J. Vis. Commun. Image R. 22 (2011) 297–312

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Perceptual visual quality metrics: A survey

Weisi Lin [a,*], C.-C. Jay Kuo [b]

[a] School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore
[b] Ming Hsieh Department of Electrical Engineering and the Signal and Image Processing Institute, University of Southern California, Los Angeles, CA 90089-2564, USA

ARTICLE INFO

Article history:
Received 17 November 2009
Accepted 21 January 2011
Available online 1 February 2011

Keywords:
Human visual system (HVS)
Vision-based model
Signal-driven model
Signal decomposition
Just-noticeable distortion
Visual attention
Common feature and artifact detection
Full reference
No reference
Reduced reference

ABSTRACT

Visual quality evaluation has numerous uses in practice, and also plays a central role in shaping many visual processing algorithms and systems, as well as their implementation, optimization and testing. In this paper, we give a systematic, comprehensive and up-to-date review of perceptual visual quality metrics (PVQMs) to predict picture quality according to human perception. Several frequently used computational modules (building blocks of PVQMs) are discussed. These include signal decomposition, just-noticeable distortion, visual attention, and common feature and artifact detection. Afterwards, different types of existing PVQMs are presented, and further discussion is given toward feature pooling, viewing condition, computer-generated signal and visual attention. Six often-used image metrics (namely SSIM, VSNR, IFC, VIF, MSVD and PSNR) are also compared with seven public image databases (totally 3832 test images). We highlight the most significant research work for each topic and provide the links to the extensive relevant literature.

1. Introduction

Quality evaluation for digital visual signals is one of the basic and challenging problems in the field of image and video processing as well as many practical situations, such as process evaluation, implementation, optimization (e.g., video encoding), testing and monitoring (e.g., in transmission and manufacturing sites). In addition, how to evaluate picture quality plays a central role in shaping most (if not all) visual processing algorithms and systems [50,114,124]. Examples of technological dependence upon visual quality evaluation include: signal acquisition, synthesis, enhancement, watermarking, compression, transmission, storage, retrieval, reconstruction, authentication, and presentation (e.g., display and printing).

Objective quality evaluation for images and video can be classified into two board types: signal fidelity measures, and perceptual visual quality metrics (PVQMs).

The signal fidelity measures refer to the traditional MAE (mean absolute error), MSE (mean square error), SNR (signal-to-noise ratio), PSNR (peak SNR), or one of their relatives [41]. Although they are simple, well defined, with clear physical meanings and widely accepted, they can be a poor predictor of perceived visual quality, especially when the noise is not additive [71,84]. Some metrics have been used to estimate delivered picture quality after transmission based on network parameters [108,138,183], such as throughput, jitter, delay, bit error and packet loss rates. However, the same network parameters may result in different degradation of visual content, and therefore different perceived quality. Quality determined by consumers' perception and satisfaction is much more complex than the statistics that a typical network management system can provide. It has been well acknowledged that a signal fidelity measure does not align well with human visual perception of natural images or computer generated graphics [41,52,97,149,161].

Since the human visual system (HVS) is the ultimate receiver and appreciator for the majority of processed images, video and graphics, it would be more logical, economical and user-oriented to develop a perceptual quality metric in system design and optimization. Naturally, perceptual visual quality (or distortion) can be evaluated by subjective viewing tests with appropriate standard procedures [65]. This is however time consuming, laborious and expensive, since the resultant mean opinion score (MOS) needs to be obtained by many observers through repeated viewing sessions. Moreover, incorporation of subjective viewing tests is not feasible for on-line visual signal manipulations (such as encoding, transmission, relaying, etc.). Even in situations where human examiners are allowed (e.g., visual inspection in a factory environment) and the manpower cost is not a problem, the assessment results still depend upon viewers' physical conditions, emotional states, personal experience, and the context of preceding display. Hence, it is necessary to build computational models to predict the evaluation of an average observer. In other words, objective means are sought to approximate human perception results (e.g.,

* Corresponding author.
E-mail addresses: wslin@ntu.edu.sg (W. Lin), cckuo@sipi.usc.edu (C.-C. Jay Kuo).

1047-3203/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvcir.2011.01.005

MOS, when the number of subjects is sufficiently large). In comparison with the subjective viewing tests, objective metrics are advantageous in repeatability due to the nature of objective measurement.

Although physical variations in terms of MSE, SNR, PSNR, etc. reflect picture quality change, these traditional signal fidelity metrics fail to predict the HVS perception because of the following problems:

(1) Not every change in an image is noticeable;
(2) Not every pixel/region in an image receives the same attention level;
(3) Not every change leads to distortion (otherwise, many edge sharpening and post-processing algorithms would have not been developed);
(4) Not every change yields a same extent of perceptual effect with a same magnitude of change (due to spatial/temporal/chrominance masking).

A significant amount of research efforts has been made toward HVS-based picture quality evaluation during the past decade [26,27,51,70,106,118,123,156,157,160,183,188] so as to tackle the abovementioned four problems of traditional measures.

2. The problem

2.1. Nature of the problem

Visual quality assessment can be of the first party (the photographer or image maker), the second party (the subject of an image) and the third party (neither the photographer nor the subject) [72]. The concern in this survey is the perception of third-party observers, since this represents the most general and meaningful situation in modeling and applications.

PVQMs refer to the objective models for predicting subjective visual quality scores (i.e., the MOS). In this paper, we will focus on surveying the PVQMs developed so far that carry out direct evaluation of the actual picture under consideration, rather than some predefined signal patterns that go through the same processing [66]. This is because picture quality is a function of visual contents, so the change of predefined test signals through a system is not necessarily a reliable source of visual quality measurement for actual signals; and in addition, the predefined visual signal adds to the overheads of transmission/storage.

In spite of the recent progress in related fields, objective evaluation of picture quality in line with human perception is still a long and difficult odyssey [38,123,156,157,163,183] due to the complex, multi-disciplinary nature of the problem (related to physiology, psychology, vision research and computer science), the limited understanding of the HVS mechanism, and the diversified scope of applications and requirements.

Despite the difficulties, perceptual visual quality evaluation should be less demanding than computer vision in general, since it can be performed without the need of emulating "*the process of discovering from images what is present in the world, and where it is*" (Marr's words on vision [99]), in most meaningful and practical situations for visual quality evaluation. With proper modeling of major underlying physiological and psychological phenomena, it is possible to develop better visual quality metrics to replace non-perceptual criteria widely used nowadays, in various specific practical situations.

2.2. Organization of this paper

Due to the vast scope of this survey, we divide the main body of the survey that follows into two parts for clearer presentation: in Section 3 below, a review will be given on basic computational modules in building various PVQMs; in Section 4, two major categories of PVQMs will be then discussed. The further rationale for such a 2-step organization strategy is as follows.

The basic computational modules include signal decomposition (decomposing an image or video into different color, spatial and temporal channels), detection of common features (like contrast and motion) and artifacts (like blockiness and blurring), just-noticeable distortion (JND) (i.e., the maximum change in visual content that cannot be detected by the majority of viewers), and visual attention (VA) (i.e., the HVS's selectivity to respond to the most attractive activities in the visual field). First, many of these are based upon the related physiological and psychological knowledge. Second, most of them are independent research topics themselves, like JND and VA modelling, and have other applications (image/video coding [10,194], watermarking [187], error resilience [48], computer graphics [136], just to name a few) in addition to PVQMs. Third, these modules can be simple PVQMs themselves in specific situations (e.g., blockness and burring). After the discussion of these basic building modules, we will be able to focus on system-level issues related to the major PVQMs in Section 4.

In Section 5, we will compare six existing image quality metrics (SSIM [167], VSNR [17], IFC [139], VIF [140], MSVD [42], and PSNR) against the subjective viewing data, from seven publicly available databases. These databases are with a wide variety of visual contents and distortion types to enable a meaningful and convincing benchmarking.

Before going to the main body of this paper, let us briefly explain several psychophysical phenomena that have been commonly used in PVQM development. The contrast sensitivity function (CSF) denotes the HVS's sensitivity toward signal contrast with spatial frequencies and temporal motion velocities [73,155], and exhibits a parabola-like curve with the increase of spatial and temporal frequencies, respectively. Luminance adaptation refers to the noticeable luminance contrast as a function of background luminance; for digital images, luminance adaptation takes a parabola-like curve [23,67]. Visual masking is usually the increase of the HVS's contrast threshold for a signal in the presence of another one; it can be divided into intra-channel masking [7] by the signal itself, and inter-channel masking [13,82] by signals with different frequencies and orientations.

For the convenience of the reader, the major abbreviations and notations used in this paper are listed in Tables 1 and 2.

3. Basic computational modules

There have been basically two categories of PVQMs [183]: the vision-based modeling and signal-driven approach. For the first category [30,93,174,178], PVQMs are developed based upon systematical modeling of relevant psychophysical properties and physiological knowledge, including temporal/spatial/color decomposition, CSF, luminance adaptation, and various masking effects. The second category attempts to tackle the problem from the viewpoint of signal extraction and analysis, such as statistical features [185], structural similarity [162], luminance/color distortion [107], and the common visual artifacts (e.g., blockiness and blurring) [100,189]. These metrics look at how pronounced the related features are in image/video to estimate overall quality. This does not necessarily mean that such metrics disregard human vision knowledge, as they often consider psychophysical effects as well (e.g., a JND model), but image content and distortion analysis rather than fundamental vision modeling is the basis for design.

There are metrics making use of both classes. For example, a scheme was proposed in [148] to switch between a model-based scheme and a signal-driven one according to the extent of blocki-

Table 1
Major abbreviations used in this paper.

Abbreviation	Explanation	First appearance
A57	An perceptual image quality database [32]	Section 5.1
CG	Computer graphics	Section 4.3.3
CSF	Contrast sensitivity function	Section 2.2
CSIQ	An perceptual image quality database [80]	Section 5.1
DCT	Discrete cosine transform	Section 3.2.1
DMOS	Differential MOS	Section 5
DWT	Discrete Wavelet Transform	Section 3.3.1
FT	Fourier Transform	Section 3.4
FR	Full-reference	Section 4
HVS	The human visual system	Section 1
IFC	An image quality metric [139]	Section 2.2
IFT	Inverse Fourier Transform	Section 3.4
IVC	An perceptual image quality database [12]	Section 5.1
JND	Just-noticeable distortion	Section 2.2
LIVE	An perceptual image quality database [141]	Section 5.1
MAE	Mean absolute error	Section 1
MAE	Mean opinion score	Section 1
MOS	Mean square error	Section 1
MSVD	An image quality metric [42]	Section 2.2
NR	No-reference	Section 4
PSF	Point spread function	Section 3.2
PSNR	Peak SNR	Section 1
PVQM	Perceptual Visual Quality Metric	Section 1
QoE	Quality of Experience	Section 6
RF	Random field	Section 4.2.1
RMSE	Root-mean-square error	Section 5.2
RR	Reduced-reference	Section 4
SNR	Signal-to-noise ratio	Section 1
SSIM	An image quality metric [167]	Section 2.2
SVD	Singular Value Decomposition	Section 4.2.1
TID	An perceptual image quality database [131]	Section 5.1
VA	visual attention	Section 2.2
VIF	An image quality metric [140]	Section 2.2
VSNR	An image quality metric [17]	Section 2.2
WIQ	An perceptual image quality database [40]	Section 5.1

Table 2
Major notations used in this paper.

Notation	Explanation	First appearance
A	The FT amplitude	Section 3.4
c	A color (or luminance) channel	Section 3.1
C	DCT coefficient	Section 3.3.1
C_P	Pearson linear correlation coefficient	Section 5.2
C_S	Spearman rank order correlation coefficient	Section 5.2
f_s	Spatial frequency	Section 3.1
f_t	Temporal frequency	Section 3.1
F	The intensity, color or orientation map for an image	Section 3.4
$\widehat{F}$	The pixel-by-pixel contrast for F	Section 3.4
F^l	The interpolation of F to the finer scale	Section 3.4
g	The weighted average of gradients	Section 3.3.2
H	Height of an image	Section 3.2.2
(i,j)	DCT subband index	Section 3.3.1
I	an image	Section 3.2.2
JND_D	DCT-based luminance JND	Section 3.3.1
JND_p	Pixel-based JND	Section 3.3.2
M_h	Horizontal blockiness	Section 3.2.2
M_v	Vertical blockiness	Section 3.2.2
n	DCT block index	Section 3.3.1
N	Image block size	Section 3.2.2
r	Orientation	Section 3.1
R	The VA map	Section 3.4
s	A decomposed signal component	Section 3.1
s_p	Perceptual effect of s	Section 4.1
T^l	Luminance adaptation	Section 3.3.2
T^t	Texture masking	Section 3.3.2
T_{s_csf}	The base threshold for DCT-based JND due to the spatial CSF	Section 3.3.1
v	The object velocity perceived by the retina	Section 3.3.1
v_o	The object velocity in the image	Section 3.3.1
v_e	The eye movement velocity	Section 3.3.1
W	width of an image	Section 3.2.2
(x,y)	Pixel location	Section 3.1
α_φ	The elevation parameter for DCT-based JND due to an effect φ	Section 3.3.1
α_{inter}	The elevation parameter for DCT-based JND due to inter-band masking	Section 3.3.1
α_{intra}	The elevation parameter for DCT-based JND due to intra-band masking	Section 3.3.1
α_{lum}	The elevation parameter for DCT-based JND due to luminance adaptation	Section 3.3.1
ΔC	The DCT coefficient difference between the reference and test images	Section 4.2.1
ΔC_p	The perceptual distortion for ΔC	Section 4.2.1
κ^{lt}	A parameter accounting for the overlapping effect between $T^l(x,y)$ and T^t	Section 3.3.2
ϕ	The FT phase	Section 3.4

ness in decoded video, and a model-based metric was applied to blockiness-dominant areas in [197], with the help of a signal-driven measure.

Most vision-based PVQMs use signal decomposition in images. The signal feature extraction and common artifact detection are the core for many signal-driven PVQMs; the perceptual effect of common imaging and compression artifacts far exceeds the extent of their representation in the MSE or PSNR. The JND and VA models have been used either independently or jointly to evaluate the visibility and the perceived extent of visual signal difference. Therefore, all these techniques help to address the four basic problems (as mentioned in the Introduction) to be overcome against the traditional signal fidelity metrics, since they enable the differentiation of various visual signal changes for perceptual quality evaluation purpose.

In this section, we review the existing work on these four topics, namely, image/video decomposition, visual feature and artifact detection, JND modelling, and VA map generation. We intend only to a brief coverage of the first two topics since they are primarily based upon adapting traditional filtering and image processing techniques for the purpose of PVQMs. Based upon the major ideas presented, the reader should be able to find more details with references provided. The emphasis of this section is therefore for the last two topics, with more analysis and discussion. A clear view of all these modules facilitates not only the presentation of this paper but also future research and new system development.

3.1. Image/video decomposition

It is well known [69,137,159] that the HVS has separate processing for achromatic and chromatic signals, different visual pathways for signals with fast and slow motion, and special cells in the visual cortex for distinctive orientations. Psychophysical experiments also demonstrate that visual signals are differentiated with frequency [74,111] and orientation [82,129]. Therefore, decomposition of an image or video into different color, spatial and temporal channels facilitates the evaluation of signal changes for unequal treatment of each signal component to emulate the

HVS response, by enabling a system to address the fourth problem of traditional metrics mentioned in the Introduction. A standard, general process of visual signal decomposition is to derive $s(c,f_t,f_s,r,x,y)$, which represents a decomposed signal component, with color (or luminance) c, temporal frequency f_t, spatial frequency f_s, orientation r and location (x,y), respectively.

For the color representation, the opponent-color (B-W, R-G, and Y-B) space [130,178], which is based on the physiological evidence of the opponent cells in the parvocellular pathway, and the CIELAB space [202], which is more perceptually uniform, can be used for visual quality evaluation. The YC_bC_r space is more convenient if the compressed signal is dealt with (e.g., [87,194]) due to its wide use in the image/video compression standards. Other color spaces have also been used, e.g., the YOZ space [174]. However, it is often that only the luminance component of the signal is used for efficiency [119,145,197] because of its more important role in human visual perception than chrominance components, especially in quality evaluation of compressed images (it is worthwhile to point out that most coding decisions are made based on the luminance manipulations in the current image/video compression algorithms).

Temporal decomposition can be made via a sustained (low-pass, with lower f_t) and transient (band-pass, with higher f_t) filters [47,179] to stimulate two different visual pathways. Then, based upon the fact that receptive fields in the primary visual cortex resemble Gabor patterns [33] that can be characterized by particular spatial frequency f_s and orientation r, many types of spatial filters (e.g., Gabor filters, Cortex filters [172], wavelets, Gaussian pyramid [11], steerable pyramid filters [143,179]) can be used to decompose each temporal channel.

3.2. Features and artifact detection

Detection of a number of signal features and artifacts is common to visual quality evaluation in diversified scenarios. First of all, meaningful visual information is conveyed by contrast (that of luminance, color, motion, etc.). A largely uniform picture carries little or no information. The HVS perceives much more of signal contrast than the absolute signal strength, because it has specialized cells to process signal contrast (rather than absolute signal) [59,77]. This is also why contrast is central to CSF, luminance adaptation, contrast masking, visual attention, and so on.

In addition, certain structural artifacts occur in the prevalent signal compression and delivery process, which result in annoying effects to the viewer. Major coding artifacts include blockiness, blurring, edge damage and ringing [142,198]. The traditional measures (MSE or PSNR) fail to reflect the perceptual effect of such structural artifacts. In fact, blurring exists even in uncompressed images and video due to the imperfect PSF (point spread function) and out-of-focus of the imaging system, as well as object motion during the signal capturing process [3,39]. Moreover, the effects of motion and jerkiness have also been investigated [122,134,192] since they distinguish the evaluation of video from that of images.

3.2.1. Contrast

In Peli's work [128], local image contrast is evaluated by a band-pass-filtered image and a lowpass-filtered one. Following a similar methodology, contrast has been evaluated as the ratio of the combined analytic oriented filter response to the low-pass filtered image in the wavelet domain [184] or the ratio of high-pass and low-pass responses in the Harr wavelet space [79]. Luminance contrast was estimated as the ratio of the noticeable pixel change to the average luminance in a neighborhood [87]. The contrast was also calculated as a local difference of the reference video frame and the processed one with the Gaussian pyramid decomposition in [28], or by comparing DCT (discrete cosine transform) amplitudes with the amplitude of the DC coefficient for the corresponding block in [174].

For color and texture contrast [92], the k-means clustering algorithm can be used to group all image blocks. The largest cluster is viewed as the image backgroud. The contrast is then calculated as the Euclidean distance from the means of the corresponding background cluster. Motion contrast is obtained by evaluating relative motion [92,191] (i.e., object motion against the background).

3.2.2. Blockiness

The blocking artifact is a prevailing degradation caused by the block-based DCT coding technique, especially under low bit-rate conditions, due to the different quantization sizes used in the neighboring blocks and the lack of consideration for inter-block correlation. If an image I of size $W \times H$ is divided into $N \times N$ blocks, the horizontal and vertical difference (discontinuity) at block boundaries can be evaluated as [189]:

$$M_h = \left[\sum_{k=1}^{H/N-1}\sum_{x=0}^{W-1}(I(x,k\cdot N-1)-I(x,k\cdot N))^2\right]^{1/2} \quad (1)$$

for horizontal blockiness, and

$$M_v = \left[\sum_{l=1}^{W/N-1}\sum_{y=0}^{H-1}(I(l\cdot N-1,y)-I(l\cdot N,y))^2\right]^{1/2} \quad (2)$$

for vertical blockiness.

Variations of this method can be found in [107,121]. Object edges at block boundaries can be excluded in blockiness consideration [197]. Luminance adaptation and texture masking have recently been considered for blockiness evaluation [199].

An alternative method for gauging blockiness is through harmonic analysis [147], which may be used when block boundary positions are unknown beforehand (e.g., with video being cropped, re-taken by a camera, or coded with variable block sizes).

3.2.3. Blurring

Blurring can be effectively measured around edges in an image, since it is most noticeable there and such detection is efficient (because edges only constitute a small fraction of all image pixels). When the reference image is available, the extent of blur can be estimated via contrast decrease on edges [87]. In the cases with only a test image available, various blind methods to measure the blur/sharpness in an image have been proposed with edge spread detection [100,120], kurtosis [16,201], frequency domain analysis [78,98], PSF estimation [190], width/amplitude of lines and edges [35], and local contrast via 2-D analytic filters [180].

3.2.4. Motion and jerkiness

For visual quality evaluation of coded video, the major temporal distortion is jerkiness which is mainly caused by frame dropping [127] and is very annoying to the viewer that prefers continuous and smooth temporal transition. For decoded video without availability of the coding parameters, frame freeze can be simply detected by frame difference [85]; in the cases when the frame rate is available, the jerkiness effect can be evaluated using the frame rate [122,134,192] or more comprehensively, both the frame rate and temporal activity (i.e., motion) [60,91]. The location, number and duration of lost frames were estimated via inter-frame correlation analysis in [109], while lost frames and the density of group dropping were detected by inter-frame dissimilarity to measure fluidity in [126,127], in which it was concluded that, for the same level of frame loss, scattered fluidity breaks introduce less quality degradation than aggregated ones. The impact of time interval be-

tween occurrences of significant visual artifacts was studied in [144].

3.3. Just-noticeable distortion (JND) modeling

As we mentioned in the Introduction, not every change in an image is noticeable. The JND refers to a visibility threshold below which a change cannot be detected by the majority (e.g., 75 %) of viewers [67,72,85]. The JND modeling tackles the problem of visual similarity, and reflects the local characteristics of the HVS. Obviously, if a difference is below the JND value, it can be ignored in visual quality evaluation.

3.3.1. Subband-based JND function

DCT-based JND is the most investigated topic among all subband-based JND functions since DCT has been used in all existing image/video compression standards such as JPEG, H.261/3/4, MPEG-1/2/4, and SVC. The general form of the DCT-subband luminance JND function can be expressed as [85,150,173]:

$$JND_D(n,i,j) = T_{s\_csf}(n,i,j)\prod_{\wp}\alpha_{\wp}(n,i,j), \tag{3}$$

where n is the index of the DCT block and (i,j) represents a subband in this block, $T_{s\_csf}(n,i,j)$ is the base threshold due to the spatial CSF, and $\alpha_{\wp}(n,i,j)$ is the elevation parameter due to an effect $\wp$, such as luminance adaptation, intra-band masking, inter-band masking, temporal masking, and chrominance masking. The most practical solutions so far in the literature for determining different parameters in Eq. (3) are introduced as follows.

The widely-used formula developed by Ahumada and Peterson [2] for the base-line threshold $T_{s\_csf}(n,i,j)$ firstly fits spatial CSF curves in Fig. 1 with a parabola equation, which is a function of spatial frequencies specified by (i,j) and background luminance, and then compensates for the fact that the psychophysical experiments for determining CSF were conducted with a single signal at a time and with only spatial frequencies along one direction.

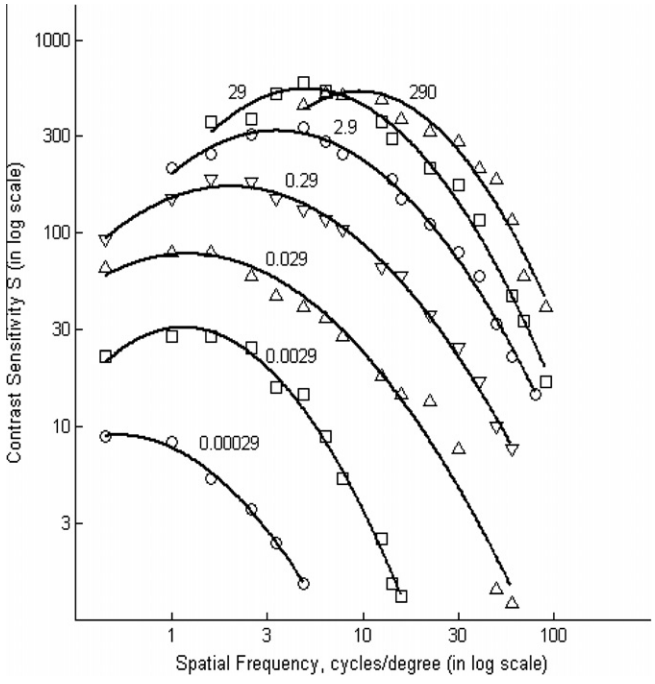

Fig. 1. Spatial CSF curves [2] (based upon the experiments in [155]) with seven different background luminance levels (labeled in cd/m^2).

Luminance adaptation factor $\alpha_{lum}(n)$ has been determined [203] to represent the variation versus background luminance shown in Fig. 2, for being more consistent with the findings of subjective viewing for digital images [23,67,114].

Intra-band masking effect $\alpha_{intra}(n,i,j)$ was addressed in [54,173] by comparing the corresponding DCT coefficient $C(n,i,j)$ against $T_{s\_csf}(n,i,j)\cdot\alpha_{lum}(n)$:

$$\alpha_{intra}(n,i,j) = \max\left(1, \left|\frac{C(n,i,j)}{T_{s\_csf}(n,i,j)\cdot\alpha_{lum}(n)}\right|^{\zeta}\right), \tag{4}$$

where ζ lies between 0 and 1.

Inter-band masking effect $\alpha_{inter}(n,i,j)$ can be assigned with low-, medium- or high-masking after classifying DCT blocks into smooth, edge and texture ones [150,203], according to energy distribution among subbands.

As for the inclusion of temporal CSF effect, the velocity $v(n)$ perceived by the retina for an image block needs to be estimated [31], and it is the object velocity $v_o(n)$ in the image (detected via block matching or optical flow) after off-setting by the eye movement velocity $v_e(n)$ (as determined in [31]):

$$v(n) = v_o(n) - v_e(n). \tag{5}$$

This is because the eye tracking toward a moving object reduces $v(n)$ appearing on the retina. A formula for incorporating the effect of $v(n)$ for temporal CSF in JND is given in [68].

The JND can also be defined in other frequency bands (e.g., Laplacian pyramid image decomposition [136], discrete wavelet transform (DWT) [176,9,166]). In comparison with DCT-based JND, significantly more research is needed for DWT-based JND, because DWT is a popular alternative transform, and more importantly, has similarity to the HVS in its multiple subchannel structure and the frequency-varying resolution. In addition, chrominance masking [1] needs more convincing investigation for all subband domains.

3.3.2. Pixel-based JND function

There are situations where JND estimated from pixels is more convenient and efficient to use, e.g., motion search [194], video replenishment [21], filtering of motion-compensated residuals [195] and edge enhancement [88], since the operations are usually performed on pixels rather than subbands. For quality evaluation of images and video, pixel-domain JND models avoid unnecessary subband decomposition.

Most pixel-based JND functions developed so far have used luminance adaptation $T^l(x,y)$ and texture masking $T^t(x,y)$ (see

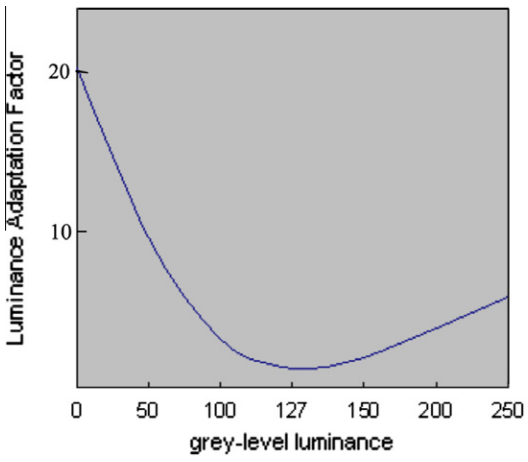

Fig. 2. Luminance adaptation for 8-bit images.

description below). The general pixel-based JND model can be expressed as [195]:

$$JND_p(x,y) = T^l(x,y) + T^t(x,y) - \kappa^{lt}(x,y) \cdot \min\left\{T^l(x,y), T^t(x,y)\right\}, \tag{6}$$

where (x,y) represents a pixel position, $\kappa^{lt}(x,y)$ accounts for the overlapping effect between $T^l(x,y)$ and $T^t(x,y)$, and $0 < \kappa^{lt}(x,y) \leqslant 1$. $T^l(x,y)$ can be determined by an approximation of Fig. 2 [23]. $T^t(x,y)$ can be estimated as:

$$T^t(x,y) = \beta \cdot g(x,y), \tag{7}$$

where $g(x,y)$ denotes the weighted average of gradients around (x,y) [23]; β takes smallest, medium and largest values for smooth, edge and texture neighborhood, respectively [89,195], because the masking effect in smooth regions can be largely neglected and distortion around edge is easier to be noticed than that in textured regions.

The models in [21,23] appear to be two special cases of this generalized formula (6), when $T^l(x,y)$ is assumed to have the dominant effect and the maximum effect of $T^l(x,y)$ and $T^t(x,y)$ is used, respectively.

The temporal effect was addressed in [22] by multiplying (6) with an elevation parameter increasing with inter-frame changes. The major shortcoming of pixel-based JND modeling lies in the difficulty of incorporating CSF explicitly, except for the case with conversion from a subband domain [204].

3.4. Visual attention (VA) map generation

As we mentioned in the Introduction, not every difference of an image receives the same attention level (the second problem). This is due to the fact that the HVS selects a part of the visual signal for detailed analysis and then responds. The VA refers to the selective awareness/responsiveness to visual stimuli [24,76], as a consequence of the human evolution.

There are two types of cues that direct attention to a particular point [133]: the bottom-up ones that refer to the external stimuli, while the top-down ones caused by a voluntary shift in attention (e.g., when the subject is given a prior information/instruction for directing the attention to a location/object). The VA process can be regarded as two stages [125]: in the pre-attentive stage, all information is processed across the entire visual field; in the attention stage, the features may be bound together (feature integration [152], especially for a bottom-up process) or the dominant feature is selected [34] (especially for a top-down process).

Most existing computational VA models are bottom-up, i.e., based upon contrast evaluation of various low-level features in images, in order to determine which locations *stand out* from their surroundings. As to the top-down (or task-oriented) attention, there is still a call of more focused research, although some initial work has been done, e.g., [55,113].

An influential bottom-up VA computational model was proposed by Itti et al. [62] for still images. An image is firstly low-pass filtered and down-sampled progressively from scale 0 (the original image size) to scale 8 (1:256 along each dimension); so the higher the scale index, the smaller the image size. This is to facilitate the calculation of feature contrast, which is defined as:

$$\widehat{F}(e,q) = \left|F(e) - F^l(q)\right|, \tag{8}$$

where F represents the map for one of the image features as follows: intensity, color and orientation; $e \in 2,3,4$ and $F(e)$ denotes the feature map at scale e; $q = e + \delta$, with $\delta \in 3,4$, and $F^l(q)$ is the interpolation to the finer scale e from the coarse scale q. In essence, $\widehat{F}(e,q)$ evaluates pixel-by-pixel contrast for a feature, since $F(e)$ represents the local information while $F^l(q)$ approximates the surroundings.

With one intensity channel, two color channels and four orientation channels (0°,45°,90°,135°; detected by Gabor filters), there are 42 feature maps computed: six for intensity, 12 for color, and 24 for orientation, as illustrated in Fig. 3. After cross-scale combination and normalization, the winner-take-all strategy localizes the most interested location on the map.

There is an alternative approach to detect bottom-up VA. Given an image $I(x,y)$, its Fourier Transform (FT) is $FT(I(x,y)) = A(u,v)e^{j\phi(u,v)}$, where $A(u,v)$ and $\phi(u,v)$ represent the FT amplitude and phase respectively. The VA map can be determined as [58]:

$$R(x,y) = IFT\left(e^{j\phi(u,v)}\right), \tag{9}$$

where IFT denotes the Inverse Fourier Transform. Eq. (9) implies that $A(u,v)$ is forced to be unity. Since $A(u,v)$ is the spectrum of spatial distribution (u,v) in the image, a unity $A(u,v)$ in Eq. (9) actually treats all (u,v) components to be with equal occurrence after IFT, i.e., the spatial components which have big difference (contrast) with the rest stand out; this is of similar objectives with Itti et al.'s model [62], although the approach adopted is different. Itti et al.'s model [62] is preferred if certain features (like orientation) need to be treated differently in a specific circumstance or new features are added (since the architecture in Fig. 3 is open), while the FT based approach [58] is more computationally efficient.

The VA map along the temporal dimension (over multiple consecutive video frames) can also be estimated. In the scheme proposed in [92] for video, different features (such as color, texture, motion, human skin/face) were detected and integrated for the continuous (rather than winner-take-all) salience map. In the work of Ma et al. [95], aural attention was also considered and integrated with visual factors. It was done by evaluating sound loudness and its sudden change, and the support vector machine (SVM) was employed to classify each audio segment into speech, music, silence, and other sounds; the ratio of speech/music to other sounds was measured for saliency detection.

The contrast sensitivity reaches its maximum at the fovea and decreases towards the peripheral retina. The JND model represents

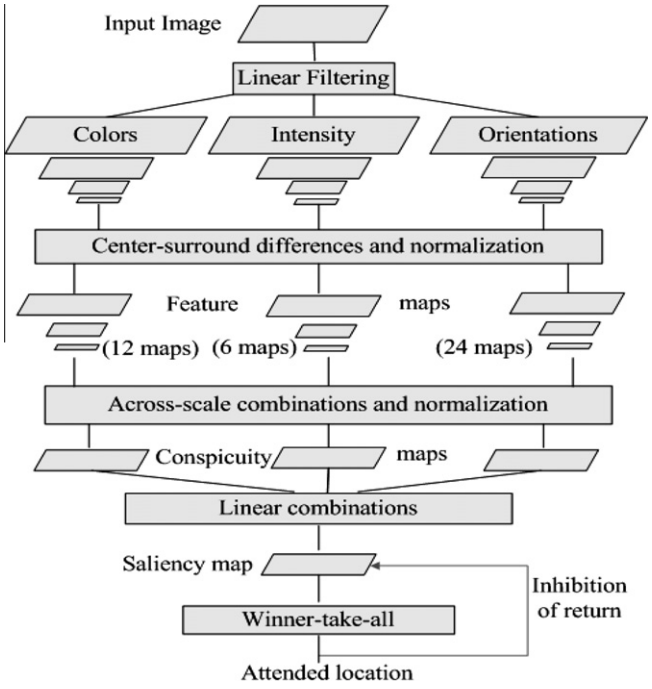

Fig. 3. Architecture of computational bottom-up VA model [62].

the visibility threshold when the attention is there. In other words, JND and VA account for the local and global responses of the HVS in appreciating an image, respectively. The overall visual sensitivity at a location in the image could be the JND modulated by the VA map [92]. Alternatively, the overall visual sensitivity may be derived by modifying the JND at every location according to its eccentricity away from the foveal points, with the foveation model in [169].

VA modeling is more meaningful for video than still images. If an observer has time long enough to perceive an image, many points of the image may become the attention center eventually. The perception to video is different. Every video frame is displayed to an observer within a limited time interval. Furthermore, motion may cause the viewer to pay attention to the moving part and trigger subsequent eye movement.

4. Perceptual visual quality metrics (PVQMs)

As introduced at the beginning of Section 3, there are two major categories of PVQMs: the vision-based modeling and signal-driven approach, according to the methodology being used. We can also classify PVQMs with regard to reference requirements: double-ended and single-ended. Double-ended metrics require both the reference (original) signal and the test (processed) signal, and can be further divided into two subclasses: reduced-reference (RR) metrics [57,185] that need only part of the reference signal and full-reference (FR) ones [17,139,140,167]) that need the complete reference signal. Single-ended metrics use only the processed signal, and are therefore also called no-reference (NR) ones [35,120,189]. Most existing PVQMs (almost all vision-based PVQMs and many signal-driven ones) are FR ones, which are expected to predict visual quality more accurately because more information is available as the ground of prediction.

The task of developing PVQMs can be usually considered as a two stage process: (a) feature detection and (b) pooling the features into a single number to represent the quality score. For the rest of this section, we will elaborate on several well referenced PVQMs (with emphasis on signal-driven ones because they represent more for the recent development), compare their differences, provide the links to some alternative solutions, and discuss some related issues that currently draw interests of the research communities and deserve more in-depth research.

4.1. Model-based PVQMs

As the simplest model-based approach, the HVS is regarded as a single spatial filter characterized by the spatial CSF, and such early models have been developed for images [43,97] and video [94,151]. More sophisticated model-based approaches have been researched [30,93,174,178], and are usually with FR and multi-channel signal decomposition as described in Section 3.1 together with evaluation of local contrast, spatiotemporal CSF and contrast/activity masking.

An early FR and multi-channel model called the visible difference predictor (VDP) was reported by Daly [30], where the HVS model accounts for sensitivity variations due to luminance adaptation, spatial CSF and contrast masking. Cortex transform is performed for signal decomposition, and different orientations are distinguished. Most existing schemes in this category follow a similar methodology as illustrated in Fig. 4, with difference in the color space adopted, spatio-temporal decomposition, and rules for error pooling. In the JNDmetrix model [28,93], the Gaussian pyramid [11] was used for decomposition with luminance and chrominance components in video. Gabor-related filters were used in [154], while the opponent color space (W-B, R-G and B-Y) and steerable pyramid filters were adopted in [178]. The work of [119] emphasizes on the distortion in spatial transitions. Decomposition are also carried out in the DCT [174] and wavelet [102] domains, respectively.

After signal decomposition (as described in Section 3.1 and the previous paragraph) in FR situations and if we follow the notations in Section 3.1, $s(c,f_t,f_s,r,x,y)$ represents the decomposed signal component, and its perceptual effect can be derived by considering the inter-channel masking [175,179], as the contrast gain control part in Fig. 4:

$$s_p(c,f_t,f_s,r,x,y)=\xi\frac{s(c,f_t,f_s,r,x,y)^{\rho}}{\varrho+\psi(c,f_t,f_s,r,x,y)\otimes s(t,c,f,r,x,y)^{v}}, \tag{10}$$

where ξ is a constant gain factor, ϱ is another constant to prevent division by zero, the excitatory part (the numerator) consists of a power-law nonlinearity of $s(c,f_t,f_s,r,x,y)$ with exponent ρ, and the inhibitory part (the denominator) basically is $s(c,f_t,f_s,r,x,y)$ with another exponent v convoluted with a pooling function $\psi(c,f_t,f_s,r,x,y)$ (e.g., a Gaussian kernel [179]). Eq. (10) aims at emulating the HVS masking phenomenon with different subbands, orientations, colors, frames and locations, for both reference and distorted video as illustrated in Fig. 4; it provides the inputs for feature pooling. The weighting parameters of channels in Section 3.1 and all parameters in Eq. (10) can be determined via the fitting of the model to CSF and contrast masking curves [179], or subjective viewing test scores [182,197].

4.2. Signal-driven PVQMs

More recent research effort has been directed to signal-driven PVQMs because model-based ones involve expensive computation and difficulties due to the gap between the knowledge for vision research (usually with simplistic, single stimulus (or two stimuli)) and the need for engineering modeling (for real-world, multiple stimuli).

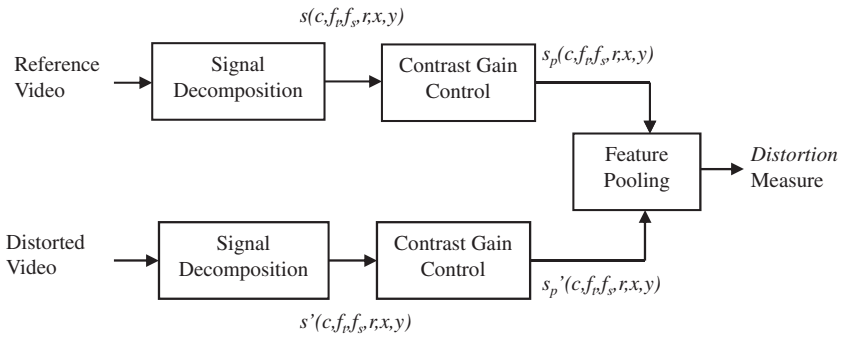

Fig. 4. Illustration of model-based PVQMs.

In comparison with model-based PVQMs, signal-driven ones do not attempt to build a comprehensive HVS model with regard to quality evaluation. Instead, they concentrate on signal modelling or processing of visual signals under consideration, and may incorporate appropriate domain knowledge (such as relevant compression/transmission artifacts). This approach is relatively less sophisticated and, therefore, computationally inexpensive. The signal-driven PVQMs can be of FR [140,162], RR [185] or NR [15,100,189].

4.2.1. Metrics with reference

For FR cases, if we follow the notations in Section 3.3.1 for JND, the DCT coefficient difference of the reference image $C_a(n,i,j)$ and that of the test one $C_b(n,i,j)$ is: $\Delta C(n,i,j) = C_a(n,i,j) - C_b(n,i,j)$; the DCTune system [173] measures perceptual distortion with the DCT-based JND as:

$$\Delta C_p(n,i,j) = \Delta C(n,i,j)/JND_D(n,i,j). \tag{11}$$

When $JND_D(n,i,j)$ is larger, the perceived distortion is less for a same amount of $\Delta C(n,i,j)$. A similar measure is defined with the pixel-based JND [87], and the resultant metric can be used for video with visual distortion and/or enhancement (to tackle the third problem highlighted in the Introduction), since noticeable contrast increase and decrease are distinguished for edges and non-edge.

An important aspect of the HVS perception is its sensitivity to image structure. The well cited SSIM (Structural SIMilarity) index was introduced by Wang and Bovik [165,167], and this FR metric can be expressed as

$$Q = \frac{\sigma_{ab}}{\sigma_a \cdot \sigma_b} \cdot \frac{2\sigma_a \cdot \sigma_b}{(\sigma_a)^2 + (\sigma_b)^2} \cdot \frac{2\bar{a} \cdot \bar{b}}{(\bar{a})^2 + (\bar{b})^2}, \tag{12}$$

where a and b denote the original and the test images, $\bar{a}$ and $\bar{b}$ are their means, σ_a and σ_b are the corresponding standard deviations, and σ_{ab} is the cross covariance. The three terms in Eq. (12) measure the loss of correlation, contrast distortion and luminance distortion, respectively. The dynamic range of Q is $[-1,1]$, and the best value is achieved if and only if $a = b$. Although SSIM bears certain similarity with MSE [36], the differentiating factor is the consideration of structural information. Singular Value Decomposition (SVD) is another way for feature detection with structural consideration to evaluate image quality using singular values (as MSVD [42]) or singular vectors [112].

With more theoretical ground, the VIF [140] (as an extension of IFC [139]) is based upon the assumption that the random field (RF) from a subband from the test image, D, can be expressed as:

$$D = GU + V, \tag{13}$$

where U denotes the RF from the corresponding subband from the reference image, G is a deterministic scale gain field, and V is a stationary additive zero-mean Gaussian noise RF. The proposed model takes into account additive noise and blur distortion; it is argued that most distortion types prevalent in real world systems can be roughly described locally by a combination of these two. The resultant metric based upon this model measures the amount of information that can be extracted about the reference image from the test one. In other words, the amount of information lost from a reference image as a result of distortion gives the loss of visual quality.

Another image metric with more theoretical ground is the VSNR [17] which operates in two stages. In the first stage, the contrast threshold for distortion detection in the presence of the image is computed via wavelet-based models of visual masking and visual summation, in order to determine whether the distortion in the test image is visible. If the distortion is below the threshold of detection, the test image is deemed to be of perfect visual fidelity (VSNR = infinity). If the distortion is above the threshold, a second stage is applied, which operates based on the property of perceived contrast, and the mid-level visual property of global precedence. These two properties are modeled as Euclidean distances in distortion-contrast space of a multiscale wavelet decomposition, and VSNR is computed based on a simple linear sum of these distances.

For RR PVQM development for video, low-level spatial–temporal features from the original video are extracted as the reference (instead of the whole image). In the work performed by Wolf and Pinson [185,186], both spatial and temporal luminance gradients are computed, to represent contrast, motion, amount and orientation of activity. Temporal gradients due to motion facilitate detecting and quantifying related impairments (e.g., jerkiness) using the time history of temporal features. Features from the reference video can be compared with those from the test video. The metric performed well in the VQEG FR-TV Phase II Test [157]. For another RR metric proposed in [57], spatial features are computed for the luminance image, and temporal features are obtained by the frame difference and global motion detection.

4.2.2. Metrics without reference

NR evaluation is closer to the HVS perception since human being does not need an explicit reference in judging picture quality. It also tends to be computationally efficient since processing reference data is not required. However, this often requires *a priori* knowledge about the distortions of visual signals to be evaluated. Thus, it is likely to be domain specific.

The models discussed in Section 3.2 can serve as or combine to be NR metrics. Examples of single-factor NR PVQMs are those for blockiness [147,189], blurring [100,190] and jerkiness [18,192]. A metric was devised in [101] for combining both blur and ringing measures in JPEG200 coded images. Blockiness, blurring and jerkiness were assessed together in [181]. Six spatial and temporal factors, namely, jerkiness, blurring, orientation, brightness and unstableness, were used in [104] for home video quality evaluation. Recently, video is classified into four visual content classes [117], and then a different NR metric is applied to a different class of video, since a single metric cannot be well tuned for all situations.

The concepts of just noticeable blockiness and blur have been proposed in [199,46], respectively. Blockiness detection was combined with consideration of HVS sensitivity, luminance adaptation, temporal masking, and intra- and inter-coefficient spatial masking (all these are factors being accounted for JND), in the DCT domain in [29]. In [83], picture quality is evaluated based on blur/sharpness and ringing measurements weighted by the pixel-based JND.

Another challenge for NR metrics is the possibility of mistaking the actual content as distortion. Note that separation of the content from distortion can be performed in the FR case (e.g., object edges are separated from blockiness [197]), and this is more difficult in the NR circumstance. In addition, end-to-end VPQMs would be useful by combining bitstream (without full decoding) and network loss analysis [183] for real-time, multi-channel quality monitoring (e.g., over IP networks), and therefore remain as a challenging research area.

4.3. More discussion on related issues

4.3.1. Feature pooling

Besides feature detection, feature pooling plays a crucial role to the success of PVQMs. For quality gauge of images and video, evaluation of all features (for both model-based and signal-driven metrics) has to be summarized to a single-number result, in analogy with the integration of various channels with the primary cortex in the brain.

For the majority of the existing models developed so far, integration has been accomplished by a percentile evaluation [28], simple summation [42,71], linear (i.e., weighted) combination

[107,28] or Minkowski pooling [38,174,178]. These pooling techniques, however, impose constraints on the relationship between the features and the quality score. For example, a simple summation or a weighted combination of features implicitly constraints the relationship to be linear, while the use of Minkowski metric for spatial pooling of the features/errors implicitly assumes that errors at different locations are statistically independent.

In [170], a method known as information content-weighted pooling has been presented, in which the weights are determined by local image content, assuming that the image source is a local Gaussian model and the visual channel is an additive Gaussian model. Pooling can be accomplished by accounting for VA and quality difference within an image (i.e., giving higher weights to low quality portion) [110]. In [107], since five types of impairments after low-pass filtering, blockiness assessment, correlation evaluation, masking analysis, and luminance adaptation and spatial CSF consideration are not entirely independent, the principal component analysis (PCA) is carried out to yield a quality index for a decoded image against its original. More recently, machine learning techniques have emerged as a way for pooling [112,117], due to their generalization capability with massive, high dimensional data, through training. More in-depth research is needed to explore if such a data driven approach provides better solutions.

Another aspect of pooling is to account for temporal effects. So far video evaluation is largely formulated as some form of *averaging* of multiple video frames. There have been continuing effort to combine spatial and temporal factors (spatiotemporal CSF formulation [68,177] and temporal error evaluation [4,116,168] in a better-grounded manner).

4.3.2. Variations of viewing conditions

External factors to be considered in PVQMs are variations of viewing conditions, e.g., ambient illumination, display resolution and viewing distance. The effect of viewing distance is actually related to display resolution. There has been limited research on the influence of the viewing distance, and the issues related to ambient illumination are largely uninvestigated. The VSNR metric [17] has been devised with the viewing distance of roughly 3.5 times of the image height, and claimed to provide a reasonable approximation of typical viewing conditions. The SSIM has been extended to multiple scales [171] by firstly downsampling both the reference and test images into different image resolutions (i.e., scales), and then replacing the first two terms in Eq. (12) (i.e., the measures of loss of correlation and contrast distortion) with the product of their counterparts in all scales, while the last term (luminance distortion) remains unchanged (still evaluated at the original resolution). However, the multi-scale SSIM does not always yield better results than its single-scale version [132]. Multi-scale has also been exploited in IFC [139] and VIF [140] via the steerable pyramid transform. Multi-scale is just a way more to compromise the effect of different viewing settings than to cater for a particular setting. In addition, it is still a problem on how to pool the calculated errors from different scales and decouple the overlapping among different scales.

A simple, empirical method [164] has also been proposed for SSIM to determine the downsampling scale Z for evaluating images viewed from a typical distance:

$$Z = max(1, round(H/256)), \tag{14}$$

where H is the image height. There are not explicit parametric choices for viewing condition variations in the major existing PVQMs. Obviously it remains a challenge in future research for metrics to account for viewing-condition variations (display resolution, ambient illumination and viewing distance) more convincingly for both benchmarking and practical use. The publicly available image databases and their viewing conditions will be presented and discussed in Section 5.1.

4.3.3. PVQMs for computer-generated visual signal

For computer graphics and animation, existing PVQMs have begun to find applications, with new, specific metrics to be developed. The model-based metric originally devised for natural images [93] was used in [8] for visual signal synthesis. In [14], the signal-driven metric devised with spatiotemporal CSF model and compensated by the eye movement [31] has been combined with a VA model, to predict error visibility in image rendering. A JND-based metric was used in indirect illumination calculation [136], accounting for the spatial CSF and luminance adaption. Perceptual quality criteria have been devised [19,149,193] according to mesh and texture resolutions, for transmission of 3D geometric mesh vertices and texture data. Other relevant work can be found in [5,75,135].

Human perception modeling can play an important role [20,37,45,149] in most computer graphics (CG) tasks. As pointed out by Tumblin and Ferwerda in [153], "*the goal of computer graphics is not to control light, but to control our perception of it*". Unlike in the cases of natural images and video, we do not have the original visual signal as the reference in CG so perception is the only criterion for processing.

Computer-generated signals have their own characteristics, statistics and requirements, in comparison with those acquired via cameras. In addition, some information that is hard to obtain in natural images is actually available in the CG cases; examples of such information are segmentation, depth, etc. The PVQMs specific to CG and animation are relatively primitive. So it is reasonable to expect more research to emerge in the area; this is also because graphics and animation become increasingly indispensable in many applications and services.

4.3.4. The role of VA

There is no doubt that VA is important to the HVS perception. However, there are diversified and even controversial views toward its role in visual quality evaluation. Improvement has been reported by using the VA map to weight a quality map for perceptual quality prediction [90,92,110] and to guide CG rendering [14,75]. On the other hand, it has been argued that VA is not always beneficial for PVQMs (at least for simple weighting) [115]. Even when metrics were reported to be improved using recorded VA data (via a eye tracker) [81], it has been also observed that greater improvement was found with VA recorded in task-free viewing than in the cases of subjects being asked to assess the picture quality. This seems to be related to top-down (task-oriented) aspect of VA.

Visual quality may be influenced by not only attentional regions, but also non-attentional ones, since as introduced in Section 3.4, the HVS's visual information processing is over the whole visual field in the pre-attentive stage of VA. Therefore, besides the quality of attended regions, that of unattended regions needs to be properly fused into the overall quality index, as an early attempt in [196]. Some research argued that distortion in image compression (with JPEG and JPEG 2000 artifacts) and transmission (with packet loss) change the subjects' eye fixation and the associated duration [158], while another work indicated that there is not obvious difference in the VA maps between a test video sequence and its original [105]. In summary, VA's influence on visual quality evaluation is still an open issue for research.

5. Databases and performance evaluation for PVQMs

Subjective viewing data are essential for verification of various VPQMs. The ITU has standardized methods to conduct subjective

viewing tests [63–65], to promote acceptance and facilitate sharing of the resultant MOS and DMOS (differential MOS). For fair benchmarking, the developed PVQMs must be evaluated with a wide variety of visual contents and distortion types to make meaningful conclusions about their performance. It is therefore more convincing to use multiple databases with MOS/DMOS from different sources since evaluation with one single database may not be comprehensive and general [146]. In this section, we demonstrate this for FR signal-driven image metrics.

We choose the better cited metrics for comparison and these metrics are formulated for quality evaluation in general (i.e., without prior knowledge on distortion), rather than that for certain type(s) of distortion. The objective of the benchmarking in this work is the comprehensive evaluation across different databases and with various distortion types together.

5.1. Public image databases with MOS/DMOS

There are publicly available subjective viewing databases for images: LIVE [141], CSIQ [80], IVC [12], Toyama [56], A57 [32], TID [131] and WIQ [40]. The important information about these seven databases is listed in Table 3. As can be seen, the types of distortion vary across the databases; there are two different types of subjective quality scores being used: MOS and DMOS, and their ranges are different. There are a total of 3832 distorted (test) images with all these databases.

There has been some control on the viewing distance and illumination in most of the databases we used in this work. The viewing distance is 4 times of the image height for IVC and WIQ, and 6 times of the image height for Toyama, while it is kept at about 3–3.75 times of the image height for most images in LIVE. For CSIQ, subjects were instructed to keep the viewing distance stable of approximately 80 cm, with image resolution of 1920×1200. Viewing conditions were not fully controlled in A57 and TID, to emulate practical scenarios where variations in viewing distance are difficult to control. The experiments in the majority of the databases are performed with normal indoor lighting, while those of WIQ (with 80 images) are done in dark rooms.

The TID database contains the largest number of test images among the seven; its MOS is the result of 654 subjects from three different countries (Finland, Italy and Ukraine), and more than 200 subjects rated each image, while 7–30 subjects were used for the other databases. In addition, TID covers more distortion types, and contains more less common distortion types (to be discussed in Section 5.3 below).

5.2. Image metric benchmarking

We will demonstrate the performance comparison for several existing major FR image metrics described earlier: SSIM [167], VSNR [17], IFC [139] and VIF [140] and MSVD [42], with reference of PSNR. For VSNR, VIF, IFC and SSIM implementation, we have used the publicly accessible Matlab package [49]; they are the original codes provided by the respective algorithm designers. The MSVD method was also implemented in MATLAB. The image scale for SSIM has been decided via Eq. (14), and this provides better performance than the cases without scaling. In the experiments of this work, we have maintained all metric parameters the same for different databases since there are not explicit parametric choices for viewing condition variations with these metrics, as mentioned in Section 4.3.2.

A 5-parameter logistic mapping specified in [156] between the objective outputs and the subjective quality ratings was employed, to remove any nonlinearity due to a subjective rating process and to facilitate the metric comparison in a common analysis space. There are two criteria commonly used for performance comparison, namely: Pearson linear correlation coefficient C_P (for prediction accuracy), and Spearman rank order correlation coefficient C_S (for monotonicity), between the MOS/DMOS and the objective prediction, and $0 \leqslant C_P, C_S \leqslant 1$. For a perfect match between the objective and subjective scores, $C_P = C_S = 1$. The performance can be also measured by the root-mean-square error (RMSE) between the MOS (or DMOS) and the metric output.

From the results with above-mentioned databases, we plotted C_P, C_S and the associated 95% confidence interval (CI) [25,156] in Figs. 5 and 6, for different metrics with luminance images. We can see that C_P and C_S give fairly consistent results. The RMSE measure also tell a similar story as C_P and C_S so its results are not presented here to save space. We can see from the figure that the perceptual metrics under comparison outperform PSNR in general. We note that VIF outperforms its predecessor IFC in most cases as expected, and SSIM and VIF are better metrics due to their overall performance over different databases.

As can be seen from the comparison of C_P, C_S and the CI in Figs. 5 and 6 between the two better performing metrics, VIF outperforms SSIM with a clear margin (i.e., non-overlapped CIs) in LIVE and CSIQ; for the other five databases, VIF and SSIM tells a mixed story: one metric is better than the other in some cases and the CIs are overlapped for all cases.

The performance of all five perceptual metrics is relatively bad on WIQ database which contains more than one artifact (like

Table 3
Description of image databases (n_0: number of original images; n: number of test images; R: image resolution (notes: LIVE has many images of size 768×512, but also of other size like 480×720, 632×505, 634×505, 618×453 and 610×488); **S**: type of subjective quality score).

Name	n_0	n	R	S (range)	Distortion types
LIVE	29	779	(See notes)	DMOS (0–100)	JPEG-2 K compression; JPEG compression; White Gaussian noise; Gaussian blurring; Rayleigh-distributed bit errors of JPEG -2 K stream or Fast fading distortion
CSIQ	30	866	512×512	DMOS (0–1)	JPEG compression; JPEG-2 K compression; global contrast decrements; additive pink Gaussian noise; additive white Gaussian noise
	10	185	512×512	MOS (0–5)	JPEG-2 K compression; JPEG compression;LAR (locally adaptive resolution) coding; blurring
Toyama	14	168	768×512	MOS (1–5)	JPEG-2 K compression (with JasPer s/w);JPEG compression (with cjpeg s/w)
A57	3	54	512×512	DMOS (0–1)	LH-subband quantization of a 5-level DWT with 9/7 filters; additive Gaussian white noise; baseline JPEG compression; JPEG-2 K compression; JPEG-2 K compression with greater to fine spatial scales to preserve global precedence; blurring
TID	25	1700	512×384	MOS (0–9)	Additive Gaussian noise; spatially correlated noise; masked noise; high frequency noise; impulse noise; quantization noise; Gaussian blur; image denoising; JPEG compression; JPEG-2 K compression; JPEG transmission; JPEG-2 K transmission; non eccentricity pattern noise; block-wise distortion of different intensity; mean shift; overall contrast change
WIQ	7	80	512×512	DMOS (0–100)	Wireless imaging artifacts, which are not considered in other publicly available image quality databases

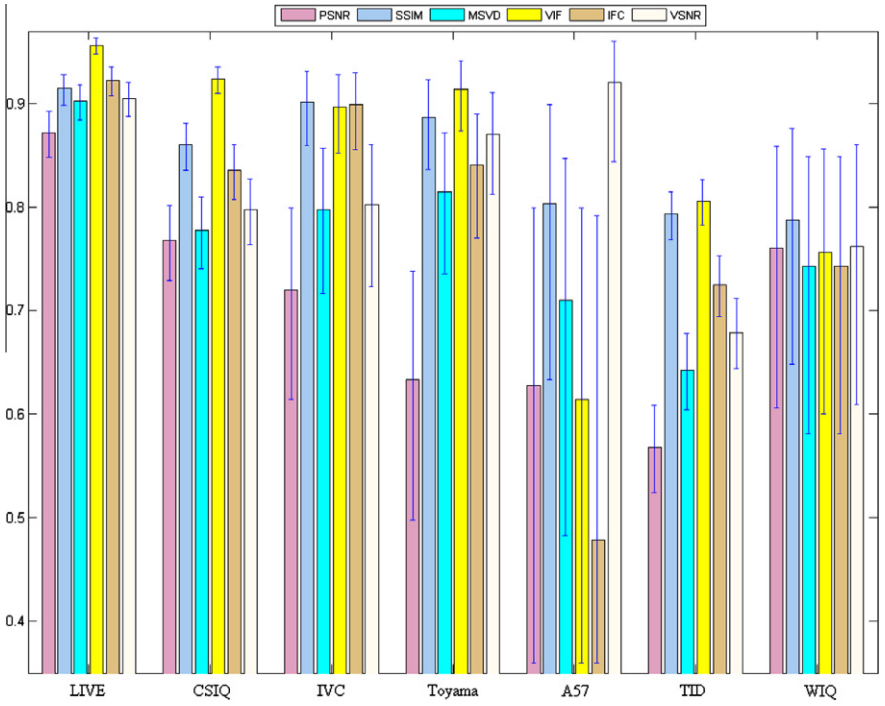

Fig. 5. C_P of different metrics for various databases (with 95% CI indicated).

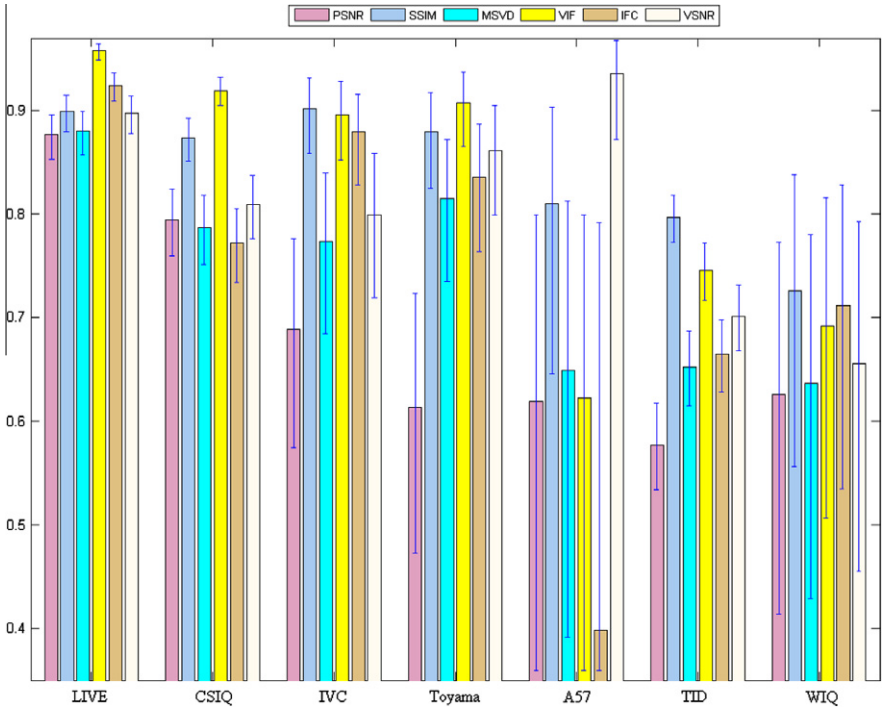

Fig. 6. C_S of different metrics for various databases (with 95% CI indicated).

blocking and ringing together in a same image) due to the complex nature of a wireless communication link. For the A57 database, the performance of VIF, SSIM, MSVD and IFC is also relatively poor for the similar reason; however, VSNR which performs well for A57 database does not perform as well with the other databases. For TID, the metric performance is not high because it contains more distortion types, as well as some less-common data (to be explained next). From the results presented above, we can see that testing with just a single database is not sufficient to evaluate a metric.

5.3. Evaluation with the less-common dataset in TID

We tested a subset of 500 images in the TID database, with 100 images from each of the following five distortion/change types [131,132]: mean (intensity) shift, contrast change, image denois-

ing, non eccentricity pattern noise, and local block-wise distortion of different intensity. These types of distortion are less common because they are only available in TID.

The mean shift and contrast change (up to a certain level) generally do not affect the visual quality substantially although the PSNR may change considerably. The denoised images were obtained after applying different denoising filters. The PSNR of denoised images is generally higher than that of the noisy images, although some denoised images visually look worse than the corresponding noisy images [131]. For non-eccentricity distortion, a small image fragment of size 15×15 pixels has been randomly copied from a nearby location (with distance of a few pixels) in the reference image. Due to the high correlation between the copied block and the replaced one, such error is not easy to be perceived by the HVS (so PSNR tells little about image quality). As for the local block-wise distortion, there are four levels of distortion: the 1st to 4th levels of distortion have 16, 8, 4 and 2 blocks being distorted in each image, but are associated with decreasing PSNR each level upward. An image in which two blocks were corrupted (i.e., the 4th distortion level) is perceived as having a better visual quality (although it has smaller PSNR) than the image with 16 corrupted blocks (i.e., the 1st distortion level); this is because a lower amount of distortion spreads over a larger area is likely to cause more quality degradation than a higher amount of distortion spreads over a smaller area.

Fig. 7 shows C_P (C_S is not shown because of its similar results) for different metrics with the set of 500 images. All PVQMs under consideration have lower C_P, compared with the corresponding results in Fig. 5, although they do much better than PSNR. The reason for SSIM to have relatively better performance than other metrics is its consistence (as to be explained next) with different types of distortion. In [80], SSIM, VSNR, and VIF have been compared with four individual types of the aforementioned distortion: image denoising, non eccentricity pattern noise, local block-wise distortion, and mean shift; according to the C_S results reported in [80], VSNR and VIF are the best metrics for the first two types of distortion respectively, while SSIM is the best metric for the last two types of distortion; more importantly, SSIM performs fairly consistently over the four distortion types, as contrasted with VIF and VSNR [80]. As can be seen from Figs. 5–7, although both VIF and VSNR have better theoretical grounding, they do not perform equally well for all the test datasets, and do not perform better

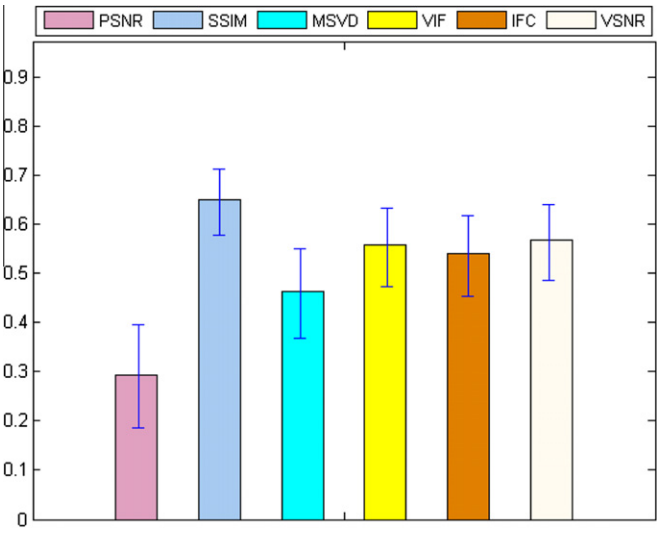

Fig. 7. C_P for different metrics with the subset of 500 images in TID (with 95% CI indicated).

than the simpler SSIM in terms of consistency. This may be attributed to the limitation of the assumptions made for these two metrics (please refer to the relevant formula and description for VIF and VSNR in Section 4.2.1).

6. Concluding remarks

Perceptual visual quality assessment aims at quantifying the quality of visual information, including still pictures and video. It can be extended to 3D models, computer graphics, animation, and 3D and multi-view visual data. This is an interdisciplinary field involving vision science, color science, signal processing, physiology, psychology and computer engineering. Many processes can affect and impair the quality of visual signals, including acquisition, compression, transmission, display, printing, and reproduction. Automatic visual quality assessment is crucial to multimedia systems by providing objective metrics for use during the design, implementation, optimization and testing stages, to avoid or reduce the need for extensive evaluation with human subjects (even in cases of human evaluation being possible).

This work has highlighted the importance, the challenges and the advances in objective, automatic visual quality assessment to align well with the human perception. Although the perceptual visual quality evaluation proves to be a difficult task, a considerable amount of research and development efforts has been directed to it and its applications, as surveyed in this paper and evidenced by the large number of cited references, as well as the recent dedicated annual workshops [26,27], special journals issues [51,70,106,118], and many special sessions in related conferences. A significant progress has been achieved in transferring relevant latest physiological and psychological findings, modeling various basic computational modules, and designing useful perceptual visual quality metrics (PVQMs). In addition, in spite of the fact that the PVQM technology is still in its infancy stage, there have been some industrial deployments [53,61,86], especially in the test equipment and manufacturing sectors, since a reasonable PVQM can be a differentiating factor to gain the competitive advantage. Based on the information collected from industrial contacts of the authors and Internet search, much more companies buy into PVQM-related ideas than five years ago, toward better consumer QoE (quality of experience).

Performance benchmarking for the commonly used image metrics has been demonstrated with open source codes and publicly available databases. The state of the art techniques can perform better than PSNR as being shown in this work, and need further advancement in terms of prediction accuracy, consistence and robustness.

With the survey presented earlier in this paper, we know that the following important aspects are relatively less investigated: temporal modeling for video (for JND, VA and others), chrominance evaluation toward a complete model, and joint multimedia (video, speech, audio, text, and so on) modeling. Signal-driven PVQMs have attracted substantial research efforts during the recent years. No-reference (NR) metrics can be less computationally expensive and adopted in a wider range of circumstances (even when the reference is not available), and therefore deserve more attention from researchers. Compressed-domain quality evaluation is useful because of the presence of a large number of coded visual signals at user/relay sites nowadays, as well as joint consideration of network loss (so that end-to-end quality control is possible). More research is therefore expected for joint distortion consideration of compression and transmission. For feature pooling, the machine-learning approach has good potential for generalization of the quality evaluating function from available data (with possible extension to the massive web data, as a result of the recent trends

of innovative use of visual content in the Internet as the ground truth of visual analysis [200]). As another dimension of development, metrics can be built for a specific codec (e.g., H.264, SVC) or application (e.g., mobile communication and hand-held devices) by incorporating the proper domain knowledge in the model. In addition, there is a call for new methodology to assess perceptual quality of IPTV, HDTV, 3D and multi-view data, with the progress in the related technology [6,44,96,103], since most of the existing metrics are for standard-definition and single-view visual data. We have also discussed the opportunities in visual attention, feature pooling, viewing condition handling, and computer graphics and animation in more details with Section 4.3.

Acknowledgement

The authors are grateful to Mr. Manish Narwaria and Mr. Yuming Fang, in the CemNet Lab, Nanyang Technological University, Singapore, for the help in preparing the experimental data for Section 5, and reproducing Figs. 1 and 3, respectively. This work is partially supported by MoE AcRF Tire 2 Grant, Singapore, Grant No.: T208B1218. The authors appreciate the editor and anonymous reviewers' constructive advice that has prompted us for two new rounds of re-thinking of our work and toward clearer presentation of the technical content in this paper.

References

[1] A.J. Ahumada, W.K. Krebs, Masking in color images, Proc. SPIE Human Vision Electron. Imaging VI (2001) 4299.
[2] A.J. Ahumada, H.A. Peterson, Luminance-model-based DCT quantization for color image compression, in: SPIE Proceedings of the Human Vision, Visual Processing, and Digital Display III, 1992, pp. 365–374.
[3] M.B. Banham, A.K. Katsaggelos, Digital image restoration, IEEE Signal Process. Mag. 14 (1997) 24–41.
[4] M. Barkowsky, B.E.J. Bialkowski, R. Bitto, A. Kaup, Temporal trajectory aware video quality measure, IEEE J. Sel. Top. Signal Process. 3 (2009) 266–279.
[5] D. Bartz, D. Cunningham, J. Fischer, C. Wallraven, The role of perception for computer graphics, in: Eurographics 2008 Annex to the Conference Proceedings (State-of-the-Art Reports), 2008, pp. 65–86.
[6] P. Benzie, J. Watson, P. Surman, I. Rakkolainen, K. Hopf, H. Urey, V. Sainov, C. von Kopylow, A survey of 3dtv displays: techniques and technologies, IEEE Trans. Circuits Syst. Video Technol. 17 (11) (2007) 1647–1658.
[7] C. Blakemore, F.W. campbell, Adaptation to spatial stimuli, J. Physiol. 200 (1969) 11–13.
[8] M.R. Bolin, G.W. Meyer, A visual difference metric for realistic image synthesis, SPIE Proc. Human Vision Electron. Imaging 3644 (1999) 106–120.
[9] A.P. Bradley, A wavelet visible difference predictor, IEEE Trans. Image Process. 8 (1999) 717–730.
[10] A.P. Bradley, F.W.M. Stentiford, Visual attention for region of interest coding in JPEG 2000, J. Visual Commun. Image Representation, 2003.
[11] P.J. Burt, E.H. Adelson, The laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (4) (1983) 532–540.
[12] P. Le Callet, F. Autrusseau, Subjective quality assessment irccyn/ivc database. From <http://www2.irccyn.ec-nantes.fr/ivcdb/>.
[13] F.W. Campbell, J.J. Kulikowski, Orientational selectivity of the human visual system, J. Physiol. 187 (2) (1966) 437–445.
[14] K. Cater, A. Chalmers, G. Ward, Detail to attention: exploiting visual tasks for selective rendering, in: Proceedings of the Eurographics Symposium on Rendering, 2003, pp. 270–280.
[15] J. Caviedes, S. Gurbuz, No-reference sharpness metric based on local edge kurtosis, Proc. IEEE Int. Conf. Image Process. (ICIP) 3 (2002) 53–56.
[16] J. Caviedes, F. Oberti, A new sharpness metric based on local kurtosis, edge and energy information, Signal Process.: Image Commun. 19 (2004) 147–161.
[17] D.M. Chandler, S.S. Hemami, Vsnr: a wavelet-based visual signal-to-noise ratio for natural images, IEEE Trans. Image Process. 16 (9) (2007) 2284–2298.
[18] J.Y.C. Chen, J.E. Thropp, Review of low frame rate effects on human performance, IEEE Trans. Systems Man Cybernet. 37 (2007).
[19] I. Cheng, A. Basu, Perceptually optimized 3-D transmission over wireless networks, IEEE Trans. Multimedia 9 (2) (2007) 386–396.
[20] K. Chiu, P. Shirley, Rendering, complexity, and perception, in: Proceedings of the 5th Eurographics Rendering Workshop (Darmstadt), 1994, pp. 19–34.
[21] Y.J. Chiu, T. Berger, A software-only videocodec using pixelwise conditional differential replenishment and perceptual enhancements, IEEE Trans. Circuits Syst. Video Technol. 9 (3) (1999) 438–450.
[22] C.H. Chou, C.W. Chen, A perceptually optimized 3-D subband image codec for video communication over wireless channels, IEEE Trans. Circuits Syst. Video Technol. 6 (2) (1996) 143–156.
[23] C.H. Chou, Y.C. Li, A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile, IEEE Trans. Circuits Systems Video Technol. 5 (6) (1995) 467–476.
[24] M.M. Chun, J.M. Wolfe, Visual attention, in: B. Goldstein (Ed.), Blackwell Handbook of Perception, Blackwell, Oxford, UK, 2001, pp. 272–310.
[25] N. Cliff, Analyzing Multivariate Data, Harcourt Brace Jovanovich, San Diego, CA, 1987.
[26] QOMEX Committee. International workshop on quality of multimedia experience QOMEX, 2009, 2010. From <http://qomex.org/>.
[27] VPQM Committee. International workshop on video processing and quality metrics for consumer electronics VPQM, 2005/06/07/08/09/10. From <http://www.vpqm.org>.
[28] Sarnoff Corporation. Sarnoff JND vision model, J. Lubin (Ed.), Contribution to IEEE G-2.1.6 Compression and Processing Subcommittee, 1997.
[29] F.-X. Coudoux, M.G. Gazalet, C. Derviaux, P. Corlay, Picture quality measurement based on block visibility in discrete cosine transform coded video sequences, J. Electron. Imaging 10 (2) (2001) 498–510.
[30] S. Daly, The visible differences predictor: an algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, 1993, pp. 179–206.
[31] S. Daly, Engineering observations from spatiovelocity and spatiotemporal visual models, in: C.J. van den Branden Lambrecht (Ed.), Vision Models and Applications to Image and Video Processing, Kluwer Academic Publishers, Norwell, MA, 2001.
[32] A57 dataset. From <http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html>.
[33] J.G. Daugman, Two-dimensional spectral analysis of cortical receptive field profiles, Vision Research 20 (10) (1980) 847–856.
[34] R. Desimone, J. Duncan, Neural mechanisms of selective visual attention, Ann. Rev. Neurosci. 18 (1995) 193–222.
[35] 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: Proceedings of the Internatioanl Conference on Computational Analysis on Images and Patterns (CAIP), Lecture Notes in Computer Science, vol. 2756, Springer, 2003, pp. 149–156.
[36] R. Dosselmann, X.D. Yang, A comprehensive assessment of the structural similarity index. Signal Image Video Process. (available on-line), 2010.
[37] D.S. Ebert, B. Buxton, P. Davies, E.K. Fishman, A. Glassner. The future of computer graphics: an enabling technology? in: SIGGRAPH, 2002.
[38] M.P. Eckert, A.P. Bradley, Perceptual quality metrics applied to still image compression, Signal Process. 70 (1998) 177–200.
[39] J.H. Elder, S.W. Zucker, Local scale control for edge detection and blur estimation, IEEE Trans. Pattern Anal. Mach. Intell. 20 (7) (1998) 699–716.
[40] U. Engelke, H.-J. Zepernick, M. Kusuma, Wireless imaging quality database. From <http://www.bth.se/tek/rcg.nsf/pages/wiq-db>.
[41] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.
[42] A.M. Eskicioglu, A. Gusev, A. Shnayderman, An SVD-based gray-scale image quality measure for local and global assessment, IEEE Trans. Image Process. 15 (2) (2006) 422–429.
[43] O.D. Faugeras, Digital color image processing within the framework of a human visual model, IEEE Trans. Acoust. Speech Signal Process. 27 (1979) 380–393.
[44] C. Fehn, 3D TV broadcasting, in: O. Schreer, P. Kauff, T. Sikora (Eds.), 3D Video Communication, Wiley, 2005.
[45] J.A. Ferwerda, Elements of early vision for computer graphics, IEEE Comput. Graph. Appl. 21 (5) (2001) 22–33.
[46] 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.
[47] R.E. Frederickson, R.F. Hess, Estimating multiple temporal mechanisms in human vison, Vision Res. 38 (7) (1998) 1023–1040.
[48] P. Frossard, O. Verscheure, Joint source/FEC rate selection for quality-optimal MPEG-2 video delivery, IEEE Trans. Image Process. 10 (12) (2001) 1815–1825.
[49] M. Gaubatz. Metrix mux visual quality assessment package. From <http://foulard.ece.cornell.edu/gaubatz/metrix_mux/>.
[50] M. Ghanbari, Standard Codes: Image Compression to Advanced Video Coding, IEE, London, UK, 2003.
[51] G. Ghinea, G.-M. Muntean, P. Frossard, M. Etoh, F. Speranza, H.R. Wu, Special issue: quality issues in multimedia broadcasting, IEEE Trans. Broadcasting 54 (3) (2008).
[52] B. Girod, What's wrong with mean-squared error?, in: A.B. Watson (Ed.), Digital Images and Human Vision, The MIT Press, 1993, pp. 207–220.
[53] OPTICOM GmbH, Perceptual voice, audio and video quality. From <http://www.opticom.de>.
[54] I. Hontsch, L.J. Karam, Adaptive image coding with perceptual distortion control, IEEE Trans. Image Process. 11 (3) (2002).
[55] J.B. Hopfinger, M.H. Buonocore, G.R. Mangun, The neural mechanisms of top-down attentional control, Nature Neurosci. 3 (2000) 284–291.
[56] Y. Horita, Y. Kawayoke, Z.M. Parvez Sazzad, Image quality evaluation database. From <http://160.26.142.130/toyama_database.zip>.
[57] Y. Horita, T. Miyata, I.P. Gunawan, T. Murai, M. Ghanbari, Evaluation model considering static-temporal quality degradation and human memory for sscqe video quality, Proc. SPIE Visual Commun. Image Process. 5150 (11) (2003) 1601–1611.
[58] X. Hou, L. Zhang, Saliency detection: a spectral residual approach, IEEE Conf. Comput. Vis. Pattern Recognition, 2007.

[59] D.H. Hubel, Eye, Brain, and Vision, Freeman, NY, 1988.
[60] G. Iacovoni, S. Morsa, R. Felice, Quality-temporal transcoder driven by the jerkiness, Proc. IEEE Int. Conf. Multimedia Expo (2005) 1452–1455.
[61] Tektronix Inc. Picture quality analysis system pqa200/500. From <http://www.tektronix.com>.
[62] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.
[63] ITU. Subjective assessment of standard definition digital television (sdtv) systems, ITU-R Recommendation BT.1129-2, 1998.
[64] ITU. Subjective video quality assessment methods for multimedia applications, ITU-T Recommendation P.910, 1999.
[65] ITU. Methodology for the subjective assessment of the quality of television pictures, ITU-R Recommendation BT 500-11, 2002.
[66] ITU-T Recommendation J.147. Objective picture quality measurement method by use of in-service test signals, ICRU Report 54, 2002.
[67] N. Jayant, J. Johnston, R. Safranek, Signal compression based on models of human perception, Proc. IEEE 81 (1993) 1385–1422.
[68] Y. Jia, W. Lin, A.A. Kassim, Estimating just-noticeable distortion for video, IEEE Trans. Circuits Syst. Video Technol. 16 (7) (2006) 820–829.
[69] P. Kaiser, R. Boynton, Human Color Vision, Optical Society of America, 1996.
[70] L. Karam, T. Ebrahimi, S. Hemami, T. Pappas, R. Safranek, Z. Wang, A.B. Watson (Guest Editors). Special issue on visual media quality assessment. IEEE J. Sel. Top. Signal Process. 3(2), April 2009.
[71] S.A. Karunasekera, N.G. Kingsbury, A distortion measure for blocking artifacts in images based on human visual sensitivity, IEEE Trans. Image Process. 4 (6) (1995) 713–724.
[72] B.W. Keelan, Handbook of Image Quality, Marcel Dekker Inc., 2002.
[73] D.H. Kelly, Motion and vision I: Stabilized images of stationary gratings, J. Opt. Soc. Am. 69 (9) (1979) 266–274.
[74] D.H. Kelly, Motion and vision II: stabilized spatio-temporal threshold surface, J. Opt. Soc. Am. 69 (10) (1979) 1340–1349.
[75] S.L. Kim, G.J.S. Choi, Real-time tracking of visually attended objects in virtual environments and its application to lod, IEEE Trans. Vis. Comput. Graph. 15 (1) (2009) 6–19.
[76] B. Kolb, I. Wishaw, Fundamentals of Human Neuropsychology, fourth ed., Freeman & Co., New York, 1996.
[77] S.W. Kuffler, Discharge patterns and functional organization of the mammalian retina, J. Neurophysiol. 16 (1953) 37–68.
[78] D. Kundur, D. Hatzinakos, Blind image deconvolutions, IEEE Signal Process. Mag. 13 (1996) 43–63.
[79] Y.-K. Lai, C.-C. Jay Kuo, A Haar wavelet approach to compressed image quality measurement, J. Vis. Commun. Image Representation 11 (1) (2000) 17–40.
[80] E.C. Larson, D.M. Chandler, Most apparent distortion: full-reference image quality assessment and the role of strategy, J. Electron. Imaging 19 (1) (2010).
[81] E.C. Larson, C. Vu, D.M. Chandler, Can visual fixation patterns improve image fidelity assessment? IEEE Int. Conf. Image Process., 2008.
[82] G.E. Legge, J.M. Foley, Contrast masking in human vision, J. Opt. Soc. Am. 70 (1980) 1458–1471.
[83] L. Liang, S. Wang, J. Chen, S. Ma, D. Zhao, W. Gao, No-reference perceptual image quality metric using gradient profiles for JPEG 2000, Signal Process.: Image Commun. (available on line), 2010.
[84] John O. Limb, Distortion criteria of the human viewer, IEEE Trans. Syst. Man Cybernet. 9 (12) (1979).
[85] W. Lin, Computational Models for Just-noticeable Difference, in: H.R. Wu, K.R. Rao (Eds.), Digital Video Image Quality and Perceptual Coding, CRC Press, 2006 (Chapter 9).
[86] W. Lin, Gauging image and video quality in industrial applications, in: Y. Liu, et al. (Eds.), Advances of Computational Intelligence in Industrial Systems, Springer-Verlag, Heidelberg, 2008.
[87] W. Lin, L. Dong, P. Xue, Visual distortion gauge based on discrimination of noticeable contrast changes, IEEE Trans. Circuits Syst. Video Technol. 15 (7) (2005) 900–909.
[88] W. Lin, Y. Gai, A.A. Kassim, A study on perceptual impact of edge sharpness in images, IEE Proc. Vision Image Signal Process. 153 (2) (2006) 215–223.
[89] A. Liu, W. Lin, M. Paul, C. Deng, F. Zhang, Just noticeable difference for images with decomposition model for separating edge and textured regions, IEEE Trans. Circuits Syst. Video Technol. 20 (11) (2010) 1648–1652.
[90] H. Liu, I. Heynderickx, Studying the added value of visual attention in objective image quality metrics based on eye movement data, IEEE Int. Conf. Image Process., 2009.
[91] Z. Lu, W. Lin, C.S. Boon, S. Kato, E. Ong, S. Yao. Perceptual quality evaluation on periodic frame-dropping video. IEEE Int. Conf. Image Process. (ICIP), 2007.
[92] Z. Lu, W. Lin, X. Yang, E. Ong, S. Yao, Modeling visual attention's modulatory aftereffects on visual sensitivity and quality evaluation, IEEE Trans. Image Process. 14 (11) (2005) 1928–1942.
[93] J. Lubin, A visual discrimination model for imaging system design and evaluation, in: E. Peli (Ed.), Vision Models for Target Detection and Recognition, World Scientific, Singapore, 1995.
[94] F. Lukas, Z. Budrikis, Picture quality prediction based on a visual model, IEEE Trans. Commun. 30 (1982) 1679–1692.
[95] Y.-F. Ma, X.-S. Hua, L. Lu, H.-J. Zhang, A generic framework of user attention model and its application in video summarization, IEEE Trans. Multimedia 7 (5) (2005) 907–919.
[96] J. Maisonneuve, M. Deschanel, J. Heiles, W. Li, H. Liu, R. Sharpe, Y. Wu, An overview of IPTV standards development, IEEE Trans. Broadcasting 55 (2) (2009) 315–328.
[97] J. Mannos, D. Sakrison, The effects of a visual fidelity criterion of the encoding of images, IEEE Trans. Inform. Theory 20 (4) (1974) 525–536.
[98] X. Marichal, W.-Y. Ma, H. Zhang, Blur determination in the compressed domain using DCT information, Proc. IEEE Int. Conf. Image Process. 2 (1999) 386–390.
[99] D. Marr, Vision, W.H. Freeman and Company, San Francisco, 1982.
[100] P. Marziliano, F. Dufaux, S. Winkler, T. Ebrahimi, A no-reference perceptual blur metric, in: Proceedings of IEEE International Conference on Image Processing (ICIP '02), September 2002, vol. 3, pp. 57–60.
[101] P. Marziliano, S. Winkler, F. Dufaux, T. Ebrahimi, Perceptual blur and ringing metrics: application to jpeg2000, Signal Process.: Image Commun. 19 (2004) 163–172.
[102] M.A. Masry, S.S. Hemami, Y. Sermadevi, A scalable wavelet-based video distortion metric and applications, IEEE Trans. Circuit Syst. Video Technol. 16 (2) (2006).
[103] L.M.J. Meesters, W.A. IJsselsteijn, P.J.H. Seuntiens, A survey of perceptual evaluations and requirements of three-dimensional TV, IEEE Trans. Circuits Syst. Video Technol. 14 (3) (2004) 381–391.
[104] T. Mei, X.-S. Hua, C.-Z. Zhu, H.-Q. Zhou, S. Li, Home video visual quality assessment with spatiotemporal factors, IEEE Trans. Circuits Systems Video Technol. 17 (6) (2007) 699–706.
[105] O. Le Meur, A. Ninassi, P. Le Callet, D. Barba, Do video coding impairments disturb the visual attention deployment?, Signal Process: Image Commun. (June) (2010).
[106] S.K. Mitra, J. Pearson, J. Caviedes (Guest Editors), Special issue on objective video quality metrics, Signal Process.: Image Commun. 19(2) (2004).
[107] M. Miyahara, K. Kotani, V.R. Algazi, Objective picture quality scale (PQS) for image coding, IEEE Trans. Commun. 46 (9) (1998) 1215–1225.
[108] 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.
[109] M. Montenovo, A. Perot, M. Carli, P. Cicchetti, A. Neri. Objective quality evaluation of video services, in: Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, January 2006.
[110] A.K. Moorthy, Alan Conrad Bovik, Visual importance pooling for image quality assessment, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 193–201.
[111] K.T. Mullen, The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings, J. Physiol. 359 (1985) 381–400.
[112] M. Narwaria, W. Lin, Scalable image quality assessment based on structural vectors. in: Proceedings of IEEE Workshop on Multimedia Signal Processing (MMSP), 2009.
[113] V. Navalpakkam, L. Itti, Top-down attention selection is fine-grained, J. Vis. 6 (11) (2006) 1180–1193.
[114] A.N. Netravali, B.G. Haskell, Digital Pictures: Representation and Compression, Plenum, New York, 1988.
[115] A. Ninassi, O.L. Meur, P.L. Callet, Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. IEEE Int. Conf. Image Process., 2007.
[116] A. Ninassi, O.L. Meur, P.L. Callet, D. Barba, Considering temporal variations of spatial visual distortions in video quality assessment, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 253–265.
[117] T. Oelbaum, C. Keimel, K. Diepold, Rule-based no-reference video quality evaluation using additionally coded videos, IEEE J. Sel. Top. Signal Process. 3 (2) (2009) 294–303.
[118] Annals of Telecommunications. Call for paper, special issue on quality of experience and socio-economic issues of network-based services, 2009.
[119] E. Ong, W. Lin, Z. Lu, S. Yao, M. Etoh, Visual distortion assessment with emphasis on spatially transitional regions, IEEE Trans. Circuits Syst. Video Technol. 14 (4) (2004) 559–566.
[120] E. Ong, W. Lin, Z. Lu, S. Yao, X. Yang, L. Jiang. No reference JPEG-2000 image quality metric, in: Proceedings of IEEE International Conference Multimedia and Expo (ICME), 2003, pp. 545–548.
[121] E. Ong, X. Yang, W. Lin, Z. Lu, S. Yao, X. Lin, S. Rahardja, C. Boon, Perceptual quality and objective quality measurements of compressed videos, J. Vis. Commun. Image Representation 17 (4) (2006) 717–737.
[122] Y.-F. Ou, T. Liu, Z. Zhao, Y. Wang, Modeling the impact of frame rate on perceptual quality of video, in: Proceedings of the IEEE International Conference Image Processing (ICIP), 2008, pp. 689–692.
[123] T.N. Pappas, R.J. Safranek, Perceptual criteria for image quality evaluation, in: A.C. Bovik (Ed.), Handbook of Image and Video Processing, Academic Press, New York, 2000.
[124] J.R. Parker, Algorithms for Image Processing and Computer Vision, John Wiley & Sons, 1996.
[125] H.E. Pashler, The Psychology of Attention, MIT Press, 1998.
[126] R. Pastrana-Vidal, J. Gicquel, Automative quality assessment of video fludity impairments using a no-reference metric, in: Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, January 2006.
[127] R. Pastrana-Vidal, J. Gicquel, C. Colomes, H. Cherifi, Sporadic frame dropping impact on quality perception, Proc. SPIE Int. Soc. Opt. Eng. 5292 (1) (2004).
[128] E. Peli, Contrast in complex images, J. Opt. Soc. Am. A 7 (1990) 2032–2040.

[129] G.C. Phillips, H.R. Wilson, Orientation bandwidths of spatial mechanisms measured by masking, J. Opt. Soc. Am. A 1 (1984) 226–232.
[130] A.B. Poirson, B.A. Wandell, Pattern-color separable pathways predict sensitivity to simple colored patterns, Vis. Res. 36 (4) (1996) 515–526.
[131] N. Ponomarenko, M. Carli, V. Lukin, K. Egiazarian, J. Astola, F. Battisti, Color image database for evaluation of image quality metrics, in: International Workshop on Multimedia Signal Processing, 2008, pp. 403–408.
[132] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, F. Battisti, Tid2008 – a database for evalua-tion of full-reference visual quality assessment metrics, Adv. Mod. Radioelectron. 10 (2009) 30–45. http://www.ponomarenko.info/tid2008.htm.
[133] M.I. Posner, Orienting of attention, Quat. J. Exp. Psychol. 32 (1980) 2–25.
[134] H.-T. Quan, M. Ghanbari, Temporal aspect of perceived quality of mobile video broadcasting, IEEE Trans. Broadcasting 54 (3) (2008) 641651.
[135] G. Ramanarayanan, J. Ferwerda, B. Walter, K. Bala, Visual equivalence: towards a new standard for image fidelity, ACM Trans. Graph. 26 (3) (2007).
[136] M. Ramasubramanian, S.N. Pattanaik, D.P. Greenberg, A perceptual based physical error metric for realistic image synthesis, Comput. Graph. (SIGGRAPH 99 Conf. Proc.) 33 (4) (1999) 73–82.
[137] R.W. Rodieck, The First Steps in Seeing, Sinauer Associates, 1998.
[138] G. Rubino, M. Varela, A new approach for the prediction of end-to-end performance of multimedia streams, in: Proceedings of the first International Conference on the Quantitative Evaluation of Systems, 2004, pp. 110–119.
[139] H.R. Sheikh, A.C. Bovik, G. de Veciana, An information fidelity criterion for image quality assessment using natural scene statistics, IEEE Trans. Image Process. 14 (12) (2005) 2117–2128.
[140] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[141] H.R. Sheikh, Z. Wang, A.C. Bovik, L.K. Cormack, Image and video quality assessment research at live. From <http://live.ece.utexas.edu/research/quality/>.
[142] M.Y. Shen, C.-C.J. Kuo, Review of postprocessing techniques for compression artifact removal, J. Vis. Commun. Image Representation 9 (1) (1998) 2–14.
[143] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, D.J. Heeger, Shiftable multi-scale transforms, IEEE Trans. Info. Theory 38 (2) (1992) 587–607.
[144] N. Suresh, N. Jayant, O. Yang, Mean time between failures: a subjectively meaningful quality metric for consumer video, in: Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, 2006.
[145] S. Winkler, Quality metric design:a closer look, in: Proceedings of the SPIE Human Vision and Electronic Imaging Conference, vol. 3959, 2000, pp. 37–44.
[146] T. Sylvain, A. Florent, Z. Parvez, H. Yuukou, Impact of subjective dataset on the performance of image quality metrics. IEEE Int. Conf. Image Process., 2008.
[147] K.T. Tan, M. Ghanbari, Blockiness detection for MPEG2-coded video, IEEE Sig. Proc. Lett. 7 (8) (2000) 213–215.
[148] K.T. Tan, M. Ghanbari, A multimetric objective picture-quality measurement model for MPEG vide, IEEE Trans. Circuits Syst. Video Technol. 10 (7) (2000) 1208–1213.
[149] D. Tian, G. AlRegib, FQM: a fast quality measure for efficient transmission of textured 3D models, in: Proceedings of the 12th Annual ACM International Conference on Multimedia, 2004.
[150] H.Y. Tong, A.N. Venetsanopoulos. A perceptual model for jpeg applications based on block classification, texture masking, and luminance masking, in: Proceedings of the IEEE International Confernce Image Processing (ICIP), vol. 3, 1998.
[151] X. Tong, D. Heeger, C.V.D.B. Lambrecht, Video quality evaluation using ST-CIELAB, SPIE Proc. Human Vision Visual Process. Digital Display 3644 (1999) 185–196.
[152] A.M. Treisman, G. Gelade, A feature integration theory of attention, Cogn. Psychol. 12 (1) (1980) 97–136.
[153] J. Tumblin, J.A. Ferwerda, Applied perception (guest editors' introduction), IEEE Comput. Graph. Appl. 21 (5) (2001).
[154] C.J. van den Branden Lambrecht, Perceptual models and architectures for video coding applications. Ph.D. dissertation, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1996.
[155] F.L. van Nes, M.A. Bouman, Spatial modulation transfer in the human eye, J. Opt. Soc. Am. 57 (1967) 401–406.
[156] Video Quality Expert Group (VQEG), Final report from the video quality expert group on the validation of objective models of video quality assessment, March 2000. Available: <www.vqeg.org>.
[157] Video Quality Expert Group (VQEG). Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II. Aug. 2003. Available: <www.vqeg.org>.
[158] C.T. Vu, E.C. Larson, D.M. Chandler, Visual fixation pattern when judging image quality: effects of distortion type, amount, and subject experience. in: IEEE Southwest Symp. Image Anal. Interp., 2008, pp. 73–76.
[159] B. Wandell, Foundamentions of Vision, Sinauer Associates, 1995.
[160] Z. Wang, A.C. Bovik, Modern Image Quality Assessment, Morgan & Claypool Publishers, 2006.
[161] Z. Wang, A.C. Bovik, Mean squared error: love it or leave it? – a new look at fidelity measures, IEEE Signal Process. Mag. (Jan.) (2009).
[162] Z. Wang, A.C. Bovik, B.L. Evan, Blind measurement of blocking artifacts in images, Proc. Internat. Conf. Image Process. 3 (2002) 981–984.
[163] Z. Wang, A.C. Bovik, and L. Lu, Why is image quality assessment so difficult? IEEE Int. Conf. Acoust. Speech Signal Process., May 2002.
[164] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, The ssim index for image quality assessment. <http://ece.uwaterloo.ca/~z70wang/research/ssim/>.
[165] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (3) (2002) 81–84.
[166] Z. Wang, A.C. Bovik, L. Lu, Wavelet-based foveated image quality measurement for region of interest image coding, Proc. Int. Conf. Image Process. 2 (2001) 89–92.
[167] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[168] Z. Wang, Q. Li, Video quality assessment using a statistical model of human visual speed perception, J. Opt. Soc. Am. 24 (12) (2007) B61–B69.
[169] Z. Wang, L. Lu, A.C. Bovik, Foveation scalable video coding with automatic fixation selection, IEEE Trans. Image Process. 12 (2003) 1703–1705.
[170] Z. Wang, X. Shang, Spatial pooling strategies for perceptual image quality assessment. IEEE Int. Conf. Image Process., Sept. 2006.
[171] Z. Wang, E.P. Simoncelli, A.C. Bovik, Multi-scale structural similarity for image quality assessment, in: 37 th IEEE Asilomar Conference on Signals, Systems and Computers, Nov. 2003.
[172] A.B. Watson, The cortex transform: rapid computation of simulated neural images, Comput. Vis. Graph. Imaging Proc. 39 (3) (1987) 311–327.
[173] A.B. Watson, DCTune: a technique for visual optimization of DCT quantization matrices for individual images, in: Society for Information Display Digest of Technical Papers, vol. XXIV, 1993, pp. 946–949.
[174] A.B. Watson, J. Hu, J.F. McGowan III, DVQ: a digital video quality metric based on human vision, J. Electron. Imaging 10 (1) (2001) 20–29.
[175] A.B. Watson, J.A. Solomon, Model of visual contrast gain control and pattern masking, J. Opt. Soc. Am. A 14 (9) (1997) 2379–2391.
[176] A.B. Watson, G.Y. Yang, J.A. Solomon, J. Villasenor, Visibility of wavelet quantization noise, IEEE Trans. Image Process. 6 (8) (1997) 1164–1175.
[177] Z. Wei, K.N. Ngan, Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain, IEEE Trans. Circuits Syst. Video Technol. 19 (3) (2009) 337–346.
[178] S. Winkler, A perceptual distortion metric for digital color video, Proc. SPIE 3644 (1999) 175–184.
[179] S. Winkler, Vision models and quality metrics for image processing applications, Lausanne, Switzerland: Ecole Polytecnique Federale De Lausanne (EPFL), Swiss Federal Inst. of Technol., Thesis 2313, December 2000.
[180] S. Winkler, Visual fidelity and perceived quality: towards comprehensive metrics, Proc. SPIE 4299 (2001) 114–125.
[181] S. Winkler, F. Dufaux, Video quality evaluation for internet streaming applications, Proc. SPIE Human Vision Electron. Imaging Conf. 5007 (2003) 104–115.
[182] S. Winkler, F. Dufaux, Video quality evaluation for mobile applications, Proc. SPIE/IS T Visual Commun. Image Process. 5150 (2003) 593–603.
[183] S. Winkler, P. Mohandas, The evolution of video quality measurement: from PSNR to hybrid metrics, IEEE Trans. Broadcasting 54 (3) (2008) 660–668.
[184] S. Winkler, P. Vandergheynst, Computing isotropic local contrast from oriented pyramid decompositions, in: Proceedings of International Conference Image Processing, 1999, pp. 420–424.
[185] S. Wolf, Measuring the end-to-end performance of digital video systems, IEEE Trans. Broadcast. 43 (3) (1997) 320–328.
[186] S. Wolf, M.H. Pinson, Video quality measurement techniques, NTIA Report02-392, June 2002.
[187] R.B. Wolfgang, C.I. Podilchuk, E.J. Delp, Perceptual watermarks for digital images and video, Proc. IEEE 87 (7) (1999) 1108–1126.
[188] H.R. Wu, K.R. Rao (Eds.), Digital Video Image Quality and Perceptual Coding, CRC Press, 2006.
[189] H.R. Wu, M. Yuen, A generalized block-edge impairment metric (GBIM) for video coding, IEEE Signal Process. Lett. 4 (11) (1997) 317–320.
[190] S. Wu, W. Lin, S. Xie, Z. Lu, E. Ong, S. Yao, Blind blur assessment for vision-based applications, J. Vis. Commun. Image Representation 20 (4) (2009) 231–241.
[191] S. Xu, W. Lin, C.-C. Jay Kuo, Mobile video processing for visual saliency map determination video, in: Conference on Applications of Digital Image Processing XXXI, SPIE Optics and Photonics, August 2008.
[192] K.C. Yang, C.C. Guest, K. El-Maleh, P.K. Das, Perceptual temporal quality metric for compressed video, IEEE Trans. Multimedia 9 (2007).
[193] S. Yang, C.-H. Lee, C.-C. Jay Kuo, Optimized mesh and texture multiplexing for progressive textured model transmission, in: Proc. 12th Annual ACM International Conference on Multimedia, 2004.
[194] X. Yang, W. Lin, Z. Lu, E. Ong, S. Yao, Just noticeable distortion model and its applications in video coding, Signal Process.: Image Commun. 20 (7) (2005) 662–680.
[195] X. Yang, W. Lin, Z. Lu, E. Ong, S. Yao, Motion-compensated residue pre-processing in video coding based on just-noticeable-distortion profile, IEEE Trans. Circuits Syst. Video Technol. 15 (6) (2005) 742–750.
[196] J. You, J. Korhonen, A. Perkis, Attention modeling for video quality assessment: Balancing global quality and local quality, in: Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), 2010.
[197] Z. Yu, H.R. Wu, S. Winkler, T. Chen, Vision-model-based impairment metric to evaluate blocking artifacts in digital video, Proc. IEEE 90 (2002) 154–169.
[198] M. Yuen, H.R. Wu, A survey of MC/DPCM/DCT video coding distortions, Signal Process. 70 (3) (1998) 247–278.
[199] G. Zhai, W. Zhang, X. Yang, W. Lin, Y. Xu, No-reference noticeable blockiness estimation in images, Signal Process.: Image Commun. (2008).

[200] L. Zhang, Q. Tian, Multimedia content analysis: model-based approaches vs. data-driven approaches, in: ACM Conference of Multimedia, 2009.
[201] N. Zhang, A.E. Vladar, M.T. Postek, B. Larrabee, A kurtosis-based statistical measure for two-dimensional processes and its application to image sharpness, Proc. Sect. Phys. Eng. Sci. Am. Stat. Soc. (2003) 4730–4736.
[202] X. Zhang, B.A. Wandell, Color image fidelity metrics evaluated using image distortion maps, Signal Process. 70 (3) (1998) 201–214.
[203] X. Zhang, W. Lin, P. Xue, Improved estimation for just-noticeable visual distortion, Signal Process. 85 (4) (2005) 795–808.
[204] X. Zhang, W. Lin, P. Xue, Just-noticeable difference estimation with pixels in images, J. Vis. Commun. Image Representation 19 (1) (2008) 30–41.