_{test} and assume that any variation within those smooth areas is noise [99–102].

The strategy of estimating MSE is motivated by a desire to statistically estimate the distortion d in Eq. (1). Unfortunately, most contributions in this area are limited in that they only estimate MSE; they do not further partition the estimated MSE into $d_{visible}$ and $d_{nonvisible}$. However, Brandão and Queluz [88] also incorporate their estimated error into a HVS-based NR QE relying on Watson's DCT-based perceptual model [103]. Whereas [103] uses the actual quantization error computed in a FR framework, [88] uses the NR estimated quantization error. In addition, the results of [93] were later incorporated by Kanumuri et al. [63] into a NR estimator of the visibility of packet losses, whose pooling step uses subjective data for training.

6.2. Feature-based approaches

Two strategies have been used to design features to be extracted in the measurement step. Both strategies assume that the statistics of v_{test} differ from those of v_{ref} and use features extracted from v_{test} to evaluate model compliance. The first strategy is to develop a model for the specific artifacts that may contribute to $d_{visible}$, focusing on those artifacts introduced by the processing chain. As such, NR measurements designed using this strategy may generalize to different classes of content $\mathcal{V}$, but they are unlikely to be able to characterize quality degradations caused by different artifacts. This approach will fail if $\mathcal{V}$ contains v_{ref} that mimic the artifacts (for example, periodic structure of vertical edges near block boundaries).

The second strategy is to model specific signal attributes that characterize v_{ref}. The goal is then to find violations of the signal model. This strategy focuses on a specific class of $\mathcal{V}$ (for example, scenes without man-made structures, or scenes with consistent lighting), and is likely to be effective for a variety of artifact types. This approach will fail if the added distortions do not cause v_{test} to violate the signal model.

Artifact and signal models can be developed in either the spatial and the transform domain, where the latter includes DCT, wavelet, and polynomial transforms. Complementary features extracted from each domain can be combined to improve overall QE accuracy.

6.2.1. Spatial artifacts due to compression

Many *blockiness* features have been surveyed in [104,15]. Among the NR QE that consider some perceptual masking, [105–107] compute a local gradient, [108] estimates the power of an ideal blocking signal using the FFT, [109,110] explore features of the DCT, and [111] apply the first three coefficients of a polynomial transform. Most blockiness detectors assume the grid-location is known; [112] detects the grid in case the image has been resized or cropped, and [15] considers motion-compensated edge artifacts, which result when block edges are motion-compensated away from block boundaries.

A detailed overview of 13 QEs that measure *blurriness* (or sharpness) of images is presented in [113], which introduces the notion of just noticeable blur. Many approaches measure physical attributes of the edge profile, including acutance [114], horizontal and vertical edge extent [115,113,116], and diagonal edge extent [117]. Using polynomial transforms, [118] designs a multiscale blur estimation algorithm to estimate the spread of a Gaussian blurring kernel. A similar goal was addressed in [119] using spatial gradients for both lines and edges. Blur has also been measured using features in the DCT domain [120,121] and using the phase of the Fourier transform [122]. A combined spatial and frequency domain approach is presented in [123], where features about the edge profile are combined with the local frequency spectrum around image edges.

With the exception of [98], which operates in the frequency domain, *ringing* features are typically extracted from the pixel domain. Oguz et al. [124] compute the variance of pixels in regions near strong edges to characterize ringing. Visible ringing regions are detected in smooth regions near edges in [125], which also incorporates luminance masking. To increase the accuracy of edge localization, [126,127] apply a bilateral filter, before computing a ringing annoyance score which is a nonlinear function of the local variance of ringing artifacts.

Blockiness, blurriness, and ringing features have also been combined with other features, including bitstream features [128,129] and edge gradient and luminance masking features [130].

Sheikh et al. [131] characterize artifacts for JPEG-2000 by exploring deviations from a signal model. Using a recent model of wavelet coefficients for "natural scenes" [132], they calibrate deviations of this model in the presence of JPEG-2000 compression against human quality ratings.

6.2.2. *Features for other spatial artifacts*

A variety of other spatial artifacts have been considered. Chang et al. [19] design a detector for *rainbow artifacts* using features that extract information about both high-frequency luminance and chrominance components. A method to identify aliasing energy in v_{test} that contributes to visible *jagginess* $d_{visible}$ has recently been presented [133] for integer downsampling. This method is only capable of identifying aliasing energy near strong directional edges that is not masked by the edge. *Color* features have been defined by [134,52,135], including others.

6.2.3. *Temporal features*

The simplest approach to quality estimation for video is to average the estimated quality of individual video frames. However, this approach ignores many known properties of temporal vision, as discussed in Section 5.1.3. More accurate QEs consider both additional temporal features and nonlinear temporal pooling.

The temporal consistency (or its opposite, temporal variability) of luminance levels measures the impact of *flicker* and *blackout* artifacts. While flicker-removal algorithms [136,137,20] use sophisticated models that include motion, their performance is often evaluated using only intensity mean and variance. A no-reference flicker-score was proposed in [138] with the goal of reducing flicker in H.264 encoded videos.

Frame freezes can be identified using several methods that differ based on the type of input information available. If the bitstream is available, frame freezes can be detected using picture time-stamps of received frames, and if only the pixels v_{test} are available, frame freezes may be detected using inter-frame correlation [139]. The duration and regularity of the frame freezes, as well as the intensity of the fluidity break, are incorporated in NR QE for frame freezes [139–142]. A dropping severity indicator, a scene boundary detector, and a motion activity estimator are combined in [142] to design a NR QE that accounts for the perceptual impact of local quality fluctuations.

Strategies for NR QE given *packet loss* depend heavily on the type of available inputs. Those NR QE that rely solely on bitstream *parameters* [143,144,3] are most widely applicable, but they rely on the strong assumption that all packet losses have equivalent perceptual impact. Packet loss rate (PLR) [143,144], quantizer step size [143] and bit-rate and frame-rate [144] have all been incorporated.

Those NR QE that can parse the video bitstream using a variable-length decoder (but do not use v_{test}) can obtain precise information about the location, spatial extent, and temporal extent of the packet loss artifacts [93], but must estimate the strength of the resulting error. A NR model of visibility of packet losses [63] also considers motion predictability, spatial motion smoothness, and the estimated MSE due to packet loss [93]. These methods must rely on assumptions about the error concealment strategy of the decoder.

Finally, those NR QE that rely solely on the video pixels v_{test} no longer rely on assumptions about error concealment, but face the challenge of estimating the location and extent of the artifact. If a loss affects an entire frame, it can be detected as described above for frame freezes. Approaches that search for edges along macroblock boundaries [145,146] assume the error does not affect the entire frame and may fail to identify artifacts caused by error propagation. The additional features of vertical gradients and edges in horizontal gradients are extracted in [147] to detect artifacts of spatial error concealment.

Hybrid methods that use both the pixels v_{test} and its bitstream avoid many of the challenges of using only one input, although they require both inputs to be available and additional processing [3]. The effectiveness of error concealment is evaluated in [148] which identifies corrupted macroblocks using bitstream information and combines bitstream-based motion information with and pixel-based vertical and horizontal luminance discontinuities. Pixel-based features are extracted from bitstream segments that have incorrect checksums in [149] to detect bit-error artifacts.

7. Evaluation

Thorough evaluation of a QE can provide insight into potential improvements, identification of specific failure cases, and eventually more robust performance. In contrast, inadequate evaluation can lead to false performance claims and inevitable QE failure. Any QE should be provided with a complete discussion of evaluation, as described below.

7.1. Assumption and operation verification

Assumptions made at each step of the QE design should be validated or refuted. If a measurement evaluates violations of a signal model, it should be tested on a wide variety of undegraded inputs v_{ref}. If a measurement evaluates conformance with an artifact model, it should be tested to verify not only that it does not unwittingly measure other artifacts [10], and also that it does not detect artifacts inside undegraded inputs. Monotonicity of the individual measurements should be verified. For example, blockiness should increase as quantization increases. The pooling step should be tested to verify correct operation across multiple artifacts. NR QEs which rely on restricted inputs (for example, only v_{test} or only bitstream parameters) should be tested to understand how the input constraints limit performance.

Operation should be evaluated using synthetic inputs (or so-called "toy examples"). Such inputs allow a QE to be stressed in carefully controlled directions. Artifacts can be added, amplified, or spatially and temporally distributed, for example. Inputs which either violate or exactly match the statistical signal models assumed in design can identify whether the QE adequately estimates desired statistical quantities.

7.2. Classical numerical measures

Evaluating the performance of QEs nearly always begins with a quantification of the differences between $Q_{subj}(\mathcal{V})$ and $Q_{obj}(\mathcal{V})$. Obviously, the ground truth data used to *tune and/or train* the QE should not be part of the *test set*. Pearson linear correlation, outlier ratio, and RMS error quantify performance based on how well the QE predicts *individual* subjective quality scores on an absolute scale. Spearman rank-order correlation quantifies how well the QE maintains the relative ranking of scores. These four parameters are the most commonly used quantities (see [77] for a discussion of these quantities, along with several others).

While computation of these quantities over the entire test set provides performance information, evaluation should not stop with these simple computations. Computation of these quantities over meaningful subsets of the test set should also be performed. Such subsets can include classification based upon presence or absence of specific artifacts, sources with more or less observer variability in scores, observers, or any other subset which is reasonable for the particular QE.

Specific examination of outliers is important, as it can provide identification of particular failures within a QE or evidence that a QE will fail on some percentage of unforeseeable cases.

7.3. Resolving power and classification errors

A QE's accuracy regarding differences between pairs of scores in a test set $Q_{subj}(v_1)$ and $Q_{subj}(v_2)$ can be further quantified using *resolving power* and *classification errors*. Brill et al. define the *resolving power* of a QE [150], which computes a confidence in the difference between QE scores $\Delta Q_{obj} = Q_{obj}(v_1) - Q_{obj}(v_2)$ and facilitates an understanding of whether a difference of a given size is meaningful. Resolving power is dependent on the subjective data; a QE can have different resolving powers for different data sets.

Classification errors occur when differences in subjective scores $\Delta Q_{subj} = Q_{subj}(v_1) - Q_{subj}(v_2)$ and QE outputs ΔQ_{obj} for two different sources disagree, in one of three different ways [150,151]:

- *false ties* occur when $|\Delta Q_{subj}| > \gamma$ but $|\Delta Q_{obj}| < \gamma$;
- *false differences* occur when $|\Delta Q_{subj}| < \gamma$ but $|\Delta Q_{obj}| > \gamma$;
- *false ranking* occurs when $Q_{subj}(v_1) > Q_{subj}(v_2)$ but $Q_{obj}(v_1) < Q_{obj}(v_2)$.

The threshold γ may depend on the application (for example, in the stopping criterion for algorithm optimization), but it can be related to the minimum desired quality difference, which is often the JND. (Here, we have assumed that $Q_{obj}(\cdot)$ and $Q_{subj}(\cdot)$ have been normalized to exist on the same scale.)

7.4. Application-specific evaluation

Lastly, the QE should be evaluated in the application for which it was designed. If a QE is designed for algorithm optimization, it should be placed in the loop of an actual algorithm to verify that the algorithm's outputs demonstrate superior quality to those generated without the QE in the loop. Such verification requires a subjective experiment. If a QE is designed for troubleshooting, it should be tested with a multi-component system in which various components are forced to fail. QE designed for other applications should be tested by verifying that Eqs. (4)–(8) are correct when $Q_{obj}(\cdot)$ is substituted for $Q_{subj}(\cdot)$.

8. Concluding thoughts

We have introduced the reader to a variety of applications and a broad range of artifacts. We have described a three-stage framework for NR quality estimation that provides not only the opportunity for including a target application appropriately, but also substantial opportunity for incorporating characteristics of humans as viewers at multiple levels. We have provided a generous survey of approaches, primarily to the *measurement* stage of the framework, and have also enumerated a variety of performance metrics for evaluation of any proposed quality estimator.

An adoption of this framework by the community would facilitate collaborative effort toward effective solutions for the very challenging problem of NR quality estimation. We believe that "the whole is greater than a sum of the parts" and that through joint efforts, substantial progress can be made toward effective no-reference quality estimation.

References

[1] M.P. Eckert, A.P. Bradley, Perceptual quality metrics applied to still image compression, Signal Process. 70 (1998) 177–200.

[2] S. Winkler, Issues in vision modeling for perceptual video quality assessment, Signal Process. 78 (2) (1999) 231–252.

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

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

[5] A.E. Savakis, S.P. Etz, A.C. Loui, Evaluation of image appeal in consumer photography, SPIE, vol. 3959, , 2000, pp. 111–120.

[6] R. Datta, J. Li, J.Z. Wang, Algorithmic inferencing of aesthetics and emotion in natural images: an exposition, in: Proceedings of ICIP , 2008, pp. 105–108.

[7] C. Li, T. Chen, Aesthetic visual quality assessment of paintings, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 236–252.

[8] E. Fedorovskaya, C. Neustaedter, W. Hao, Image harmony for consumer images, in: Proceedings of ICIP, , 2008, pp. 121–124.

[9] R. Samadani, T. Mauer, D. Berfanger, J. Clark, B. Bausk, Representative image thumbnails: automatic and manual, SPIE, vol. 6806, 2008.

[10] M.C.Q. Farias, S.K. Mitra, A methodology for designing no-reference video quality metrics, in: VPQM, 2009.

[11] M. Yuen, H.R. Wu, A survey of hybrid MC/DPCM/DCT video coding distortions, Signal Process. 70 (1998) 247–278.

[12] A. Punchihewa, D.G. Bailey, Artefacts in image and video systems: classification and mitigation, in: Proceedings of Image and Vision Computing, New Zealand, 2002, pp. 197–202.

[13] ANSI T1.801.02-1996, Digital transport of video teleconferencing/video telephony signals—performance terms, definitions, and examples, Technical Report, American National Standard for Telecommunication, 1996.

[14] E. Vicario, I. Heynderickx, G. Ferretti, P. Carrai, Design of a tool to benchmark scaling algorithms on LCD monitors, in: SID Digest of Technical Papers, vol. 33, May 2002, pp. 704–707.

[15] A. Leontaris, P.C. Cosman, A.R. Reibman, Quality evaluation of motion-compensated edge artifacts in compressed video, IEEE Trans. Image Process. 16 (4) (2007) 943–956.

[16] E. Akyol, A.M. Tekalp, M.R. Civanlar, Content-aware scalability-type selection for rate adaptation of scalable video, EURASIP J. Appl. Signal Process. 2007 (1) (2007), pp. 214–214.

[17] I. Chong, H. Cheong, A. Ortega, New quality metric for multimedia compression using faulty hardware, in: International Workshop on Video Processing and Quality Metrics, 2006.

[18] D. Nowroth, I. Polian, B. Becker, A study of cognitive resilience in a JPEG compressor, in: IEEE International Conference on Dependable Systems and Networks, , June 2008, pp. 32–41.

[19] L. Chang, Y.-P. Tan, H.-C. Chua, Detection and removal of rainbow effects artifacts, in: IEEE International Conference on Image Processing, , 2007, pp. I-297–I-300.

[20] A.C. Kokaram, On missing data treatment for degraded video and film archives: a survey and a new Bayesian approach, IEEE Trans. Image Process. 13 (3) (2004) 397–415.

[21] C.J. van den Branden Lambrecht, Color moving pictures quality metric, in: Proceedings of ICIP, , 1996, pp. 885–888.

[22] J. Lubin, M.H. Brill, A. De Vries, O. Finard, Method and apparatus for assessing the visibility of differences between two image sequences, US Patent 5,974,159, October 1999.

[23] A.B. Watson, Toward a perceptual video quality metric, Human Vision and Electronic Imaging, vol. 3299, , 1998, pp. 139–147.

[24] M. Masry, S.S. Hemami, Y. Sermadevi, A scalable wavelet-based video distortion metric and applications, IEEE Trans. Circuits Syst. Video Technol. 16 (2) (2006) 260–273.

[25] A.A. Michelson, Studies in Optics, University of Chicago Press, 1927.

[26] R.F. Hess, A. Bradley, L. Piotrowski, Contrast-coding in amblyopia. I. Differences in the neural basis of human amblyopia, Proc. R. Soc. London Ser. B 217 (1983) 309–330.

[27] E. Peli, Contrast in complex images, J. Opt. Soc. Am. A 7 (1990) 2032–2040.

[28] B. Moulden, F.A.A. Kingdom, L.F. Gatley, The standard deviation of luminance as a metric for contrast in random-dot images, Perception 19 (1990) 79–101.

[29] N. Graham, Visual Pattern Analyzers, Oxford University Press, New York, 1989.

[30] G.E. Legge, J.M. Foley, Contrast masking in human vision, J. Opt. Soc. Am. 70 (1980) 1458–1470.

[31] D.J. Heeger, Normalization of cell responses in cat striate cortex, Visual Neurosci. 9 (1992) 181–197.

[32] J.M. Foley, Human luminance pattern mechanisms: masking experiments require a new model, J. Opt. Soc. Am. A 11 (1994) 1710–1719.

[33] J.M. Foley, C.C. Chen, Pattern detection in the presence of maskers that differ in spatial phase and temporal offset: threshold measurements and a model, Vision Res. 39 (1999) 3855–3872.

[34] A.B. Watson, J.A. Solomon, A model of visual contrast gain control and pattern masking, J. Opt. Soc. Am. A 14 (1997) 2379–2391.

[35] A.B. Watson, M. Taylor, R. Borthwick, Image quality and entropy masking, in: Human Vision, Visual Processing, and Digital Display VIII, Proceedings of SPIE, vol. 3016, 1997, pp. 2–12.

[36] S.S. Hemami, D.M. Chandler, B.G. Chern, J.A. Moses, Suprathreshold visual psychophysics and structure-based visual masking, in: Visual Communications and Image Processing, 2006.

[37] J.G. Robson, Spatial and temporal contrast-sensitivity functions of the visual system, J. Opt. Soc. Am. 56 (8) (1966) 1441–1442.

[38] A. Watanabe, T. Mori, S. Nagata, K. Hiwatashi, Spatial and temporal contrast-sensitivity functions of the visual system, Vision Res. 8 (9) (1968) 1245–1263.

[39] D.H. Kelly, Motion and vision. II. Stabilized spatio-temporal threshold surface, J. Opt. Soc. Am. 69 (10) (1979) 1340–1349.

[40] R.F. Hess, R.J. Snowden, Temporal properties of human visual filters: number shapes and spatial covariation, Vision Res. 32 (1) (1992) 47–59.

[41] R.E. Fredericksen, R.F. Hess, Temporal detection in human vision: dependence on stimulus energy, J. Opt. Soc. Am. A 14 (10) (1997) 2557–2569.

[42] R.E. Fredericksen, R.F. Hess, Estimating multiple temporal mechanisms in human vision, Vision Res. 38 (7) (1998) 1023–1040.

[43] W.G. Owen, Spatio-temporal integration in the human peripheral retina, Vision Res. 12 (1972) 1011–1026.

[44] D. Kahneman, Time intensity reciprocity under various conditions of adaptation and backward masking, J. Exp. Psychol. 71 (1966) 543–549.

[45] E. Hering, Zur lehre vom Lichtsinne, Carl Gerolds and Sohn, 1878.

[46] D. Jameson, L.M. Hurvich, Some quantitative aspects of an opponent-colors theory. I. chromatic responses and spectral saturation, J. Opt. Soc. Am. 45 (7) (1955) 546–552.

[47] L.M. Hurvich, D. Jameson, An opponent-process theory of color vision, Psychol. Rev. 64 (10) (1957) 384–404.

[48] G.J.C. van der Horst, M.A. Bouman, Spatiotemporal chromaticity discrimination, J. Opt. Soc. Am. 59 (11) (1969) 1482–1488.

[49] E.M. Granger, J.C. Heurtley, Spatiotemporal chromaticity discrimination, J. Opt. Soc. Am. 63 (9) (1973) 1173–1174.

[50] D.H. Kelly, Luminous and chromatic flickering patterns have opposite effects, Science 188 (4186) (1975) 371–372.

[51] R.L. DeValois, K.K. DeValois, Spatial Vision, Oxford University Press, 1990.

[52] C.C. Koh, J.M. Foley, S.K. Mitra, Color preference and perceived color naturalness of digital videos, SPIE, vol. 6057, , 2006, pp. 257–267.

[53] C.C. Koh, J.M. Foley, S.K. Mitra, Color preference, color naturalness, and annoyance of compressed and color scaled digital videos, SPIE, vol. 6492, 2007.

[54] B. Zhang, J.P. Allebach, Z. Pizlo, An investigation of perceived sharpness and sharpness metrics, Proceedings of the SPIE, vol. 5668, , 2005, pp. 98–110.

[55] B. Keelan, Handbook of Image Quality: Characterization and Prediction, Marcel Dekker, Inc, 2002.

[56] International Telecommunication Union, ITU-T recommendation P.930: principles of a reference impairment system for video, 1996.

[57] M.C.Q. Farias, J.M. Foley, S.K. Mitra, Detectability and annoyance of synthetic blocky, blurry, noisy, and ringing artifacts, IEEE Trans. Signal Process. 55 (6) (2007) 2954–2964.

[58] R.R. Pastrana-Vidal, J.-C. Gicquel, C. Colomes, H. Cherifi, Sporadic frame dropping impact on quality perception, SPIE, vol. 5292, 2004.

[59] R.R. Pastrana-Vidal, J.-C. Gicquel, C. Colomes, H. Cherifi, Temporal masking effect on dropped frames at video scene cuts, SPIE, vol. 5292, 2004.

[60] Q. Huynh-Thu, M. Ghanbari, Temporal aspects of perceived quality in mobile video broadcasting, IEEE Trans. Broadcast. 54 (3) (2008) 641–651.

[61] M.S. Moore, J.M. Foley, S.K. Mitra, Defect visibility and content importance: effects on perceived impairment, Signal Process. Image Commun. 19 (2004) 185–203.

[62] F. Boulos, B. Parrein, P. Le Callet, D.S. Hands, Perceptual effects of packet loss on H.264/AVC encoded videos, in: International Workshop on Video Processing and Quality Metrics, 2009.

[63] S. Kanumuri, P.C. Cosman, A.R. Reibman, V.A. Vaishampayan, Modeling packet-loss visibility in MPEG-2 video, IEEE Trans. Multimedia 8 (2) (2006) 341–355.

[64] S. Kanumuri, S.G. Subramanian, P.C. Cosman, A.R. Reibman, Predicting H.264 packet loss visibility using a generalized linear model, in: IEEE International Conference on Image Processing, 8–11 October 2006, pp. 2245–2248.

[65] G. Yadavalli, M. Masry, S.S. Hemami, Frame rate preferences in low bit rate video, in: IEEE International Conference on Image Processing, vol. 1, 14–17 September 2003, pp. I-441–I-444.

[66] C.J. Hughes, M. Ghanbari, D.E. Pearson, V. Seferidis, J. Xiong, Modeling and subjective assessment of cell discard in atm video, IEEE Trans. Image Process. 2 (1993) 212–222.

[67] S. Winkler, R. Campos, Video quality evaluation for Internet streaming applications, in: SPIE, , 2003, pp. 104–115.

[68] J.B. Kruskal, M. Wish, Multidimensional Scaling, Sage, 1986.

[69] J.-B. Martens, Multidimensional modeling of image quality, Proc. IEEE 90 (1) (2002) 133–153.

[70] J.S. Goodman, D.E. Pearson, Multidimensional scaling of multiply-impaired television pictures, IEEE Trans. Syst. Man Cybern. SMC-9 (6) (1979) 353–356.

[71] J. Allnatt, Transmitted Picture Assessment, J. Wiley and Sons, New York, NY, 1983.

[72] V. Kayargadde, J.-B. Martens, Perceptual characterization of images degraded by blur and noise: experiments, J. Opt. Soc. Am. A 13 (6) (1996) 1166–1177.

[73] D.M. Chandler, K.H. Lim, S.S. Hemami, Effects of spatial correlations and global precedence on the visual fidelity of distorted images, in: Human Vision and Electronic Imaging, 2006.

[74] F. Kozamernik, P. Sunna, E. Wyckens, D.I. Pettersen, Subjective quality of internet video codecs phase II evaluations using SAMVIQ, EBU Technical Review, January 2005.

[75] B. Keelan, H. Urabe, ISO 20462, a psychological image quality measurement standard, SPIE, vol. 5294, , 2004, pp. 181–189.

[76] Video Quality Experts Group (VQEG), Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, March 2000.

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

[78] Video Quality Experts Group (VQEG), Final report from the video quality experts group on the validation of objective models of multimedia quality assessment, phase I, September 2008.

[79] CCIR, Recommendation 500-3: method for the subjective assessment of the quality of television pictures, Recommendations and Reports of the CCIR, International Telecommunication Union, Geneva, 1996.

[80] ATIS-0800025, Test plan for evaluation of quality models for IPTV services, October 2009.

[81] P. Corriveau, Video quality testing, in: H.R. Wu, K.R. Rao (Eds.), Digital Video Image Quality and Perceptual Coding, CRC Press, 2006, pp. 125–154 (Chapter 4).

[82] C. Lee, H. Choi, E. Lee, S. Lee, J. Choe, Comparison of various subjective video quality assessment methods, SPIE, vol. 6059, 2006.

[83] M. Knee, The picture appraisal rating (PAR)—a single-ended picture quality measure for MPEG-2, in: Proceedings of International Broadcasting Convention, September 2000, pp. 95–100.

[84] D.S. Turaga, Y. Chen, J. Caviedes, No reference PSNR estimation for compressed pictures, Signal Process. Image Commun. 19 (2004) 173–184.

[85] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, E. Nakasu, A method of estimating coding PSNR using quantized DCT coefficients, IEEE Trans. Circuits Syst. Video Technol. 16 (2) (2006) 251–259.

[86] A. Ichigaya, Y. Nishida, E. Nakasu, Nonreference method for estimating PSNR of MPEG-2 coded video by using DCT coefficients and picture energy, IEEE Trans. Circuits Syst. Video Technol. 18 (6) (2008) 817–826.

[87] T. Brandão, M.P. Queluz, Blind PSNR estimation of video sequences using quantized DCT coefficient data, in: Picture Coding Symposium, 2007.

[88] T. Brandão, M.P. Queluz, No-reference image quality assessment based on DCT domain statistics, Signal Process. 88 (2008) 822–833.

[89] A. Eden, No-reference estimation of the coding PSNR for H.264-coded sequences, IEEE Trans. Consum. Electron. 53 (2) (2007) 667–674.

[90] J. Choe, C. Lee, Estimation of the peak signal-to-noise ratio for compressed video based on generalized Gaussian modeling, Opt. Eng. 46 (10) (2007).

[91] T. Brandão, M.P. Queluz, No-reference PSNR estimation algorithm for H.264 encoded video sequences, in: EUSIPCO'08, 2008.

[92] S.-Y. Shim, J.-H. Moon, J.-K. Han, PSNR estimation scheme using coefficient distribution of frequency domain in H.264 decoder, Electron. Lett. 44 (2) (2008).

[93] A.R. Reibman, V.A. Vaishampayan, Y. Sermadevi, Quality monitoring of video over a packet network, IEEE Trans. Multimedia 6 (2) (2004) 327–334.

[94] A.R. Reibman, V. Vaishampayan, Low complexity quality monitoring of MPEG-2 video in a network, in: IEEE International Conference on Image Processing, vol. 3, 14–17 September 2003, pp. III-261–III-264.

[95] M. Naccari, M. Tagliasacchi, F. Pereira, S. Tubaro, No-reference modeling of the channel induced distortion at the decoder for H.264/AVC video coding, in: IEEE ICIP, 2008.

[96] K. Nishikawa, K. Munadi, H. Kiya, No-reference PSNR estimation for quality monitoring of motion JPEG2000 video over lossy packet networks, IEEE Trans. Multimedia 10 (4) (2008) 637–645.

[97] S.I. Olsen, Estimation of noise in images: an evaluation, in: CVGIP: Graphical Models Image Processing, vol. 55 (4), July 1993, pp. 319–323.

[98] X. Li, Blind image quality assessment, in: IEEE ICIP02, 2002.

[99] A. Amer, E. Dubois, Fast and reliable structure-oriented video noise estimation, IEEE Trans. Circuits Syst. Video Technol. 15 (1) (2005).

[100] M. Ghazal, A. Amer, A. Ghrayeb, A real-time technique for spatio-temporal video noise estimation, IEEE Trans. Circuits Syst. Video Technol. 17 (12) (2007) 1690–1699.

[101] V. Kayargadde, J.-B. Martens, An objective measure for perceived noise, Signal Process. 49 (3) (1996) 187–206.

[102] K. Rank, M. Lendl, R. Unbehauen, Estimation of image noise variance, IEE Visual Image Signal Process. 164 (1999) 80–84.

[103] A.B. Watson, DCT quantization matrices optimized for individual images, in: SPIE Human Vision, Visual Processing, and Digital Display IV, 1993.

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

[105] H.R. Wu, M. Yuen, A generalized block-edge impairment metric for video coding, IEEE Signal Process. Lett. 4 (11) (1997) 317–320.

[106] S. Suthaharan, A perceptually significant block-edge impairment metric for digital video coding, in: Proceedings of International Conference on Acoustics, Speech, and Signal Processing, , 2003, pp. III-681–III-684.

[107] R.V. Babu, S. Suresh, A. Perkis, No-reference JPEG-image quality assessment using GAP-RBF, Signal Process. 87 (2007) 1493–1503.

[108] Z. Wang, A.C. Bovik, B.L. Evans, Blind measurement of blocking artifacts in images, in: IEEE International Conference on Image Processing, 2000.

[109] S. Liu, A.C. Bovik, Efficient DCT-domain blind measurement and reduction of blocking artifacts, IEEE Trans. Circuits Syst. Video Technol. 12 (12) (2002) 1139–1149.

[110] G. Zhai, W. Zhang, X. Yang, W. Lin, Y. Xu, No-reference noticeable blockiness estimation in images, Signal Process. Image Commun. 23 (2008) 417–432.

[111] L. Meesters, J.-B. Martens, A single-ended blockiness measure for JPEG-coded images, Signal Process. 82 (2002) 369–387.

[112] R. Muijs, I. Kirenko, A no-reference block artifact measure for adaptive video processing, in: EUSIPCO ’05, 2005.

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

[114] R.M. Rangayyan, S.G. Elkadiki, Algorithm for the computation of region-based image edge profile acutance, J. Electron. Imag. 4 (1) (1995) 62–70.

[115] P. Marziliano et al., Perceptual blur and ringing metrics: application to JPEG2000, Sig. Process. Image Commun., February 2004, pp. 163–172.

[116] N.G. Sadaka, L.J. Karam, R. Ferzli, G.P. Abousleman, A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling, in: IEEE International Conference on Image Processing, 12–15 October 2008, pp. 369–372.

[117] E. Ong, W. Lin, Z. Lu, X. Yang, S. Yao, L. Jiang, F. Moschetti, A no-reference quality metric for measuring image blur, in: IEEE International Symposium on Signal Processing and its Applications, 2003, pp. 469–472.

[118] V. Kayargadde, J.-B. Martens, Estimation of edge parameters and imager blur using polynomial transforms, in: CVGIP: Graphical Models Image Process., vol. 56 (6), November 1994, pp. 442–461.

[119] J. Dijk, M. van Grinkel, R.J. van Asselt, L.J. van Vliet, P.W. Verbeek, A new sharpness measure based on Gaussian lines and edges, in: International Conference on Computer Analysis on Images and Patterns (CAIP), vol. 2756, 2003, pp. 149–156.

[120] X. Marichal, W.-Y. Ma, H.-J. Zhang, Blur determination in the compressed domain using DCT information, in: IEEE International Conference on Image Processing, 1999, pp. 386–390.

[121] K.-C. Yang, C.C. Guest, P.K. Das, Perceptual sharpness metric (PSM) for compressed video, in: IEEE International Conference on Multimedia and Expo, 2006.

[122] G. Blanchet, L. Moisan, B. Rougé, Measuring the global phase coherence of an image, in: IEEE International Conference on Image Processing, 2008, pp. 1176–1179.

[123] J. Caviedes, F. Oberti, A new sharpness metric based on local kurtosis edge and energy information, Signal Process. Image Commun. 19 (10) (2004) 147–161.

[124] S.H. Oguz, Y.H. Hu, T.Q. Nguyen, Image coding ringing artifact reduction using morphologicalpost-filtering, in: IEEE International Workshop on Multimedia Signal Processing, , 1998, pp. 628–633.

[125] X. Feng, J.P. Allebach, Measurement of ringing artifacts in JPEG images, SPIE, vol. 6076, 2006.

[126] H. Liu, N. Klomp, I. Heynderickx, Perceptually relevant ringing region detection method, in: EUSIPCO ’08, 2008.

[127] H. Liu, N. Klomp, I. Heynderickx, A no-reference metric for perceived ringing, in: International Workshop on Video Processing and Quality Metrics, 2009.

[128] A.G. Davis, D. Bayart, D.S. Hands, Hybrid no-reference video quality prediction, in: IEEE International Symposium on Broadband Multimedia Systems, 2009.

[129] D. Hands, D. Bayart, A. Davis, A. Bourret, No reference perceptual quality metrics: approaches and limitations, in: Human Vision and Electronic Imaging XIV, 2009, p. 72400Y.

[130] U. Engelke, H.-J. Zepernick, Pareto optimal weighting of structural impairments for wireless imaging quality assessment, in: IEEE International Conference on Image Processing, 2008, pp. 373–376.

[131] H.R. Sheikh, A.C. Bovik, L. Cormak, No-reference quality assessment using natural scene statistics: JPEG 2000, IEEE Trans. Image Process. 14 (11) (2005) 1918–1927.

[132] R.W. Buccigrossi, E.P. Simoncelli, Image compression via joint statistical characterization in the wavelet domain, IEEE Trans. Image Process. 8 (12) (1999) 1688–1701.

[133] A.R. Reibman, S. Suthaharan, A no-reference spatial aliasing measure for digital image resizing, in: IEEE International Conference on Image Processing, 12–15, October 2008, pp. 1184–1187.

[134] S.E. Susstrunk, S. Winkler, Color image quality on the Internet, SPIE, vol. 5304, 2004.

[135] T.J. Janssen, F.J. Blommaert, Predicting the usefulness and naturalness of color reproductions, J. Imaging Sci. Technol. 44 (2) (2000) 93–104.

[136] P.M.B. van Roosmalen, R.L. Lagendijk, J. Biemond, Correction of intensity flicker in old film sequences, IEEE Trans. Circuits Syst. Video Technol. 9 (7) (1999) 1013–1019.

[137] T. Vlachos, Flicker correction for archived film sequences using a nonlinear model, IEEE Trans. Circuits Syst. Video Technol. 14 (4) (2004) 508–516.

[138] Y. Kuszpet, D. Kletsel, Y. Moshe, A. Levy, Post-processing for flicker reduction in H.264/AVC, in: Picture Coding Symposium, 2007.

[139] M. Montenovo, A. Perot, M. Carli, P. Cicchetti, A. Neri, Objective quality evaluation of video services, in: International Workshop on Video Processing and Quality Metrics, 2006.

[140] R.R. Pastrana-Vidal, J.-C. Gicquel, Automatic quality assessment of video fluidity impairments using a no-reference metric, in: International Workshop on Video Processing and Quality Metrics, 2006.

[141] K. Watanabe, J. Okamoto, T. Kurita, Objective video quality assessment method for evaluating effects of freeze distortion in arbitrary video scenes, SPIE Electronic Imaging, vol. 6494, 2007.

[142] K.-C. Yang, C.C. Guest, K. El-Maleh, P.K. Das, Perceptual temporal quality metric for compressed video, IEEE Trans. Multimedia 9 (7) (2007) 1528–1535.

[143] O. Verscheure, P. Frossard, M. Hamdi, User-oriented QoS analysis in MPEG-2 video delivery, Real-Time Imaging 5 (1999) 305–314.

[144] S. Mohamed, G. Rubino, A study of real-time packet video quality using random neural networks, IEEE Trans. Circuits Syst. Video Technol. 12 (12) (2002) 1071–1083.

[145] R.V. Babu, A.S. Bopardikar, A. Perkis, O.I. Hillestad, No-reference metrics for video streaming applications, in: International Workshop on Packet Video, 2004.

[146] H. Rui, C. Li, S. Qiu, Evaluation of packet loss impairment on streaming video, Journal of Zhejiang University SCIENCE B 7 (April) (2006).

[147] D. Shabtay, N. Raviv, Y. Moshe, Video packet loss concealment detection based on image content, in: EUSIPCO ’08, 2008.

[148] T. Yamada, M. Yoshihiro, M. Serizawa, No-reference video quality estimation based on error-concealment effectiveness, in: Packet Video, 2007, pp. 288–293.

[149] R.A. Farrugia, C.J. Debono, A robust error detection mechanism for H.264/AVC coded video sequences based on support vector machines, IEEE Trans. Circuits Syst. Video Technol. 18 (12) (2008) 1766–1770.

[150] M.H. Brill, J. Lubin, P. Costa, S. Wolf, J. Pearson, Accuracy and cross-calibration of video-quality metrics: new methods from ATIS/T1A1, Signal Process. Image Commun. 19 (2004) 101–107.

[151] ATIS Technical Report T1.TR.77-2002, Data and sample program code to be used with the method specified in T1.TR.72-2001 for the calculation of resolving power of the video quality metrics in T1.TR.74-2001 and T1.TR.75-2001, January 2002.