Visual Quality Evaluation for Images and Videos

Songnan Li, Lawrence Chun-Man Mak, and King Ngi Ngan

The Chinese University of Hong Kong

1 Introduction

Information is exploding with technology progress. Compared with text and audio, image and video can represent information more vividly, which makes visual quality one of the most important aspects in determining user experience. A good visual quality evaluation method can assist in monitoring the quality of multimedia services and boosting user experience.

Visual quality evaluation plays its role in different stages of the visual information distribution chain, e.g., camera filter design for visual signal acquisition [51], quality monitoring during signal relaying [36], representation at the user site by display [39], printer [19], etc. In addition, the success of visual quality evaluation will provide guidance to a large number of image and video processing algorithms, e.g., compression, watermarking, image fusion, error protection, feature enhancement and detection, restoration, retrieval, graphic illumination, and so on. Due to its fundamental role, works on visual quality metrics can date back to half a century ago, and a vast number of objective quality metrics have been proposed over time.

Since pixel-based metrics, such as Mean Square Error (MSE), Signal-to-Noise Ratio (SNR) and Peak Signal-to-Noise Ratio (PSNR), are simple to calculate and easy to be incorporated into optimization process, they have been widely used in most visual-related products and services. However, it has been well acknowledged that these pixel-based difference measures do not correlate well with the Human Visual System's perception [18]. Distortions perceived by the human being are not always captured by MSE/SNR/PSNR, because these metrics operate on a pixel-by-pixel basis without considering the signal content, the viewing condition and the characteristics of the Human Visual System (HVS). These problems make the design of a better objective quality metric necessary, and progresses in recent vision research provide us with guidance to achieve this goal.

Different from objective quality metrics which work automatically without human intervention, subjective quality evaluation acquires quality judgment from the human observers in an off-line manner, and as a matter of course is considered to be the most accurate approach to measure visual quality. Although subjective quality evaluation is time-consuming and not feasible for on-line manipulation, its role in objective quality metric design is still irreplaceable: the perceptual visual quality derived from subjective evaluation can serve as a benchmark for the performance evaluation of different objective quality assessment algorithms, and can even direct the algorithm design. More and more subjectively-rated image and

W. Lin et al. (Eds.): Multimedia Analysis, Processing & Communications, SCI 346, pp. 497–544.
springerlink.com

video database become publicly available [59, 8, 20, 56, 10], which will speed up the advent of better objective quality metrics.

The goal of this chapter is to review both the objective and subjective visual quality evaluation methods, with Section 2 dedicated to the former and Sections 3 and 4 dedicated to the latter. In Section 2, objective visual quality metrics are reviewed on the basis of two distinct design approaches: HVS modeling approach and engineering based approach. Metrics for images and for videos are presented without clear partition due to their great similarities. Section 3 briefly describes the different aspects of a typical subjective evaluation procedure, while Section 4 presents an application of the subjective quality evaluation carried out recently by the authors: a performance comparison between two video coding standards — H.264 and AVS.

2 Objective Visual Quality Metrics

2.1 Classification

Objective visual quality metrics can be classified into Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) metrics according to the availability of the reference information. For FR metrics, the original image or video sequence is fully available as the reference, and is considered to be of perfect quality. The distorted visual signal is compared against the reference and their similarity is measured so as to determine the quality of the distorted signal. FR metrics can be adopted in many image and video processing algorithms such as compression, watermarking, contrast enhancement, etc. However, in many practical applications, e.g., quality monitoring during transmission or at the user site, it is impossible to access the entire reference. RR and NR can fit in these situations. RR metrics only need partial reference signal. Features are extracted from the reference signal and transmitted in an ancillary channel for comparison against the corresponding features of the distorted signal extracted at the monitoring site. Compared with FR metrics, RR metrics are more flexible (weaker requirement on registration) and less expensive in terms of bandwidth requirement. NR metrics gauge quality without any reference information at all. They are free of registration requirement, and applicable to a wide range of applications. However, although human observers are good at NR quality evaluation, NR metric development turns out to be a difficult task, and only limited successes have been achieved so far.

From the design viewpoint, many classification methods have been proposed in literature: metrics designed by psychophysical approaches and by engineering approaches [89]; error sensitivity and structural sensitivity metrics [77]; bottom-up and top-down metrics [37], etc. Winkler et al. [90] proposed a comprehensive classification method which groups metrics into three categories: Data metrics, Picture metrics, Packet- and Bitstream-based metrics. Data metrics measure the fidelity of the signal without considering its content. The representatives are MSE/SNR/PSNR and their close families [15]. On the other hand, Picture metrics treat the visual data as the visual information that it contains. They may take the viewing conditions (viewing distance, ambient illumination, display properties, etc.), the HVS characteristics, and the signal content into consideration. Picture metrics can be further distinguished into two groups: metrics designed by vision modeling approaches (HVS-model-based metrics) and metrics designed by engineering approaches (engineering-based metrics). In this section, objective visual quality metrics will be separated into these two categories for a more detailed introduction. Packet- and Bitstream-based metrics [66, 30] are used in applications like internet streaming or IPTV. They measure the impact of network losses on visual signal quality. Different from traditional methods like bit error rate or packet loss rate, Packet- and Bitstream-based metrics distinguish the importance of the lost information to visual quality by checking the packet header and the encoded bitstream.

2.2 HVS-Model-Based Metrics

HVS-model-based metrics incorporate characteristics of the HVS which are obtained from psychophysical experiments of vision research. Although anatomy provides us with detailed physiological evidences about the front-end of the HVS (optics, retina, LGN etc.), a thorough understanding of the latter stages of the visual pathway (visual cortex, etc.) in charge of higher-level perception is still unachievable, which makes the construction of a complete physiological HVS model impossible. Consequently HVS models used by visual quality metrics are mostly based on psychophysical studies and only account for lower-level perception. Physiological and psychophysical factors typically incorporated into the HVS model include color perception, luminance adaptation, multi-channel decomposition, contrast sensitivity function (CSF), masking, pooling, etc., as shown in Fig. 1.

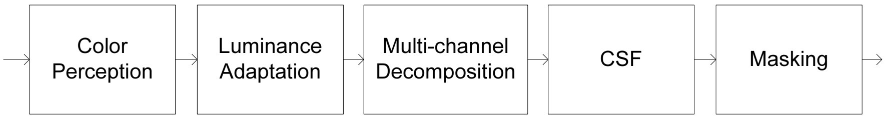


Fig. 1. A typical HVS model

Each of the perceptual factors listed above will be explained in detail in this section. But before that, it should be noted that visual quality metrics will also model different transformations of the visual signal, before it is perceived by the human eyes. For example, at the very beginning, visual signal usually is represented by pixel values. When displayed on a Cathode Ray Tube (CRT) or Liquid Crystal Display (LCD) monitor, these pixel values will be transformed into light intensities, which have a non-linear relationship with their corresponding pixel values. This non-linear relationship is determined by the gamma value of the display[1] and may be slightly different for the R, G, and B channels which will be introduced below.

[1] An excellent explanation on “gamma” can be found in [57].

2.2.1 Components of the HVS Model

- Color perception

R, G, and B stands for the three primary colors Red, Green, and Blue, which can be combined to create most of the visible colors. RGB color space is commonly employed by camera sensors and in computer graphics. There are physiological evidences that justify the use of this color space. When lights pass through the optics of the eye and arrive at the retina, they will be sampled and converted by two different types of photoreceptors: rods and cones. Rods and cones are responsible for vision at low light level and high light level, respectively. In addition, cones contribute in color perception. There are three types of cones: S cones, M cones, and L cones, which are sensitive to short, median, and long wavelengths, respectively, as shown in Fig. 2. They are often depicted as blue, green, and red receptors, although as a matter of fact they do not accurately correspond to these colors.

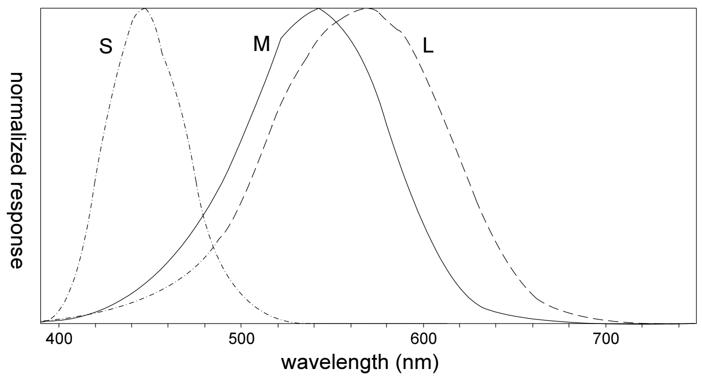


Fig. 2. Normalized response spectra of human cones, S, M, and L types, with wavelength given in nanometers [85]

Responses of the cones need to be further processed at a higher stage of the HVS for the purpose of "decorrelation". RGB channels are highly correlated with each other: by viewing the R, G, and B channels of a given image independently, you can find that each channel contains the entire image. Their correlations can also be seen from Fig. 2 where the three types of cones overlap in their sensitivities to the light wavelengths. Possibly for coding efficiency, HVS records the differences between the responses of the cones, rather than each type of cone's individual response. This is referred to as the *opponent processing theory* of color perception. According to this theory, there are three opponent channels: Black-White (B-W) channel, Red-Green (R-G) channel, and Blue-Yellow (B-Y) channel. Neural response to one color of an opponent channel is antagonistic to the other color in the same channel. This explains a number of perceptual phenomena, e.g., you can perceive a reddish yellow (orange) but you never perceive a reddish green. Physiological evidences support the existence of opponent channels: bipolar cells and ganglion cells of the retina may be involved in opponent color processing [88].

Besides the above mentioned RGB and B-W/R-G/B-Y, many other color spaces are developed for different purposes, e.g., CIELAB, HSL, YIQ, and so on. Most of these share the common characteristic that they treat visual information as a combination of luminance and chrominance components. Chrominance component is represented by two descriptors which usually have different physical meanings for different color spaces, such as a* (red-green balance) and b* (green-blue balance) for CIE L* a* b*, H (hue) and S (saturation) for HSL, and I (blue-green-orange balance) and Q (yellow-green-magenta balance) for YIQ. Luminance component on the other hand is more or less the same which is to simulate the B-W opponent channel of the HVS. Since B-W channel carries most of the visual information, many visual quality metrics only make use of luminance information for quality assessment. According to the performance comparison of different color spaces in visual quality metrics [86], there is only a slight performance loss due to abandoning the use of chrominance components, and on the other hand, the computational complexity can be reduced by a great amount.

- Luminance adaptation

It is well known that our perception is sensitive to luminance contrast rather than the luminance intensity. Given an image with a uniform background luminance *I* and a square at the center with a different luminance *l* + *dl*, if *dl* is the threshold value at which the central square can be distinguished from the background, then according to the Weber's law the ratio of *dl* divided by *l* is a constant for a wide range of luminance *l*. This implies that our sensitivity to luminance variation is dependent to the local mean luminance. In other words, the local mean luminance masks the luminance variation: the higher the local mean luminance, the stronger is the masking effect. That is the reason why the term "luminance masking" is preferred by some authors rather than "luminance adaptation". In practical implementations, the luminance of the original signal may serve as the masker to mask the luminance variation due to distortion.

If the ratio of *dl* divided by *l* is used to represent the differential change of the luminance contrast *dc,*

$$dc = \frac{K \times dl}{l}, \tag{1}$$

where *K* is a constant, then the following equation

$$c = K \times \log l + C, \tag{2}$$

can be obtained, which describes the relationship between the luminance contrast *c* and the luminance intensity *l*. Constants *K* and *C* can be determined experimentally. According to equation (1.2), the perceived luminance contrast is a non-linear function of the luminance intensity. For visual quality metrics, one approach to model luminance adaptation is by applying such a logarithmic function or other alternatives e.g., cube root, or square root function [13] before multi-channel decomposition.

Another way to simulate luminance adaptation is to convert the luminance intensity to the luminance contrast right after (rather than before) multi-channel decomposition, as shown in Fig. 3. Peli E. in [54] defines a local band limited contrast measure for complex images, which assigns a local contrast at every point of an image and at every frequency channel. Multi-channel decomposition creates a series of bandpass-filtered images and lowpass-filtered images. For bandpass-filtered image $a_k(x,y)$ and its corresponding lowpass-filtered image $l_{k-1}(x,y)$, the contrast at frequency subband k is expressed as a two-dimensional array $c_k(x,y)$:

$$c_k(x,y) = \frac{a_k(x,y)}{l_{k-1}(x,y)}. \tag{3}$$

This is one of the many attempts to define luminance contrast in complex images, and it has been adopted in visual quality metrics such as [39, 44] with some modifications.

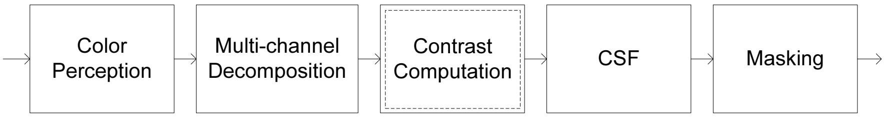


Fig. 3. HVS model using contrast computation

Other approaches to implement luminance adaptation can be found in Just Noticeable Difference (JND) modeling [35], which is closely related to visual quality assessment. In spatial domain JND, the visibility threshold of luminance variation can be obtained as a function of the background luminance, as shown in Fig. 4. In frequency domain JND, luminance adaptation effect is often implemented as a modification to the baseline threshold derived from the contrast sensitivity function, which will be introduced later.

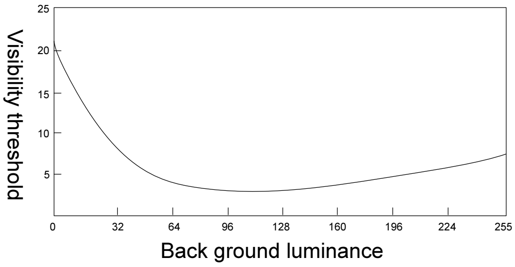


Fig. 4. Luminance adaption: visibility threshold versus background luminance [12]

- Multi-channel decomposition

Instead of employing just one channel as in the early works [41, 40], multi-channel decomposition has been widely used for HVS modeling nowadays. Multi-channel decomposition is justified by the discovery of the spatial frequency selectivity and orientation selectivity of the simple cells in the primary visual cortex. It can also be successfully used to explain empirical data from masking experiments.

Both temporal and spatial multi-channel decomposition mechanisms of the HVS have been investigated over time, with more efforts paid to the spatial one. For spatial multi-channel decomposition, most studies suggest that there exists several octave spacing radial frequency channels, each of which is further tuned by orientations with roughly 30 degree spacing [50]. Fig. 5 shows a typical decomposition scheme which is employed in [84] and is generated by cortex transform [79]. Many other decomposition algorithms serving this purpose exist, e.g., steerable pyramid transform [64], QMF (Quadrature Mirror Filters) transform [65], wavelet transform [31], DCT transform [80], etc. Some of these aim at accurately modeling the decomposition mechanism, while others are used due to their suitability for particular applications, e.g., compression [80]. A detailed comparison of these decomposition algorithms can be found in [85].

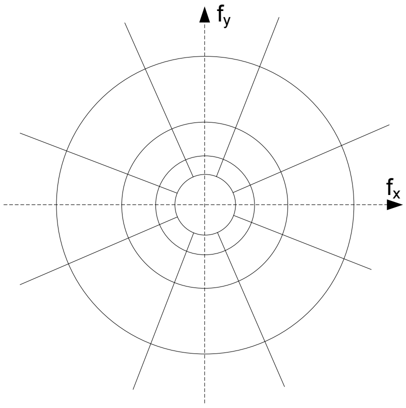


Fig. 5. Illustration of the partitioning of the spatial frequency plane by the steerable pyramid transform [31]

For temporal decomposition, it is generally believed that there exist two channels: one low-pass channel, namely sustained channel, and one band-pass channel, namely transient channel. Since most visual detailed information is carried in sustained channel, HVS models employed by some video quality metrics like those in [44, 94] only use a single low pass temporal filter to isolate the sustained channel, while the transient channel is disregarded. Temporal filters can be implemented as either Finite Impulse Response (FIR) filters [44] or Infinite Impulse Response (IIR) filters [32], either before [44] or after spatial decomposition [83].

- Contrast sensitivity function

Contrast sensitivity is the inverse of the contrast threshold – the minimum contrast value for an observer to detect a stimulus. These contrast thresholds are derived from psychophysical experiments using simple stimuli, like sine-wave gratings or Gabor patches. In these experiments, the stimulus is presented to an observer with its contrast increasing gradually. The contrast threshold is determined at the point where the observer can just detect the stimulus.

It has been proved by many psychophysical experiments that the HVS's contrast sensitivity depends on the characteristics of the visual stimulus: its spatial frequency, temporal frequency, color, and orientation, etc. Contrast sensitivity function (CSF) can be used to describe these dependences. Fig. 6 shows a typical CSF quantifying the dependency of contrast sensitivity on spatial frequency. The decreasing sensitivity for higher spatial frequency is a very important HVS property which has been widely applied in image and video compression: because the HVS is not sensitive to signals with higher spatial frequencies, larger quantization can be applied to them without introducing visible distortions. On the other hand, the decreasing sensitivity for lower frequencies is less crucial, and in many cases it has been neglected intentionally resulting in a low-pass version of the CSF [1]. CSF is more complex when the influences of other factors like temporal frequency or color are considered in conjunction with the spatial frequency [87].

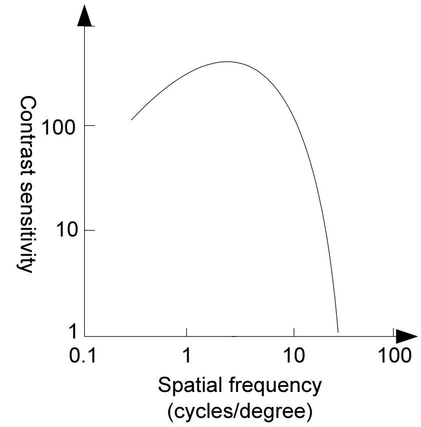


Fig. 6. A typical spatial CSF function

It should be noted that spatial frequency of a visual signal is a function of viewing distance. When the observer moves closer to the display or away from it, the spatial frequency of the visual signal will be changed. As a result, in order to make use of the CSF in the correct way, either viewing distance needs to be taken as a parameter for spatial frequency calculation, or the viewing distance should be fixed, e.g., 6 times image height for SDTV and 3 times image height for HDTV.

To incorporate it into the HVS model, CSF can be implemented either before or after the multi-channel decomposition. In the former case, CSF is implemented as linear filters with frequency response close to the CSF's. In the latter case, since visual signal has already been decomposed into different frequencies, CSF filtering can be approximated by multiplying each subband with a proper value. In JND models, CSF is often used to obtain the baseline contrast threshold, which will be further adjusted to account for luminance adaptation and the masking effect introduced below.

- Masking

Masking effect refers to the visibility threshold elevation of a target signal (the maskee) caused by the presence of a masker signal. It can be further divided into spatial masking and temporal masking.

In most spatial masking experiments, the target and masker stimuli are sinewaves or Gabor patches. The target stimulus is superposed onto the masker stimuli, and contrast threshold of the target stimulus are recorded, together with the masker information, including its contrast, spatial frequency, orientation, phase, etc. Many of these experiments verify that the threshold contrast of the target depends on the masking contrast, and also the other characteristics of the masker. Generally higher masking contrast and larger similarity between the masker and the target in their spatial frequencies, orientations, and phases will lead to higher masking effect, which is known as the contrast masking effect. Fig. 7 shows part of the contrast masking data from [33] describing the relationship between threshold contrast and the masking contrast. With the increase of the masking contrast, the threshold contrast reduces first and then increases, consistently for the three viewers. The threshold contrast reduction, referred to as *facilitation*, is often neglected in spatial masking models, resulting in a monotone increasing curve similar to Fig 8.

One way to implement contrast masking is to make the original visual content act as the masker and the distortion as the target [13, 18]. In this case, usually contrast masking is assumed to occur only between stimuli located in the same channel (intra-channel masking) characterized by its unique combination of spatial frequency, orientation, phase, etc. The output of contrast masking function will

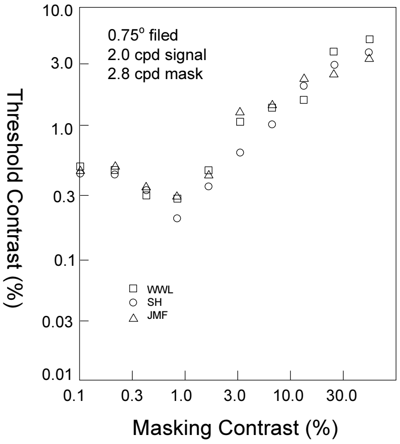


Fig. 7 Experimental data for contrast masking from [33]. WWL, SH, JMF represent three subjects.

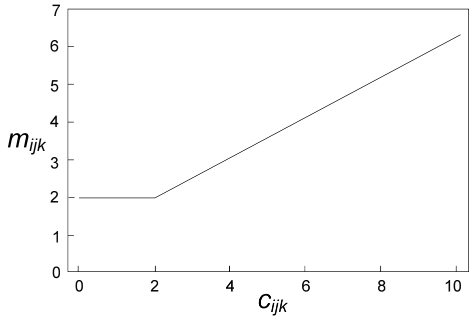


Fig. 8. Contrast masking function from [80], describing the masked threshold m_{ijk} as a function of DCT coefficient c_{ijk}.

be multiplied to the CSF baseline threshold to account for the contrast threshold elevation caused by contrast masking. In another type of contrast masking model, original visual content will no longer serve as the masker. Instead, the original and the distorted signals pass through the masking model separately. The outputs of the masking model simulate the response of cortical visual neurons to these visual contents, which will be compared directly in the next stage (pooling). The response of visual neuron can be modeled either by a saturating nonlinear transducer function [33] or by a contrast gain control process [65, 17, 81]. As an example of the latter case, Watson and Solomon integrate a variety of channel interactions into their model [81] (inter-channel masking), which is achieved by division of the excitatory signal from each neuron of one channel by an inhibitory signal that is a linear combination of responses of neurons within neighboring channels.

The sine-wave gratings or Gabor patches used to derive the above contrast masking models are oversimplifications with respect to natural images. Efforts have been made towards masking measurements with more realistic targets, e.g., quantization errors [47], and maskers, e.g., random noises, bandpass noises, and even natural images [82]. To emphasize their differences with the traditional contrast masking, different terms were used, such as noise masking, texture masking, or entropy masking, etc.

Compared with spatial masking, temporal masking has received less attention and is of less variety. In most of its implementations in video quality assessment, temporal masking strength is modeled as a function of temporal discontinuity in intensity: the higher the inter-frame difference, the stronger is the temporal masking effect. Particularly, the masking abilities of scene cut have been investigated in many experiments, with both of its "forward masking" and "backward masking" effects identified [2].

- Pooling

In vision system, pooling refers to the process of integrating information of different channels, which is believed to happen at the latter stages of the visual pathway. In visual quality assessment, pooling is used to term the error summation process which combines errors measured in different channels into a quality map or a single quality score. For image quality assessment, most approaches perform error summation across frequency and orientation channels first to produce a 2-D quality map, and then perform it across spaces to obtain a single score indicating the quality of the entire image. For video quality assessment, one more step is performed to combine quality scores for frames into a quality score for the video sequence.

Minkowski summation, as shown below, is the most popular approach to implement the error pooling process:

$$E = (\sum_i |e_i|^\beta)^{\frac{1}{\beta}} . \tag{4}$$

In the above equation, e_i represents error measured at channel/position/frame i; E is the integrated error; and β is the summation parameter which is assigned a value between 2 to 5 in most works. With a higher value of β, E will depend more on the larger e_is, which is consistent with the reality that visual quality is mostly determined by the stronger distortions.

In image quality metrics, higher-level characteristics of the vision system can be applied into quality metric by using spatial weighted pooling [74]:

$$E = \frac{\sum_i (w_i \times |e_i|)}{\sum_i w_i} , \tag{5}$$

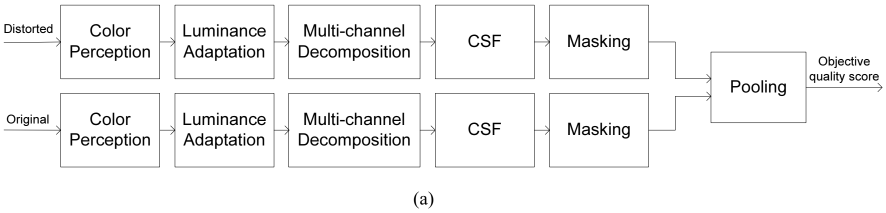
where w_i is the weight given to the error e_i at spatial position i. It represents the significance of e_i to the visual quality of the image, and can be determined by cognitive behaviors of the vision system, such as visual attention [46]. In video quality metrics, temporal pooling can also integrate cognitive factors, such as the asymmetric behavior with respect to quality changes from bad to good and reverse [63].

2.2.2 Frameworks

Fig. 9 shows two different frameworks of HVS-model-based quality metrics. It should be noted that most HVS-model-based metrics are FR metrics, so their inputs include both the original and distorted visual signals.

In the first framework shown in Fig. 9 (a), the original and distorted signals pass through each of the HVS components separately, where the representations of the visual signals are changed sequentially simulating the processing of the HVS, until their differences are calculated and summed in the error pooling process. The summed error can be further converted to detection probability by using the probability summation rule as in [13], or converted to a quality score by using nonlinear regression as in [60].

In the second framework, JNDs need to be calculated in the domain where the original and the distorted signals are compared. JND is the short form for Just Noticeable Distortion which refers to the maximum distortion under the perceivable level. As shown in Fig. 9 (b), generally JND will be calculated as the product of



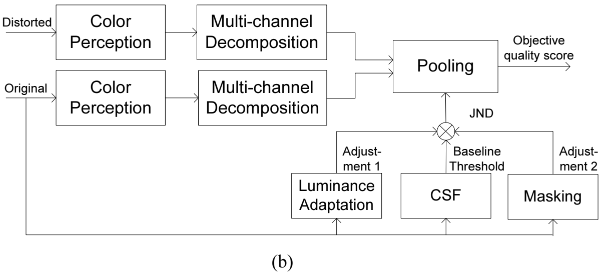


Fig. 9. Two frameworks of HVS-model-based quality metrics

the baseline contrast threshold obtained from the CSF and some adjustments obtained from luminance adaptation and various masking effects, such as contrast masking, temporal masking, and so on, which have been introduced in the last section. The errors of the distorted signal will be normalized (divided) by their corresponding JND values before they are combined in the error pooling process.

Frameworks different from the above mentioned exist, but often have slight differences. For example, the framework of DVQ [83] may be simplified as shown in Fig. 10, where the local contrast computation is added after multi-channel decomposition, and the luminance adaptation is removed compared with Fig. 9 (b).

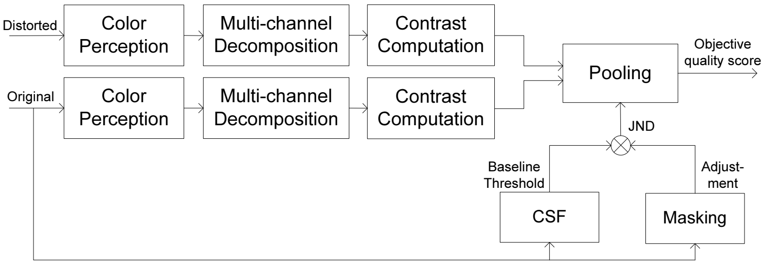


Fig. 10. DVQ's framework

2.2.3 Shortcomings

As stated above, the foundation of the HVS model mostly grounds on psychophysical experiments which use simple visual stimuli and target at contrast threshold evaluation. This leads to two major problems to visual quality metric that employs

HVS model as its kernel. Firstly, a natural image usually is a superposition of a large number of simple stimuli. Their interactions cannot be fully described by a model which is based on experimental data of only one or two simple stimuli. Secondly, there is no justification for the use of experimental data of contrast threshold evaluation in gauging visual quality, especially for images with supra-threshold distortions. For visual quality evaluation, it should be helpful to take higher-level behaviors of the vision system into consideration, but in most HVS models which target at contrast threshold prediction, only low-level perceptual factors are simulated. Besides the problems mentioned above, the high computational complexity is another disadvantage of the HVS-model-based quality metrics especially for video quality assessment.

2.3 Engineering-Based Metrics

To overcome these shortcomings brought by the vision model, recently many new visual quality metrics were designed by engineering approaches. Instead of founding on accurate experimental data from subjective viewing tests, these engineering-based quality metrics are based on (a) assumptions about, e.g., visual features that are closely related to visual quality; (b) prior knowledge about, e.g., the distortion properties or the statistics of the natural scenes. Since these features and prior knowledge are considered to be higher-level perceptual factors compared with lower-level ones used in the vision model, engineering-based quality metrics are also referred to as top-down quality metrics, and are considered to have the potential to better deal with supra-threshold distortions. In [27], International Telecommunication Union (ITU) recommends four video quality metrics after VQEG's FRTV Phase II tests [67], all of which belong to this category. This may serve as the evidence for the promising future of engineering-based quality metrics.

Unlike HVS-model-based metrics, most of which are FR, there are also many RR and NR engineering-based quality metrics. Since FR and RR metrics share great similarities in their processing routines, they will be reviewed together below, followed by the introduction of NR metrics.

2.3.1 FR and RR Metrics

Viewed conceptually, most engineering-based FR and RR quality metrics consist of three processing steps: (a) feature extraction; (b) feature comparison; and (c) quality determination, as shown in Fig. 11. The extracted features characterize the quality metric and determine its performance. These features may be either scalar ones or vectors, and their differences can be obtained in various ways, such as the absolute distance, the Euclidean distance, etc. In most quality metrics, the feature differences can quantify video distortions locally, by using one or several 2-D distortion maps. And these distortion maps will be combined together to generate a single quality score. Two methods are commonly used for the last step: in Fig. 12 (a), spatial/temporal pooling are performed first to generate several distortion factors each representing the intensity of a particular distortion type, and then these distortion factors will be combined at the end to generate a signal quality score for the entire image or video sequence; in Fig. 12 (b), distortions of different types are combined

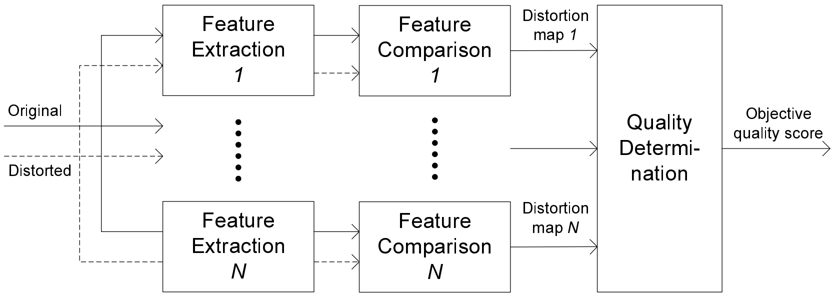


Fig. 11. A conceptual framework for FR and RR metrics

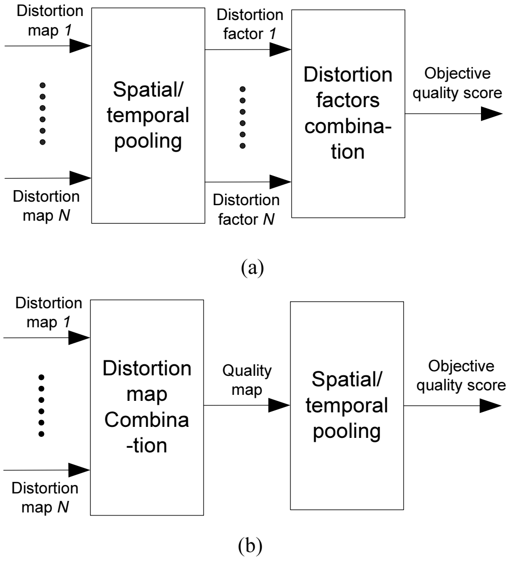


Fig. 12. Two methods for quality determination

together first to generate a quality map, and then spatial pooling is performed on this quality map to compute a single quality score.

In this section, four classic engineering-based image or video quality metrics will be briefly explained, with focus on their implementations of the three processing steps.

- Picture Quality Scale (PQS)

PQS [45] is a hybrid image quality metric employing both the HVS model and the engineering design approaches. Among the five distortion factors measured, three of them are obtained basically by using HVS models. Involved perceptual factors include luminance adaptation, CSF, and texture masking. The other two engineering-based distortion factors measure blockiness and error correlations.

To fit in the three processing steps introduced above, in PQS, non-linear mapped luminance values (to account for the luminance adaptation effect) are used as features, and feature comparison is implemented by direct subtraction. These feature differences are further processed by the CSF and by using prior knowledge about the locations of the distortions to produce two distortion maps measuring blockiness and local error correlations, respectively. In the last step, spatial pooling is performed separately on each of the two distortion maps, generating two engineering-based distortion factors. Together with the three HVS-model-based distortion factors, they are de-correlated by singular value decomposition and linearly combined to generate the PQS quality score.

Compared with the metrics that we will introduce below, the features and the comparison method used in PQS are very simple. In fact, it is not the features but the prior knowledge used, i.e., the locations of the distortions, that represents the idea of the engineering design approach.

- Video Quality Model (VQM)

VQM [55] is one of the best proponents of the VQEG FRTV Phase II tests [67]. For a video sequence, VQM generates seven distortion factors to measure the perceptual effects of a wide range of impairments, such as blurring, blockiness, jerky motion, noise and error blocks, etc. Viewed conceptually, VQM's distortion factors are all calculated in the same steps. Firstly, the video streams are divided into 3D Spatial-Temporal (S-T) sub-regions typically sized by 8 pixel × 8 lines × 0.2 second; then feature values will be extracted from each of these 3D S-T regions by using, e.g., statistics (mean, standard deviation, etc.) of the gradients obtained by a 13-coefficient spatial filter, and these feature values will be clipped to prevent them from measuring unperceivable distortions; Finally these feature values will be compared and their differences will be combined together for quality prediction.

Three feature comparison methods used by VQM are Euclidean distance, ratio comparison, and log comparison, as shown by equations (1.6), (1.7), (1.8), respectively, where f_o and f_{o2} are original feature values, and f_p and f_{p2} are the corresponding processed feature values. Euclidean distance is applied to 2D features (C_B-C_R vectors), and the other two are applied to scalar features (luminance values). The feature differences are integrated by spatial and temporal pooling first, generating seven distortion factors, which are then linearly combined at the last to yield the final VQM quality score.

$$p = \sqrt{(f_o - f_p)^2 + (f_{o2} - f_{p2})^2}\ . \tag{6}$$

$$p = (f_p - f_o) / f_o\ . \tag{7}$$

$$p = \log_{10}(\frac{f_p}{f_o})\ . \tag{8}$$

- Structural Similarity Index (SSIM)

SSIM was first proposed in [77, 73] as a FR image quality metric. It has been extended to video quality metrics [78, 58] and applied to numerous vision-related applications.

The basic assumption of SSIM is that the HVS is highly adapted to extract structural information from the viewing field. SSIM divides the input images into overlapping image patches (e.g., 8×8 pixel blocks), and from each image patch three features were extracted. The first two scalar features are the mean μ and standard deviation σ of the luminance values of the image patch. The third feature can be regarded as a vector with its elements being luminance values normalized by σ. The extracted features from the reference image patch x and the distorted image patch y are compared by using the following equations[2]:

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{{\mu_x}^2 + {\mu_y}^2 + C_1}, \tag{9}$$

$$c(x,y) = \frac{2\sigma_x\sigma_y + C_2}{{\sigma_x}^2 + {\sigma_y}^2 + C_2}, \tag{10}$$

$$s(x,y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3}, \tag{11}$$

where $l(x,y)$, $c(x,y)$ and $s(x,y)$ are termed as the luminance similarity, the contrast similarity and the structural similarity, respectively, and constants C_1, C_2 and C_3 are used to avoid division by zero. Different from PQS and VQM, SSIM adopts the quality determination method shown in Fig. 1.12 (b): the three similarity factors were combined first before the spatial pooling was performed. This makes SSIM being able to produce a spatially varying quality map which indicates quality variations across the image.

- Visual Information Fidelity (VIF)

VIF [60] is a FR image quality metric grounding on the assumption that visual quality is related to the amount of information that the HVS can extract from an image. Briefly, VIF works in the wavelet domain and uses three models to model the original natural image, the distortions, and the HVS, respectively. As shown in Fig. 13, C, D, E and F are the modeling results for the original image, the distorted image, the perceived original image, and the perceived distorted image, respectively. Each of them is represented by a set of random fields, in such a way that the mutual information between any two of them is measurable. The mutual information between C and E, $I(C,E)$, quantifies the information that the HVS can extract from the original image, whereas the mutual information between C and F,

[2] According to the explanations about the features used by SSIM, equation (1.11) actually involves both extraction and comparison of the third feature.

I(*C*,*F*), quantifies the information that can be extracted from the distorted image. The VIF quality score is given by equation (12). Viewed conceptually, the only feature used by VIF is the visual information. For implementation, the mutual visual information *I*(*C*,*E*) and *I*(*C*,*F*) are measured locally within each wavelet subband. Pooling over spaces and wavelet subbands is performed to obtain the numerator and the denominator of equation (12).

$$s_{VIF} = \frac{I(C,F)}{I(C,E)}. \tag{12}$$

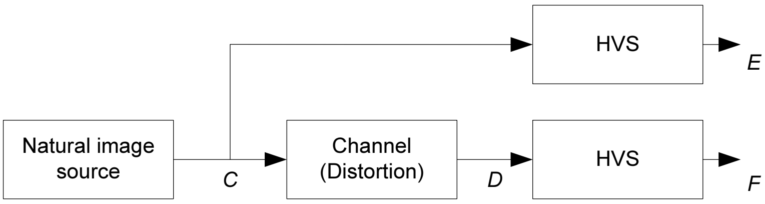


Fig. 13 Block-diagram of VIF from [60]

2.3.2 NR Metrics

As mentioned before, although human observers can easily assess quality without reference information, NR metric design is by no means an easy task. By contrary, NR quality assessment is so difficult that their applications are often limited to the cases where the prior knowledge about the distortion type is available. The distortion types that NR metrics often deal with include blocking, blurring, ringing, and jerky/jitter motion, etc., caused by signal acquisition, compression, or transmission. Another prior knowledge used by some NR metrics asserts that natural images belong to a small set in the space of all possible signals, and they can be described by statistical models fairly well. However, distortions may cause the modeling's inaccuracy, which in turn can be used as an indication of the distortion strength.

- Blocking

Blocking artifacts arise from block-based compression algorithms, like JPEG, MPEG1/2, H.26x, etc., running at low bit rates. Due to the fixed block structure commonly used, blockiness often appears as periodic horizontal and vertical edges whose positions are fixed on block boundaries. Most NR quality metrics detect and quantify blockiness in the spatial domain, by directly measuring differences of the boundary pixels. And in some NR metrics [91, 5, 95] these boundary differences are further adjusted to account for the luminance and texture masking effects. On the other hand, a handful of NR metrics detect blockiness in frequency domain. For example in [76], 1-D FFT is applied to the differences of adjacent rows or columns of the image. Periodic peaks caused by blockiness are identified in the resultant power spectrum and are used to assess blockiness. In [71], eight sub-images are constructed from the distorted image. Their similarities are measured in the Fourier transform domain to obtain two values describing inter-block similarity and intra-block similarity, respectively, with the former being closely related to the blockiness strength. The inter-block similarity is normalized by the intra-block similarity to yield the final blockiness measure.

- Blurring

Blurring artifacts can be caused by camera out-of-focus, fast camera motion, data compression, and so on. Unlike blockiness which can be easily localized, blurring is more image content dependent. Since blur affects edges most conspicuously, many spatial-domain blur metrics make use of this, by detecting step edges and then estimating the edge spread in the perpendicular direction [43, 49]. Also there are blur metrics that work in the frequency domain. For example in [42], the authors proposed a blur determination technique based on histograms of non-zero DCT coefficients of the JPEG or MPEG compressed signals. In [9], a blur metric was developed based on local frequency spectrum measurement, i.e., 2D kurtosis, around the edge regions.

- Ringing

Ringing is another common compression artifact caused by high frequency quantization. Analogous to the Gibbs phenomenon, ringing artifact manifests itself in the form of spurious oscillations, and appears most prominently in the smooth regions around the edges. Compared with blocking metrics or blurring metrics, ringing NR metrics are less investigated, and most existing ones [48, 11] follow a similar conceptual routine: identifying strong edges first and then detecting activities around them as the indication of the ringing artifact intensity. In an alternative approach [34], ringing strength is quantified by measuring the noise spectrum that is filtered out by anisotropic diffusion.

It should be noted that since ringing artifacts often coexist with other compression artifacts, e.g., blockiness or blur, and appear to be less annoying comparatively, quality metrics rarely aim at quantifying ringing artifact only. Instead most NR metrics that measure ringing artifacts will also consider other distortion types, and will balance their contributions to the final quality prediction.

- Jerky/jitter motion

In videos, besides the above mentioned spatial artifacts, temporal impairments like jerky or jitter motion may arise. Jerkiness is often used to describe the “regular” frame freezing followed by a discontinuous motion. Generally it is caused by consistent frame dropping on the encoder side (transcoder) serving as a bit rate or terminal capability adaptation strategy. On the other hand, jitter often describes “irregular” frame dropping due to packet loss during signal transmission. The influences of these motion fluidity impairments to visual quality have been investigated in [52, 38] by subjective viewing tests. Usually NR metrics quantifying motion impairments [53, 92] will consider the following factors: the frame dropping duration (the shorter the better for visual quality), the frame dropping density (with the same amount of frame loss, the more scattered the better), and the motion activity (the lower the better). Compared with spatial distortion NR metrics, temporal distortion

NR metrics can provide better quality prediction that is more consistent with the judgments of the human observers.

- Statistics of natural images

Different from texts, cartoons, computer graphics, x-ray images, CAT scans, etc., natural images and videos possess their unique statistical characteristics, which has led to the development of many NSS (Natural Scene Statistics) models to capture them. As argued in [61], since distortions may disturb the statistics of natural scenes, a deviation of a distorted signal from the expected natural statistics can be used to quantify the distortion intensity. NR metrics based on this philosophy are few but very enlightening. For example in [61], this philosophy was clearly stated for the first time, and a NR metric was developed for JPEG2000 coded images, in which a NSS model [7] was employed to capture the non-linear dependencies of wavelet coefficients across scales and orientations. In [75], the authors proposed a technique for coarse-to-fine phase prediction of wavelet coefficients. It was observed that the proposed phase prediction is highly effective in natural images and it can be used to measure blurring artifacts that will disrupt this phase coherence relationship.

3 Subjective Evaluation Standard

Objective quality metrics are developed to approximate human perceptions, but up to this moment, no single metric can completely represent human response in visual quality assessment. As a result, subjective evaluation is still the only mean to fully characterize the performance of different systems. A standard procedure is therefore needed for fair and reliable evaluations.

Several international standards are proposed for subjective video quality evaluation for different applications. The most commonly referenced standard is the ITU-R BT.500 defined by International Telecommunication Union (ITU) [26]. ITU-R BT.710 [25] is an extension of BT.500 dedicated for high-definition TV. ITU-T P.910 [29] is another standard which defines the standard procedure of digital video quality assessment with transmission rate below 1.5Mbit/s. These standards provide guidelines for various aspects of subjective evaluations, such as viewing condition, test sequence selection, assessment procedures, and statistical analysis of the results. The Video Quality Expert Group (VQEG) also proposed several subjective evaluation procedures to evaluate the performance of different objective quality metrics [68, 69, 70]. These proposed methods share many similarities to the BT.500 and P.910 standards.

A brief description of various aspects of the subjective evaluation standards will be discussed in this section. This includes the general viewing condition, observer selection, test sequence selection, test session structure, test procedure, and post-processing of scores.

3.1 Viewing Condition

The general viewing condition defines a viewing environment that is most suitable for visual quality assessment. It minimizes the environment's effect on the quality of the image or video under assessment. In BT.500, two environments (laboratory environment and home environment) are defined. Laboratory viewing environments is intended for system evaluation in critical conditions. Home viewing environment, on the other hand, is intended for quality evaluation at the consumer side of the TV chain. The viewing conditions are designed to represent a general home environment. Table 1 tabulates some parameters of viewing conditions used in different standards.

Table 1. General Viewing Conditions

	Condition	BT.710	BT.500 (lab env)	BT.500 (home env)	P.910
a	Ratio of viewing distance to picture height	3	-	Function of screen height	1-8H
b	Peak luminance on the screen (cd/m^2)	150-250	-	200	100-200
c	Ratio of luminance of inactive screen to peak luminance	≤0.02	≤0.02	≤0.02	≤0.05
d	Ratio of luminance on the screen when displaying only black level in a completely dark room, to that corresponding to peak white	Approx. 0.01	Approx. 0.01	-	≤0.1
e	Ratio of luminance of background behind picture monitor to peak luminance of picture	Approx. 0.15	Approx. 0.15	-	≤0.2
f	Illumination from other sources	low	low	200lux	≤20lux
g	Chromaticity of background	D_{65}	D_{65}	-	D_{65}
h	Arrangement of observers	Within ±30^o horizontally from the center of the display. The vertical limit is under study	Within ±30^o relative to the normal (only apply to CRT, other display are under study)	Within ±30^o relative to the normal (only apply to CRT, other display are under study)	-
i	Display size	1.4m (55 in)	-	-	-
j	Display brightness and contrast	-	Setup via PLUGE [28, 24]	Setup via PLUGE [28, 24]	-

3.2 Candidate Observer Selection

To obtain reliable assessments from observers, certain requirements on the selection of observers must be met. Firstly, their eyesight should be either normal or has been corrected to normal by spectacles. Prior to the test session, observer must be screened for normal eyesight, which includes normal visual acuity and normal color vision. Normal visual acuity can be checked by the Snellen or Landolt chart. A person is said to have normal acuity when he/she can correctly recognize the symbols on the standard sized Snellen chart 20/20 line when standing 20 feet from the chart. At 20 feet, the symbols on the 20/20 line subtend five minutes of arc to the observers, and the thickness of the lines and spaces between lines subtends one minute of arc. Normal color vision can be checked by specially designed charts, for instance, the Ishihara charts. Numbers or patterns of different colors are printed on plates with colored background. A person with color deficiency will see numbers or patterns different from the person with normal color vision. The observer is classified as having normal vision if he/she has no more than a certain number of miss-identification of the patterns. In [29], it is required that observers cannot make more than 2 mistakes out of 12 test plates.

Another requirement to the observers is that they should be familiar with the language used in the test. The observer must be able to understand the instruction and provide valid response with semantic judgment terms expressed in that language.

In a formal subjective evaluation experiment, the observers should be non-experts, i.e., they should not be directly involved in video quality assessment as part of their works, and should not be experienced assessors. At least 15 observers should be used to provide statistically reliable results. However, in early phase of the development stage, an informal subjective evaluation with 4 to 8 expert observers can provide indicative results.

3.3 Test Sequence Selection

No single set of test material can satisfy all kinds of assessment problems. Particular types of test material should be chosen for particular assessment problems. For example, in a study of the overall performances of two video coding systems for digital TV broadcast, sequences with a broad range of contents and characteristics should be used. On the other hand, if the systems being assessed are targeted for video conference on mobile network, head-and-shoulder sequences should be chosen.

For general purpose video system assessment, a broad range of contents should be chosen. Video sequences from movies, sports, music videos, advertisements, animations, broadcasting news, home videos, documentaries, etc., can be included in the test set. Different characteristics of the test sequences should also be considered. The test set should have different levels of color, luminance, motion, spatial details, scene cuts, etc.

When impairment evaluation is performed, the sequences with different levels of impairment should be produced in order to generate tractable results for performance analysis.

3.3.1 Spatial and Temporal Information

Sequences with different levels of spatial and temporal information should be chosen for general purpose assessment. The P.910 standard defined a method to measure these kinds of information in a video sequence.

The spatial information (*SI*) is a simple measure of the spatial complexity in the sequence. Each luminance plane F_n in a video frame at time *n* is first filtered with the Sobel filter to obtain a filter output $Sobel(F_n)$. The standard deviation over all pixels in the frame, $std_{space}[Sobel(F_n)]$, is then computed. This operation is repeated for every frame in the sequence. The *SI* is defined as the maximum standard deviation of all frames, i.e.,

$$SI = max_{time}\{std_{space}[Sobel(F_n)]\}. \tag{13}$$

The temporal information (*TI*) measures the change of intensity of the frames over time. Intensity change can be caused by motions of objects or background, change of lighting conditions, camera noises, etc. To compute *TI*, first we compute $M_n(i,j)$, the difference between pixel values of the luminance plane at the same location but in successive frames, i.e.,

$$M_n(i,j) = F_n(i,j)- F_{n-1}(i,j), \tag{14}$$

where $F_n\ (i,j)$ is the pixel value at i^{th} row and j^{th} column in the frame at time *n*. *TI* is then defined as

$$TI = max_{time}\{std_{space}[M_n(i,j)]\}. \tag{15}$$

If there are scene cuts in the sequence, two values of *TI* may be computed, one for sequence with scene cut, and one for sequence without scene cut. In addition, scenes with high *SI* are generally associated with relatively high *TI*, since motion in complex scenes usually results in large differences in intensity in successive frames.

3.4 Structure of Test Session

The maximum duration of a single test session is 30 minutes. A break of a few minutes should be given to the observer before the next test session starts. At the beginning of the first test session, about five stabilizing presentations should be introduced to stabilize assessor's opinions. Scores obtained from these presentations should not be taken into account. For subsequent sessions, about three stabilizing sequences should be presented at the beginning of each session. In addition, several training sequences should be introduced before the first session starts in order to familiarize the observers with the assessment procedure. The observers can ask questions regarding the assessment after the training sequences. The whole process is illustrated in Fig. 14.

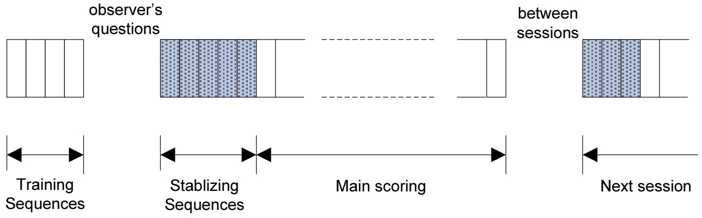


Fig. 14. Test session structure

A pseudo random order of the sequences should be used for each assessor. The order can be derived from Graeco-Latin squares or other means. This can reduce the effects of tiredness or adaptation on the grading process. Also, the same picture or sequence should not be used for consecutive presentations, even with different levels of distortions, to prevent ambiguity.

3.5 Assessment Procedure

There are numerous Full-Reference and No-Reference assessment procedures targeted for different problems. In the full-reference case, the two commonly used evaluation procedures are double-stimulus impairment scale (DSIS) and double-stimulus continuous quality-scale (DSCQS). DSIS is used for failure characterization, e.g., identifying the effect of certain impairment introduced in the video encoding or transmitting process. DSCQS is used when an overall system evaluation is need. No-Reference assessment can be performed using the single stimulus (SS) method. These three assessment procedures will be briefly described in this section. Readers should refer to [26] for more detailed descriptions of different assessment procedures.

3.5.1 Double-Stimulus Impairment Scale

DSIS method is a cyclic assessment procedure. An unimpaired reference image or sequence is presented to the observer first, followed by the same signal with impairments added by the system under test. The observer is asked to vote on the impaired one while keeping in mind the reference.

The structure of presentations is shown in Fig. 15. There are two variants of the structure. In variant I, the reference and the impaired picture or sequence are presented only once. In variant II, the reference and impaired material are presented

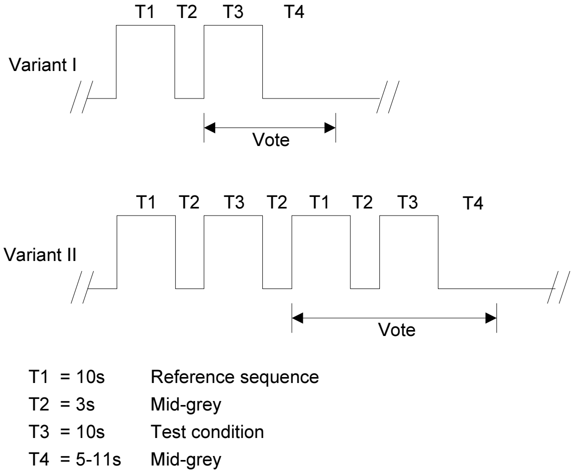


Fig. 15. Presentation structure of DSISï

twice. When the impairment is very small or when the presented material is a video sequence, variant II is preferred though it is more time consuming. The observers should be asked to look at the reference and impaired material for the whole duration of T1 and T3, respectively. A mid-gray screen (T2) is shown between T1 and T3 for a duration of about 3 seconds. If variant I is used, voting period starts immediately after the impaired material is presented. If variant II is used, voting period starts when the pair of reference and impaired material is shown at the second time. In either variant, a 5 to 11 seconds mid-gray (T4) is displayed before the next presentation.

The scores are recorded by using the five-grade impairment scale. The five grades and their distortion level descriptions are as follow:

5	imperceptible
4	perceptible, but not annoying
3	slightly annoying
2	annoying
1	very annoying

A form with clearly defined scale, numbered boxes or some other means to record the grading should be provided to the observers for score recording.

The impairments of the test material should have a broad range so that all five grades in the grading scale should be chosen by the majority of observers, and the impairments should be evenly distributed among the five grading levels.

3.5.2 Double-Stimulus Continuous Quality Scale

The double stimulus continuous quality scale (DSCQS) method is an effective evaluation method for overall system performance. The structure of the presentation is somewhat similar to that of DSIS, except that the pair of reference and impaired material is presented in random order. The observer does not have the knowledge of the display order, and he/she will give scores for both reference and impaired images or sequences.

There are two variants of the DSCQS presentation structure. In variant I, only one observer participates in a test session. The observer is free to switch between signal A and B until he/she is able to determine a quality score for each signal. This process may be performed two or three times for duration of up to 10 seconds. In variant II, at most three observers can assess the material simultaneously. If the material is a still picture, each picture will be displayed for 3 to 4 seconds with five repetitions. For video sequences, the duration of each sequence is about 10 seconds with two repetitions. This presentation structure is illustrated in Fig.16.

The observers are asked to assess the overall picture quality of each presentation by inserting a mark on a vertical scale. An example is shown in Fig. 17. The vertical scales are printed in pairs since both the reference and the impaired sequence must be assessed. The scales provide a continuous rating system for score from 0 to 100, which is different from the five-grade scale used in DSIS. They are divided into five equal lengths and associated descriptive terms are printed on the left of the first scale as general guidance to the observer. To avoid confusion, the

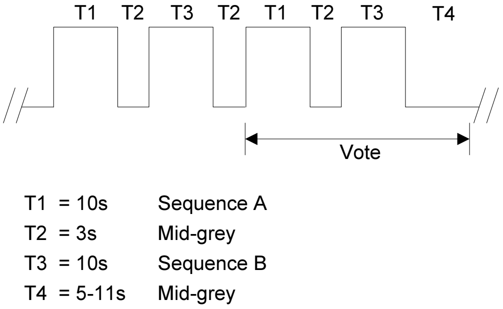


Fig. 16. Presentation structure of DSCQS

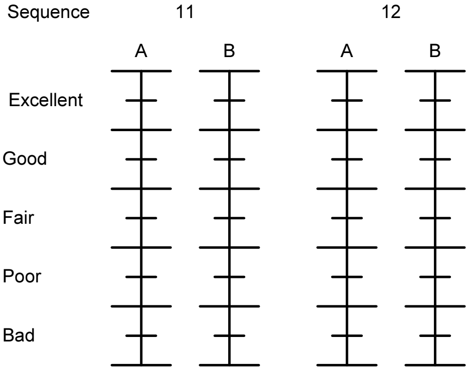


Fig. 17. Grading scale score sheet for DSCQSï

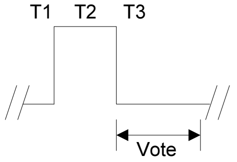
observer should use pen with color different from the printed scale. Electronic scoring tools can be used only if its display does not compromise the viewing conditions listed in Table 1.

3.5.3 Single Stimulus Methods

In single stimulus methods, the image or sequence is assessed without the reference source, and a presentation consists of three parts: a mid-gray adaptation field, a stimulus, i.e., the image or sequence being assessed, and a mid-gray post-exposure field. The durations of these parts are 3, 10, and 10 seconds, respectively. The voting of the stimulus can be performed during the display of the post-exposure field. The overall structure is illustrated in Fig. 18.

Depending on the applications, grading can be recorded by the 5-point impairment scale used in DSIS, an 11-grade numerical categorical scale described in ITU-R BT.1082 [23], or the continuous scale used in DSCQS.

Another variant of SS is that the whole set of test stimuli are presented three times, which means that a test session consists of three presentations. Each of



T1 = 3s Mid-gray adaptation
T2 = 10s Stimulus
T3 = 10s Mid-gray post-exposure

Fig. 18. Presentation structure of SS

them includes all the images or sequences to be tested only once. The first presentation serves as a stabilizing presentation. The scores obtained from this presentation will not be taken into account. The score of each image or sequence is the mean score obtained from the second and third presentation. The display orders of the images or sequences are randomized for all three presentations.

3.6 Post-Processing of Scores

After obtaining the scores from the observers, these data must be statistically summarized into meaningful form for analysis. In addition, observers should be screened and statistically unreasonable results should be discarded. A relationship between an objective measurement of the picture quality and the subjective score can also be deduced from the data obtained.

The following analysis is applicable to the results from the DSIS, DSCQS, and other methods which use numerical scales for grading. In the first case, the impairment is rated on a five-point or multi-point scale and the score range is from 1 to 5. In the second case, continuous rating scales are used and the results have integer values between 0 and 100.

3.6.1 Mean Scores and Confidence Interval Calculation

The common representation of the scores is the mean score and confidence interval. Let *L* be the number of presentations in the test, *J* be the number of test conditions applied to a picture or sequence, *K* be the number of test images or sequences, *R* be the number of repetitions of a test condition applied on a test picture or sequence. The mean score, $\overline{u}_{jkr}$, for each of the presentations is defined as

$$\overline{u}_{jkr} = \frac{1}{N}\sum_{i=1}^{N} u_{ijkr} , \tag{16}$$

where u_{ijkr} is the score of observer *i* for test condition *j*, sequence/image *k*, repetition *r*; and *N* is the number of observers. The overall mean scores, $\overline{u}_j$ and $\overline{u}_k$, could be calculated for each test condition and each test image or sequence in a similar manner.

The confidence intervals should be presented in addition to the mean scores to provide more information about the variability of the results. The confidence interval is derived from the standard deviation and sample size. [26] proposed to use the 95% confidence interval, which is defined as $\lfloor \bar{u}_{jkr} - \delta_{jkr}, \bar{u}_{jkr} + \delta_{jkr} \rfloor$ where:

$$\delta_{jkr} = 1.96 \frac{S_{jkr}}{\sqrt{N}}, \tag{17}$$

and the standard deviation for each presentation, S_{jkr}, is defined as:

$$S_{jkr} = \sqrt{\sum_{i=1}^{N} \frac{\left(\bar{u}_{jkr} - u_{ijkr}\right)^2}{(N-1)}}. \tag{18}$$

The 95% confidence interval indicates that with a probability of 95%, the mean score will be within the interval if the experiment is repeated for a large number of times. As more samples are available, the confidence interval range gets smaller and the mean score becomes more reliable.

3.6.2 Screening of Observers

Sometimes scores obtained from certain observers may deviate from the distribution of the normal scores significantly. This kind of observers must be identified and their scores discarded from the test. The β_2 test is suggested in BT.500 to accomplish such task.

For each test presentation, we first compute the mean, $\bar{u}_{jkr}$, standard deviation, S_{jkr}, and kurtosis coefficient, β_{2jkr}, where β_{2jkr} is given by:

$$\beta_{2jkr} = \frac{m_4}{(m_2)^2} \text{ with } m_x = \frac{\sum_{i=1}^{N} \left(u_{ijkr} - \bar{u}_{jkr}\right)^x}{N}. \tag{19}$$

If $\beta_{2\mathrm{jkr}}$ is between 2 and 4, the distribution of the score can be assumed to be normal. Now for each observer, i, we need to find the number of score entries that lie outside of the score distribution of each test presentation. The valid range of distribution of each test presentation is defined as $\bar{u}_{jkr} \pm 2S_{jkr}$ if the distribution is normal. For non-normal distribution, the valid range is defined as $\bar{u}_{jkr} \pm \sqrt{20}S_{jkr}$. Let P_i and Q_i be the number of times that the score from observer i is above and below the valid range, respectively. P_i and Q_i can be computed by the following procedure:

```
for $j, k, r = 1, 1, 1$ to $J, K, R$
  if $2 \le \beta_{2\mathrm{jkr}} \le 4$, then:
    if $u_{ijkr} \ge \bar{u}_{jkr} + 2S_{jkr}$ then $P_i = P_i + 1$
    if $u_{ijkr} \le \bar{u}_{jkr} - 2S_{jkr}$ then $Q_i = Q_i + 1$
```

else:
if $u_{ijkr} \geq \bar{u}_{jkr} + \sqrt{20} S_{jkr}$ then $P_i = P_i$+1
if $u_{ijkr} \leq \bar{u}_{jkr} - \sqrt{20} S_{jkr}$ then $Q_i = Q_i$+1

where *J*, *K*, and *R* have the same meaning as in section 1.3.6.1. After computing P_i and Q_i for observer *i*, if the following two conditions are met, then observer *i* will be rejected.

Condition 1: $\frac{P_i + Q_i}{J \cdot K \cdot R} > 0.05$

Condition 2: $\left| \frac{P_i - Q_i}{P_i + Q_i} \right| < 0.3$

The observer screening procedure should not be applied to the results of a given experiment more than once. In addition, it should be restricted to the experiment when there are relatively few observers (e.g., fewer than 20) participating the experiment and all of them are non-experts.

3.6.3 Relationship between the Mean Score and the Objective Measure of a Picture Distortion

When evaluating a relationship between the mean scores and a type of impairment at different levels, or between the mean scores and some objective measurements of distortion, it will be useful if the relationship can be represented by a simple continuous function with the mean score as the dependent variable.

In [26], the symmetric logistic function and a non-symmetric function are introduced to approximate the relationship. For both cases, the mean score *u* must first be normalized by taking a continuous variable *p* so that

$$p = \frac{\left(\bar{u} - u_{\min}\right)}{\left(u_{\max} - u_{\min}\right)}, \tag{20}$$

where u_{min} is the minimum score available on the *u*-scale for the worst quality; and u_{max} is the maximum score available on the *u*-scale for the best quality. The normalized mean score *p* can be estimated by a symmetric logistic function. Let $\hat{p}$ be the estimate of *p*. The function $\hat{p} = f(D)$, where *D* is the distortion parameter, can now be approximated by a judiciously chosen logistic function, as given by the general relation

$$\hat{p} = f(D) = \frac{1}{1 + e^{(D - D_M) \cdot G}}, \tag{21}$$

where D_M and *G* are constants and *G* may be positive or negative. To solve for D_M and *G*, we define

$$I = \frac{1}{p} - 1\,, \tag{22}$$

and its estimate

$$\hat{I} = \frac{1}{\hat{p}} - 1\,. \tag{23}$$

Combining (1.21) and (1.23), we obtained

$$\hat{I} = e^{(D - D_M)\cdot G}\,. \tag{24}$$

Let J be the natural log of I, and $\hat{J}$ be the natural log of $\hat{I}$, i.e.,

$$\hat{J} = \log_e I = (D - D_M)\cdot G\,. \tag{25}$$

A linear relationship between $\hat{J}$ and D is established. D_M and G can then be found by minimizing ε, the mean squared estimation error between $\hat{J}$ and J, which is defined as

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N}\left(J_i - \hat{J}_i\right)^2\,. \tag{26}$$

The simple least square method can be used to find the optimal D_M and G.

The symmetrical logistic function is particularly useful when the distortion parameter D can be measured in a related unit, e.g., the S/N (dB). If the distortion parameter was measured in a physical unit d, e.g., a time delay (ms), then the relationship between $\hat{p}$ and d can be defined as

$$\hat{p} = \frac{1}{1 + (d / d_M)^{1/G}}\,. \tag{27}$$

This is a non-symmetric approximation of the logistic function. Similar to the case of logistic function, we define $\hat{I}$ as

$$\hat{I} = \frac{1}{\hat{p}} - 1\,, \tag{28}$$

and the estimated $\hat{J}$ is

$$\hat{J} = \log\left(\hat{I}\right) = \frac{1}{G}\log\left(\frac{d}{d_M}\right). \tag{29}$$

The optimal values of d_M and G can be solved by minimizing the mean squared error ε defined in (1.26) using the Levenberg-Marquardt algorithm.

4 An Application of Quality Comparison: AVS versus H.264

4.1 Background

One important application of subjective quality evaluation is to compare the performance of two video encoding systems. Since human observers are the final judge of the quality of the encoded video, a subjective evaluation process is necessary to obtain a comprehensive understanding of the performance of the systems being tested.

In this section, we describe a comparison between two encoding systems, namely, the H.264/Advanced Video Coding (AVC) and the Audio and Video Coding Standard (AVS), based on objective distorion measurements and subjective evaluation. H.264/AVC is the most recent video coding standard jointly developed by the ITU-T Video Coding Experts Group (VCEG) and the ISO/IEC Moving Picture Experts Group (MPEG) [21]. Various profiles are defined in the standard to suite different applications. For example, simple baseline profile provides the basic tools for video conferencing and mobile applications that run on low-cost embedded systems, while the main profile is used in general applications. More complex High, High 10, and High 4:2:2 profiles are introduced in the Fidelity Range Extension (FRext) of H.264 to further improve the coding efficiency for high-definition (HD) video and studio quality video encoding [62]. New coding tools, e.g., 8×8 block size transform, support of different chroma format, and precision higher than 8 bits, are utilized in these profiles.

AVS is a new compression standard developed by AVS Workgroup of China [3, 4, 22]. AVS Part 2 (AVS-P2) is designed for high-definition digital video broadcasting and high-density storage media. It is published as the national standard of China in February, 2006. Similar to the H.264/AVC, AVS is a hybrid DPCM-DCT coding system with compression tools like spatial and temporal prediction, integer transform, in-loop deblocking filter, entropy coding, etc [93]. The target applications of AVS include HD-DVD and satellite broadcast in China. AVS-P2 also introduced the X profile, which is designed for high quality video encoding. Several new tools are introduced: Macroblock-level adaptive frame/field coding (MBAFF), adaptive weighting quantization, adaptive scan order of transform coefficients, and arithmetic coding of most syntax elements.

In general, the structure of AVS and H.264 are very similar. The major difference is that many components in the AVS are less complex than the H.264 counterpart. For example, AVS utilizes only 4 different block sizes in motion estimation, while H.264 uses 7 block sizes. AVS has only 5 luminance block intra-prediction modes, compared to 13 modes in H.264. In addition, AVS utilizes a simpler in-loop deblocking filter, shorter tap filter for sub-pixel motion estimation, and other techniques to reduce the complexity of the encoder. The AVS encoder requires only about 30% of the computation load of H.264 for encoding, but it is able to achieve similar coding efficiency.

Objective comparisons between H.264-main profile and AVS-base profile were reported in several articles, e.g., in [93, 16, 72]. The results generally show that for smaller frame sized sequences, such as QCIF, and CIF, H.264 has a slight advantage over AVS. For SD and HD video, AVS and H.264 are similar in their rate distortion performance. However, no comparison has been made between the AVS-X profile and the H.264-High profile. In this section, the results of the objective and subjective performance comparisons between these two profiles will be presented. Section 1.4.2 describes the setup of the experiment. Objective and subjective evaluation results are presented in Sections 1.4.3 and 1.4.4, respectively. In section 1.4.5, we compare the accuracy of PSNR, and two other objective quality metrics that correlate better to human perception than PSNR, i.e., the Structural Similarity (SSIM) index and the Video Quality Metric (VQM), using the results obtained from subjective evaluation.

4.2 Test Setup

The performance comparison of AVS-X profile and H.264-high profile is divided into two parts: objective comparison and subjective comparison. In objective comparison, rate-distortion performance in terms of PSNR and bit rates is first used as the evaluation metric. Subjective evaluation of visual quality is also performed to compare their coding performances as perceived by human observers.

4.2.1 Sequence Information

Table 2 shows all the video sequences used in this comparison. Since the target of this comparison is for high-fidelity video, only HD video sequences were used in the test. For 1280×720 progressive (720p) sequences, the target bit rates used for the test were 4, 8, 10, and 15 Mbps, and the frame rate was 60Hz. Both 1920×1080 progressive (1080p) and 1920×1080 interlace (1080i) sequences were encoded at target bit rates of 6, 10, 15, and 20 Mbps, and at a frame rate of 25Hz.

Table 2. Test Sequences Used in Test

720p	1080p	1080i
City	*PedestrianArea*	*NewMobileCalendar*
Crew	*Riverbed*	*Parkrun*
Harbour	*Rushhour*	*Shields*
ShuttleStart	*Sunflower*	*StockholmPan*
SpinCalendar	*ToysCalendar*	*VintageCar*

4.2.2 Encoder Setting

The JM 14.0 reference H.264/AVC encoder and the rm6.2h reference AVS encoder are used to encode the test video sequences. The IBBPBBPBBP… GOP structure was used, with intra-frames inserted in every 0.5 sec, i.e., one intra-frame in every 30 frames for 720p, and in every 12 frames for 1080p and 1080i. Table 3 shows the general settings of the encoders. Note that due to memory limitation, only 2 reference frames were used when interlace videos are encoded with the H.264 encoder.

Table 3. Encoder Parameter Settings.

Setting	H.264	AVS-P2
Encoder version	JM 14.0	rm6.2h
Profile	high	X
Number of reference frame	4 (2 for 1080i)	2
Block size	16×16, 16×8, 8×16, 8×8, 8×4, 4×8, 4×4	16×16, 16×8, 8×16, 8×8
Fast ME	Enabled	Enabled
ME search range	32	32
RD Optimization	Enabled	Enabled
Interlace mode	PAFF	PAFF
Loop filter	Enabled	Enabled
Adaptive scan	-	Enabled
Adaptive filter	-	Disabled

4.2.3 Subjective Test Setup

The subjective assessment was performed in a studio room with lighting condition satisfying the lab environment requirement of the ITU-R BT.500 standard, which is also briefly described in section 1.3. The display monitor is a 65” Panasonic plasma display (TH-65PF9WK) and the viewing distance is 3 times the picture height. Background illumination has a D65 chromaticity.

Thirty-five non-expert observers participated in the subjective test, and about half of them were male. All of them did not work in video processing related jobs, and were not involved in any video quality assessment within the past four months. Their eyesight was either normal or had been corrected to be normal with spectacles.

Each observer compared 39 pairs of “reference” (H.264) and “processed” (AVS) sequences (13 pairs each for 720p, 1080p, and 1080i). The double-stimulus continuous quality scale (DSCQS) test method as described in section 1.3.5.2 was used for this subjective test.

Note that the uncompressed sequences are not used as the references as in normal practice because we want to compare the visual quality of H.264 and AVS sequences directly. Direct comparison allows the observers to immediately identify the small differences in visual quality and record the scores.

4.3 Rate-Distortion Performance Comparisons Using PSNR, Bitrates

The average PSNR change (ΔPSNR) and bit rate change (ΔBitrate) computed by the method described in [6] are used to measure the objective performance of the two coding systems. The ΔPSNR is a measure of the difference in PSNR of the target and reference systems under the same bit rate range. Similarly, ΔBitrate measures the difference of bit rates under the same PSNR range. A sequence is encoded at four different bit rates by both the "reference" and the "target" systems, then ΔPSNR and ΔBitrate of the target system are computed from these rate-distortion points as illustrated in Fig 19. In general, a ΔPSNR of 0.3dB is equivalent to a ΔBitrate of approximately 5%.

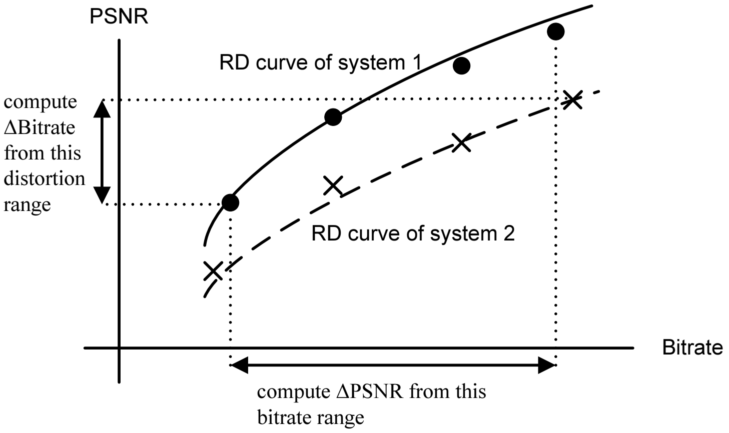


Fig. 19. Illustration of RD performance evaluation by ΔPSNR and ΔBitrate

The advantage of using this performance evaluation method is that it evaluates the performance of a system at multiple bit rates, so that a better understanding of the system under different bit rates can be acquired. In addition, both bit rate and the PSNR of encoded video sequences can be computed easily. Because of these advantages, this method is commonly used in encoding system performance comparison. In our experiment, sequences in H.264 format are defined as the reference. The ΔPSNR and ΔBitrate of all the test sequences encoded by AVS are shown in Table 4.

For the 720p sequences, the average ΔBitrate of all sequences is 0.71%, which implies that the overall coding efficiency of AVS and H.264 is very similar. However, the ΔBitrate fluctuates from -8.98% (*Crew*) to 10.72% (*SpinCalendar*). The coding efficiency of AVS depends quite heavily on the content of the sequence. For the 1080p sequences, the average ΔBitrate is -2.31%. AVS seems to have a slight advantage in encoding these sequences. The range of ΔBitrate varies only from -7.81% to 0.78%. The variation is less than that for 720p sequences. On the contrary, AVS has an average ΔBitrate of 5.48% for 1080i sequences, which means the coding efficiency is lower than that of H.264. In addition, a large variation in ΔBitrate is observed, with a range from -4.56% to 19.77%. The rate-distortion (RD) curves of several sequences are shown in Fig 20. The RD performance we obtained is similar to those reported in [16]. Although they were using older reference encoders, they also reported that AVS performs slightly worse on sequences like *City* and *SpinCalendar*, and slightly better on sequences like *Harbour* and *Crew*.

Table 4. ΔPSNR and ΔBit rate for all test sequences

Size	Sequence	ΔPSNR	ΔBitrate	*SI*
720p	*City*	-0.247	10.45%	77.41
	Crew	0.220	-8.98%	72.21
	Harbour	0.109	-3.60%	92.3
	ShuttleStart	0.092	-5.01%	35.21
	SpinCalendar	-0.205	10.72%	103.27
	Average	-0.006	0.71%	-
1080p	*PedestrianArea*	0.016	-1.10%	37.11
	Riverbed	0.402	-7.81%	39.46
	Rushhour	-0.040	-2.16%	26.74
	Sunflower	0.005	0.78%	39.39
	ToysCalendar	0.001	-1.26%	54.7
	Average	0.077	-2.31%	-
1080i	*NewMobileCalendar*	-0.398	19.77%	73.44
	Parkrun	-0.194	5.18%	123.01
	Shields	-0.188	10.08%	60.02
	StockholmPan	0.044	-3.08%	73.95
	VintageCar	0.090	-4.56%	58.3
	Average	-0.129	5.48%	-
	Overall Average	-0.019	1.29%	-

The RD performance shows that the content of the sequence has certain impact on the coding efficiency. It seems that H.264 performs better in sequences with lots of textures, such as *City*, *SpinCalendar*, *NewMobileCalendar*, and *Shields*. ΔBitrate of AVS are more than 10% in these sequences. The right-most column in Table 4 shows the spatial information (*SI*) of all sequences. *SI* is a simple measure of spatial texture introduced in section 3.1. Fig. 21 shows the relationship between ΔBitrate and *SI* for different sequences. Although there is no linear relationship between bitrate and *SI*, we can still see that positive ΔBitrate appears more frequently on sequences with high *SI*, i.e., highly textured sequences. In addition, the Pearson's correlation coefficient between ΔBitrate and SI is 0.402, suggesting that there is a positive relationship between *SI* and ΔBitrate. For other types of video, AVS is able to achieve a higher efficiency than H.264, e.g., *Crew* and *Riverbed*. The usage of more than 2 reference frames and more complex interpolation filter in H.264 does not seem to give a significant improvement in coding efficiency for these sequences. The simpler coding tools in AVS are sufficient to give similar efficiency compared with H.264.

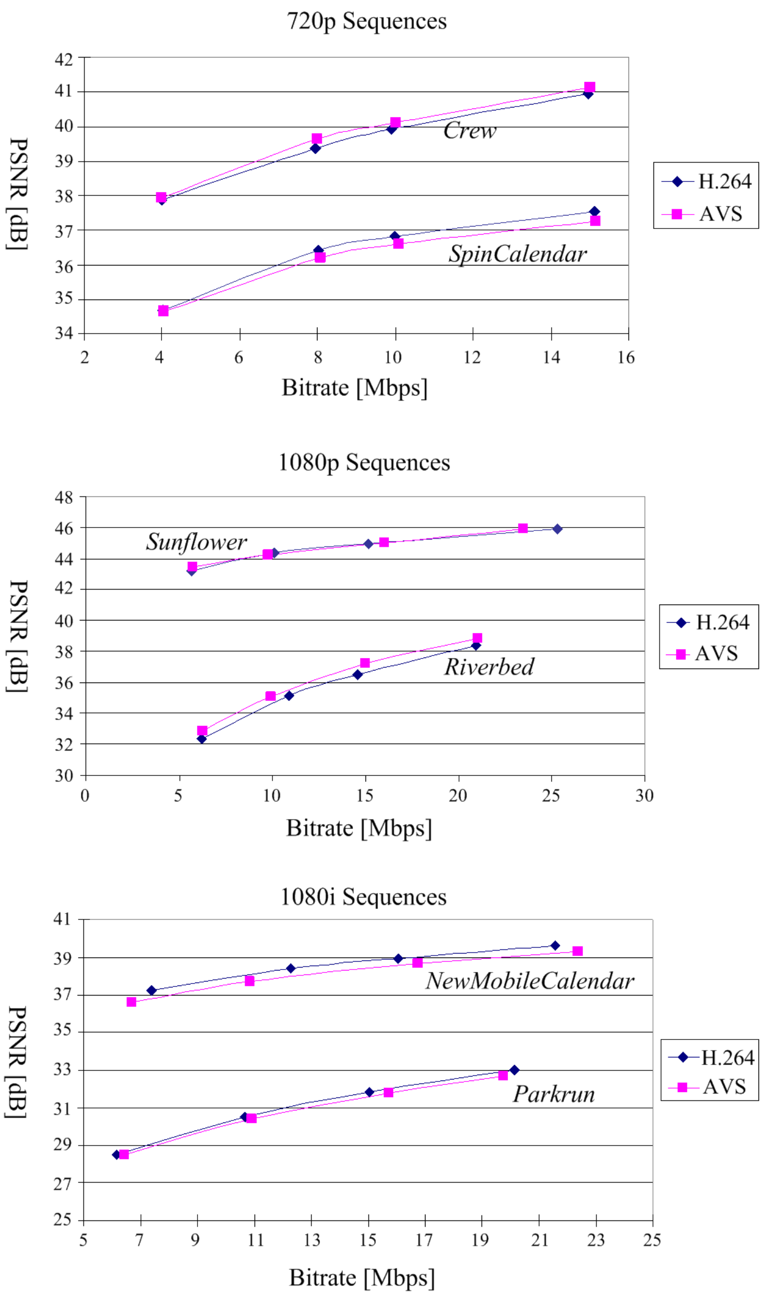


Fig. 20. RD Curves of some 720p, 1080p, and 1080i sequences

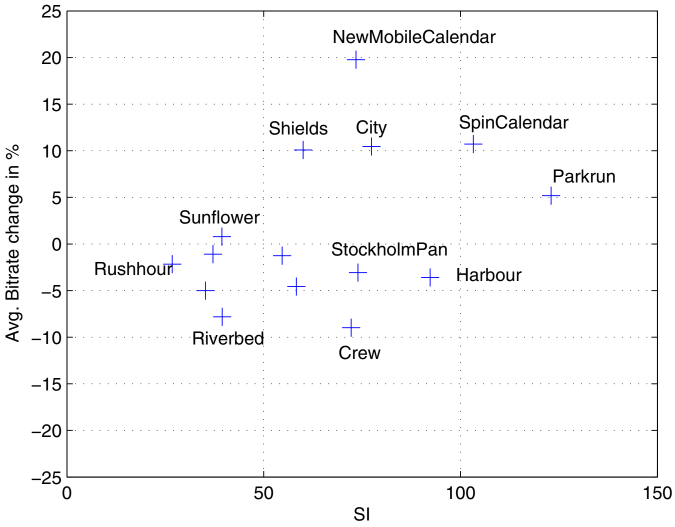


Fig. 21. ΔBitrate and SI relationship for different sequences

4.4 Subjective Evaluation Results

The scores given by the observers in the subjective test are used to evaluate the subjective quality of the sequences. The mean opinion score (MOS) of each sequence is first computed. Then the difference mean opinion score (DMOS), i.e., the difference in MOS between the AVS and the H.264 sequences, is used to compare the subjective quality of H.264 and AVS. A positive DMOS implies AVS has a better subjective quality than H.264. The DMOS for the all sequences, along with the 95% confidence level, are shown in Fig. 22.

The average DMOS of all sequences is 0.13 with standard deviation of 2.26. Note that the full range of the score is 100. This indicates that the overall visual quality of AVS and H.264 are very similar in many sequences. Fig. 23 shows the average DMOS of each sequence for the four bit rates used. The magnitudes of the average DMOS are all less than 3, which is also very small. We also computed the average DMOS of sequences with the same bit per pixel. The results are shown in Fig. 24. It is clear that on the average, AVS and H.264 have very similar performance in visual quality at different bit rates.

For 720p sequences, most of the sequences encoded by AVS have visual quality similar to those encoded by H.264. The DMOS scores range only from -5 to 5 except for one sequence: *Harbour* at 4 Mbps, which has a DMOS of -6.29. Although the PSNR of the AVS encoded sequence is only 0.17dB lower than the one encoded by H.264, the distortion is more obvious than other sequences. In fact, for *City* at 8 Mbps, the PSNR of AVS encoded sequence is 0.33 dB lower than H.264, but the DMOS is 3.24. As mentioned in Section 1.4.3, AVS performs worse in terms of RD performance for sequences containing highly-textured area.

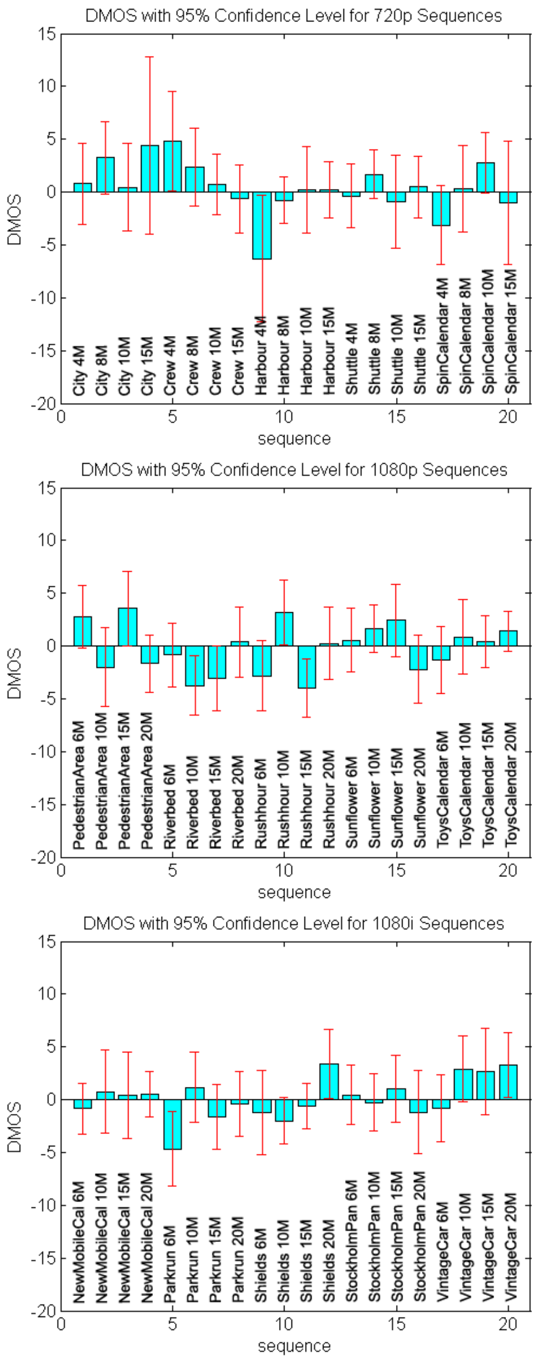


Fig. 22. DMOS for sequences in different frame sizes

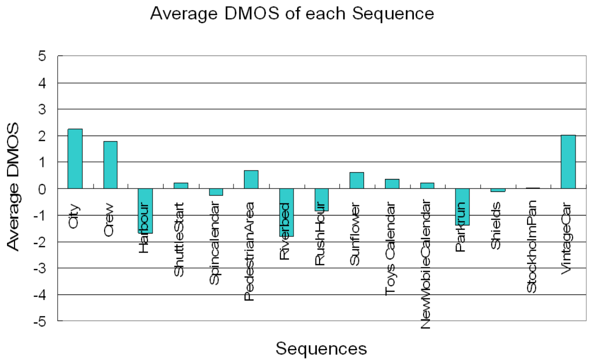


Fig. 23. Average DMOS of each sequence

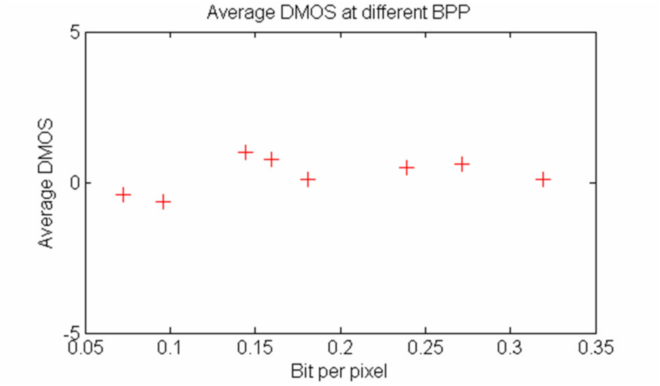


Fig. 24 Average DMOS of sequences with the same BPP

However, the subjective test results show that the visual quality is not affected by the reduction in coding efficiency. The DMOS for textured sequences, such as *City* and *SpinCalendar*, are all close to zero, even when their ΔBitrate are over 10%.

The DMOS obtained for 1080p and 1080i sequences have similar trend to that of 720p sequences. Although all sequences have non-zero DMOS, the magnitudes are all smaller than 5. No obvious difference in visual quality was observed from these sequences. Again, textured sequences such as *NewMobileCalendar* and *Shields* have DMOS close to zero. The ΔBitrate for *NewMobileCalendar* is close to 20% but the difference is visually unobservable. This phenomenon can be explained by the properties of the HVS. The spatial-temporal contrast sensitivity function of human eyes exhibits a non-separable band-pass characteristic with a low sensitivity at high spatial or temporal frequency [14]. Distortions in textured area, especially in moving regions, are less visible to human eyes than those in smooth and slow moving regions. As a result, even when ΔBitrate are over 10% for many sequences, the DMOS for them are all close to zero and the visual qualities are the same. The results clearly show that the commonly used RD performance based on PSNR and bit rates is not a good visual quality performance indicator for video coding systems.

4.5 The Use of PSNR, SSIM, and VQM in Quality Assessment

As discussed in Section 1.2, numerous objective quality metrics have been proposed to measure the visual quality of images or videos. Compared with PSNR, these metrics generally have a higher correlation with the subjective evaluation results. If an objective quality metric can accurately predict the perceived visual quality, then the difference of metric scores should be able to predict the visual difference of two sequences encoded by different systems. This can be illustrated in Fig. 25. Let D_{ref} and D_{tar} be the full-reference distortions computed for sequences encoded by the reference and target systems, respectively. The difference between D_{ref} and D_{tar}, denoted by D_{sys}, should also have a high correlation to the DMOS obtained from subjective evaluation. In this section, two popular metrics, SSIM and VQM, are tested to see if they can model human perception of distortion better than the conventional PSNR. Since SSIM is a metric designed for image, the average SSIM of all frames of the encoded sequence is used in our experiment.

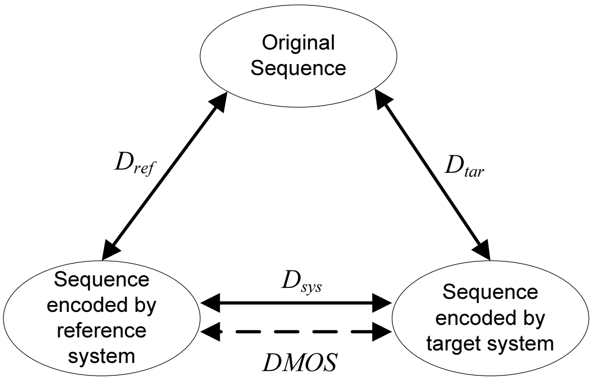

Original Sequence
D_{ref}
D_{tar}
Sequence encoded by reference system
D_{sys}
DMOS
Sequence encoded by target system

Fig. 25. Illustration of relationship of original sequence and sequence encoded by reference and target systems

The performance of an objective quality metric can be evaluated by its correlation to the MOS from subjective evaluation. Three measurements of correlation: Pearson's correlation coefficient (PCC), root mean square error (RMSE), and Spearman's rank order correlation coefficient (SROCC) are used for our evaluation. The PCC between two data sets, X and Y, is defined as

$$PCC(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}} \tag{30}$$

The RMSE between X and Y can be given as

$$RMSE(X,Y)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-Y_i)^2} \tag{31}$$

For the two sets of data X and Y, element X_i and Y_i are converted to rankings x_i and y_i, and SROCC is defined as the PCC of the ranks of X and Y.

A nonlinear mapping between the objective and subjective scores can be applied so that the objective metric can better predict the subjective quality. In both VQEG Phase-I and Phase-II testing and validation, nonlinear mapping is allowed, and the performance of the objective metric is computed after the mapping [60]. The mapping of an objective quality score x is mapped to $Q(x)$ by (32) and (33).

$$Q(x)=\beta_1\cdot\text{logistic}(\beta_2,(x-\beta_3))+\beta_4\cdot x+\beta_5 \tag{32}$$

$$\text{logistic}(\tau,x)=\frac{1}{2}-\frac{1}{1+e^{\tau\cdot x}} \tag{33}$$

The parameters $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ can be found by minimizing the sum of squared difference between the mapped score $Q(x)$ and the corresponding MOS or DMOS. An illustration of the effect of this nonlinear mapping is shown in Fig 26.

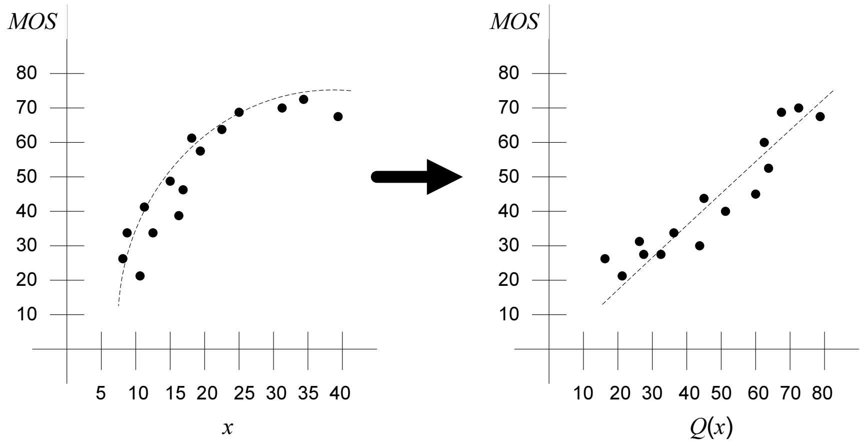


Fig. 26. Illustration of the effect of nonlinear mapping

The PSNR, SSIM, and VQM scores are first computed for all the H.264 and AVS encoded sequences. Then the score differences for the corresponding sequence pairs (that have the same content but encoded by different standards) denoted by DPSNR, DSSIM, and DVQM, are computed. Finally, the PCC, RMSE, and SROCC which measure the correlation between subjective quality scores DMOS and the nonlinearly mapped DPSNR, DSSIM and DVQM are computed. The results are shown in Table 5. VQM generates the highest PCC and SROCC, and the smallest RMSE, which indicates that VQM's correlation to the DMOS is the highest among the three quality metrics. SSIM outperforms PSNR in PCC and RMSE, but its SROCC is slightly lower than that of PSNR.

Table 5. PCC, RMSE, SROCC of three objective distortion metrics

	PCC	RMSE	SROCC
PSNR	0.1396	2.477	0.1718
SSIM	0.3421	2.350	0.1206
VQM	0.4428	2.243	0.1842

The nonlinearly mapped quality scores and the DMOS are plotted in Fig. 27 to Fig. 29. Even after the nonlinear mapping, we still cannot observe strong correlation between DMOS and the objective quality scores. The PCC and SROCC of all distortion metrics are relatively small, to be precise, below 0.5 and 0.2, respectively. This is quite different from the experimental results of other related works, e.g., [55, 77] where the correlation values are above 0.8 typically. The small correlation values are mainly due to the unperceivable differences between the H.264 and AVS encoded sequences. For many sequences, the DMOS is zero while the differences in the objective metrics, i.e., DPSNR, DSSIM, and DVQM, are nonzero. This is different from the experiments in [55, 77], where the visual quality differences between the reference and the distorted images or videos are much more obvious, and therefore the DMOS are usually non-zero. The large number of zero DMOS results in the bad performances of the tested objective quality metrics. Although VQM has comparably better performance, it still cannot accurately predict the DMOS which correspond to nearly unperceivable differences. Therefore for comparing two systems with similar performances, subjective evaluation is still a valuable tool.

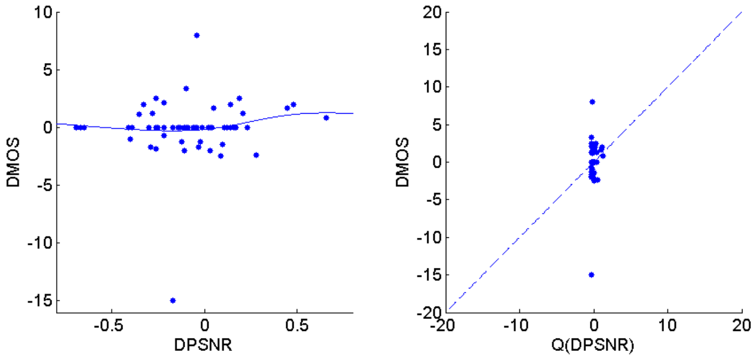


Fig. 27. DPSNR and nonlinear mapped DPSNR vs DMOS

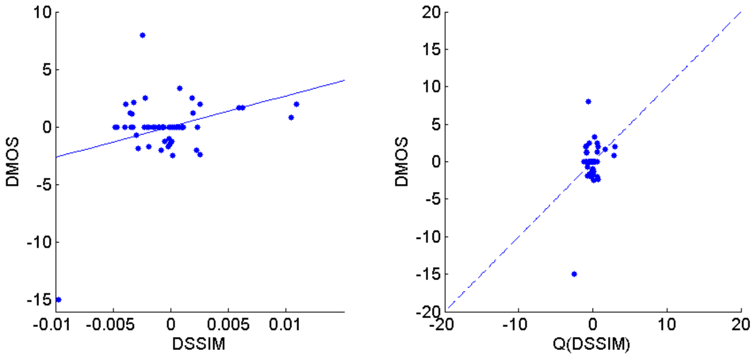


Fig. 28. DSSIM and nonlinear mapped DSSIM vs DMOS

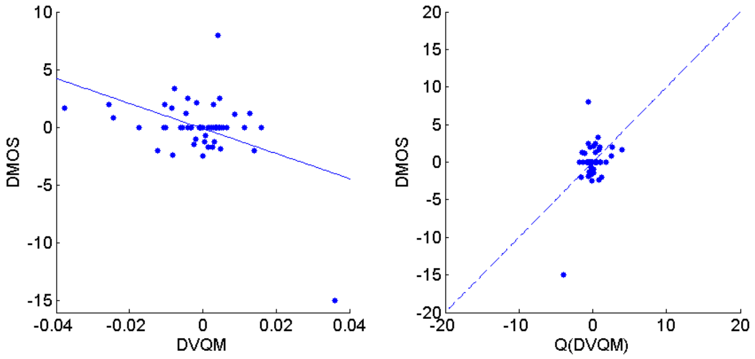


Fig. 29. DVQM and nonlinear mapped DVQMR vs DMOS

5 Conclusions

Subjective evaluation is the most accurate method for measuring perceptual visual quality; however, it is time and money consuming, and is not applicable to most real-world applications. Therefore, objective visual quality metrics are developed, and many of them are discussed in this chapter. We categorized the objective visual quality metrics into two categories, i.e., the HVS-model-based metrics and engineering-based metrics, and introduced them separately. The HVS-model-based metrics account for various low-level characteristics of the HVS, such as luminance adaptation, CSF, contrast masking, etc., which are derived from physiological or psychophysical studies. These perceptual factors and also their implementations in visual quality metrics were discussed in detail. Two quality metric frameworks are presented to illustrate how these perceptual factors cooperate. Different from HVS-model-based quality metrics, engineering-based quality metrics are generally based on assumptions and prior knowledge, e.g., assumptions about the features which the HVS most likely correlate with visual quality, and the prior knowledge about the distortion properties. A conceptual framework was presented for FR and RR metrics, and four classic FR or RR visual quality metrics were summarized by fitting them into this framework. NR metrics were reviewed on the basis of the prior knowledge that they used, including the different distortion types and the statistics of the natural scenes.

An overview of standard subjective evaluation procedure is then presented. Specifically, we described the standard evaluation procedure ITU-R BT.500, which is one of the most commonly used procedures for subjective evaluation. The viewing environment, observer selection, test sequence selection, test procedure and score analysis, which are all important factors that affect the reliability and generality of the evaluation results, are discussed.

We also described an application of subjective video quality evaluation. Two recently developed video coding standards, H.264/AVC and AVS, are compared. Standard objective comparison method utilizing the rate-distortion performance shows that AVS has comparable performance to the H.264/AVC, except in some video sequences that have more complex textures. However, subjective evaluation shows that even on these sequences, the performance of the two systems are about the same. This demonstrates that the commonly used rate-distortion performance is not an accurate performance evaluation method. Two objective metrics, SSIM and VQM, are then utilized as the distortion measures and compared with PSNR. The results show that they are more correlated to the subjective evaluation results, but still cannot completely reflect the HVS perception.

With more knowledge on the psychophysical model of HVS, more accurate objective quality metrics can be developed in the future. However, modeling the extremely complex HVS is a challenging task. Until a thorough understanding of the human perception can be established, subjective evaluation will remain to be the most reliable method we can use for visual quality evaluation.

References

1. Ahumada, J.A.J., Peterson, H.A.: A visual detection model for DCT coefficient quantization. In: The 9th AIAA Computing in Aerospace Conference, pp. 314–318 (1993)
2. Ahumada, J.A.J., Beard, B.L., Eriksson, R.: Spatio-temporal discrimination model predicts temporal masking functions. In: Proc. SPIE (1998), doi:10.1117/12.320103
3. AVS Video Expert Group, Draft of Advanced Audio Video Coding – Part 2: video, AVS_N1063 (2003)
4. AVS Video Expert Group, Information technology - Advanced coding of audio and video - Part 2: Video, GB/T 20090.2-2006 (2006)
5. Babu, R.V., Perkis, A.: An HVS-based no-reference perceptual quality assessment of JPEG coded images using neural networks. In: Proceedings of the International Conference on Image Processing, vol. 1, pp. 433–436 (2005)
6. Bjontegaard, G.: Calculation of average PSNR differences between RD curves, Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, Doc. VCEG-M33 (2001)
7. Buccigrossi, R.W., Simoncelli, E.P.: Image compression via joint statistical characterization in the wavelet domain. IEEE Transactions on Image Processing 8(12), 1688–1701 (1999)
8. Callet, P.L., Autrusseau, F.: Subjective quality assessment IRCCyN/IVC database (2005), `http://www.irccyn.ec-nantes.fr/ivcdb/`
9. Caviedes, J., Oberti, F.: A new sharpness metric based on local kurtosis, edge and energy information. Signal Processing-Image Communication 19(2), 147–161 (2004)
10. Chandler, D.M., Hemami, S.S.: A57 database, `http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html`
11. Cheng, H., Lubin, J.: Reference free objective quality metrics for MPEG coded video. In: Human Vision and Electronic Imaging X, vol. 5666, pp. 160–167 (2005)
12. Chou, C.H., Li, Y.C.: A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. IEEE Transactions on Circuits and Systems for Video Technology 5(6), 467–476 (1995)
13. Daly, S.: The Visible Differences Predictor - an Algorithm for the Assessment of Image Fidelity. In: Human Vision, Visual Processing, and Digital Display III, vol. 1666, pp. 2–15 (1992)
14. Daly, S.: Engineering observations from spatiovelocity and spatiotemporal visual models. In: Human Vision and Electronic Imaging III, vol. 3299, pp. 180–191 (1998)
15. Eskicioglu, A.M., Fisher, P.S.: Image quality measures and their performance. IEEE Transactions on Communications 43(12), 2959–2965 (1995)
16. Fan, L., Ma, S.W., Wu, F.: Overview of AVS video standard. In: Proceedings of the IEEE International Conference on Multimedia and Expo., vol. 1, pp. 423–426 (2004)
17. Foley, J.M.: Human Luminance Pattern-Vision Mechanisms - Masking Experiments Require a New Model. Journal of the Optical Society of America a-Optics Image Science and Vision 11(6), 1710–1719 (1994)
18. Girod, B.: What's Wrong With Mean Squared Error? In: Watson, A.B. (ed.) Digital Images and Human Vision. The MIT Press, Cambridge (1993)
19. Grice, J., Allebach, J.P.: The Print Quality Toolkit: An integrated print quality assessment tool. Journal of Imaging Science and Technology 43(2), 187–199 (1999)
20. Horita, Y., et al.: MICT Image Quality Evaluation Database, `http://mict.eng.u-toyama.ac.jp/mict/index2.html`

21. ISO/IEC 14496-10, Coding of audio-visual objects - Part 10: Advanced video coding. International Organization for Standardization, Geneva, Switzerland (2003)
22. ITU-T FG IPTV-ID-0082 Introductions for AVS-P2. 1st FG IPTV meeting, ITU , Geneva, Switzerland (2006)
23. ITU-R Report BT.1082-1 Studies toward the unification of picture assessment methodology. ITU, Geneva, Switzerland (1990)
24. ITU-R Recommendation BT.815-1 Specification of a signal for measurement of the contrast ratio of displays. ITU, Geneva, Switzerland (1994)
25. ITU-R Recommendation BT.710-4 Subjective assessment methods for image quality in high-definition television. ITU, Geneva, Switzerland (1998)
26. ITU-R Recommendation BT.500-11 Methodology for the subjective assessment of the quality of television pictures. ITU, Geneva, Switzerland (2002)
27. ITU-R Recommendation BT.1683 Objective perceptual video quality measurement techniques for standard definition digital broadcast television in the presence of a full reference. ITU, Geneva, Switzerland (2004)
28. ITU-R Recommendation BT.814-2 Specifications and alignment procedures for setting of brightness and contrast of displays. ITU, Geneva, Switzerland (2007)
29. ITU-T Recommendation P.910 Subjective video quality assessment methods for multimedia applications. ITU, Geneva, Switzerland (2008)
30. Kanumuri, S., et al.: Modeling packet-loss visibility in MPEG-2 video. IEEE Transactions on Multimedia 8(2), 341–355 (2006)
31. Lai, Y.K., Kuo, C.C.J.: A Haar wavelet approach to compressed image quality measurement. Journal of Visual Communication and Image Representation 11(1), 17–40 (2000)
32. Lambrecht, C.J.V.: Color moving pictures quality metric. In: Proceedings of International Conference on Image Processing, vol. I, pp. 885–888 (1996)
33. Legge, G.E., Foley, J.M.: Contrast Masking in Human-Vision. Journal of the Optical Society of America 70(12), 1458–1471 (1980)
34. Li, X.: Blind image quality assessment. In: Proceedings of the International Conference on Image Processing, vol. 1, pp. 449–452 (2002)
35. Lin, W.S.: Computational Models for Just-Noticeable Difference. In: Wu, H.R. (ed.) Digital video image quality and perceptual coding. CRC Press, Boca Raton (2005)
36. Lin, W.S.: Gauging Image and Video Quality in Industrial Applications. In: Liu, Y. (ed.), SCI, Springer, Berlin (2008)
37. Lin, W.S., Li, D., Ping, X.: Visual distortion gauge based on discrimination of noticeable contrast changes. IEEE Transactions on Circuits and Systems for Video Technology 15(7), 900–909 (2005)
38. Lu, Z.K., et al.: Perceptual quality evaluation on periodic frame-dropping video. In: Proceedings of the IEEE International Conference on Image Processing, vol. 3, pp. 433–436 (2007)
39. Lubin, J.: The use of psychophysical data and models in the analysis of display system performance. In: Watson, A.B. (ed.) Digital Images and Human Vision, The MIT Press, Cambridge (1993)
40. Lukas, F.X.J.: Picture Quality Prediction Based on a Visual Model. IEEE Transactions on Communications 30(7), 1679–1692 (1982)
41. Mannos, J.L., Sakrison, D.J.: The Effects of a Visual Fidelity Criterion on Encoding of Images. IEEE Transactions on Information Theory 20(4), 525–536 (1974)

42. Marichal, X.M., Ma, W.Y., Zhang, H.J.: Blur determination in the compressed domain using DCT information. In: Proceedings of the International Conference on Image Processing, vol. 2, pp. 386–390 (1999)
43. Marziliano, P., et al.: Perceptual blur and ringing metrics: application to JPEG2000. Signal Processing-Image Communication 19(2), 163–172 (2004)
44. Masry, M., Hemami, S.S., Sermadevi, Y.: A scalable wavelet-based video distortion metric and applications. IEEE Transactions on Circuits and Systems for Video Technology 16(2), 260–273 (2006)
45. Miyahara, M., Kotani, K., Algazi, V.R.: Objective picture quality scale (PQS) for image coding. IEEE Transactions on Communications 46(9), 1215–1226 (1998)
46. Moorthy, A.K., Bovik, A.C.: Perceptually significant spatial pooling techniques for image quality assessment. In: Proceedings of the SPIE Human Vision and Electronic Imaging XIV, vol. 7240, pp. 724012–724012 (2009)
47. Nadenau, M.J., Reichel, J., Kunt, M.: Performance comparison of masking models based on a new psychovisual test method with natural scenery stimuli. Signal Processing-Image Communication 17(10), 807–823 (2002)
48. Oguz, S.H., Hu, Y.H., Nguyen, T.Q.: Image coding ringing artifact reduction using morphological post-filtering. In: Proceedings of the IEEE Second Workshop on Multimedia Signal Processing, pp. 628–633 (1998)
49. Ong, E.P., et al.: A no-reference quality metric for measuring image blur. In: Proceedings of the Seventh International Symposium on Signal Processing and Its Applications, vol. 1, pp. 469–472 (2003)
50. Pappas, T.N., Safranek, R.J.: Perceptual criteria for image quality evaluation. In: Bovik, A.C. (ed.) Handbook of Image and Video Processing. Academic Press, Orlando (2000)
51. Parmar, M., Reeves, S.J.: A perceptually based design methodology for color filter arrays. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing III, pp. 473–476 (2004)
52. Pastrana-Vidal, R.R., et al.: Sporadic frame dropping impact on quality perception. In: Human Vision and Electronic Imaging IX, vol. 5292, pp. 182–193 (2004)
53. Pastrana-Vidal, R.R., Gicquel, J.C.: Automative quality assessment of video fluidity impairments using a no-reference metric. In: Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics (2006)
54. Peli, E.: Contrast in Complex Images. Journal of the Optical Society of America a-Optics Image Science and Vision 7(10), 2032–2040 (1990)
55. Pinson, M.H., Wolf, S.: A new standardized method for objectively measuring video quality. IEEE Transactions on Broadcasting 50(3), 312–322 (2004)
56. Ponomarenko, N., et al .: Tampere Image Database 2008 TID2008, version 1.0 (2008), `http://www.ponomarenko.info/tid2008.htm`
57. Poynton, C.: Gamma. In: Poynton, C. (ed.) A Technical Introduction to Digital Video. Wiley, New York (1996)
58. Seshadrinathan, K., Bovik, A.C.: A Structural Similarity Metric for Video Based on Motion Models. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, vol. 1, pp. 869–872 (2007)
59. Sheikh, H.R., et al.: LIVE Image Quality Assessment Database, Release 2 (2005), `http://live.ece.utexas.edu/research/quality`
60. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on Image Processing 15(2), 430–444 (2006)

61. Sheikh, H.R., Bovik, A.C., Cormack, L.: No-reference quality assessment using natural scene statistics: JPEG2000. IEEE Transactions on Image Processing 14(11), 1918–1927 (2005)
62. Sullivan, G.J., Topiwala, P.N., Luthra, A.: The H.264/AVC advanced video coding standard: Overview and introduction to the fidelity range extensions. In: Proceedings of the SPIE Applications of Digital Image Processing XXVII, vol. 5558, pp. 454–474 (2004)
63. Tan, K.T., Ghanbari, M., Pearson, D.E.: An objective measurement tool for MPEG video quality. Signal Processing 70(3), 279–294 (1998)
64. Teo, P.C., Heeger, D.J.: Perceptual Image Distortion. In: Proceedings of IEEE International Conference on Image Processing, vol. 2, pp. 982–986 (1994)
65. Teo, P.C., Heeger, D.J.: Perceptual Image Distortion. In: Human Vision, Visual Processing, and Digital Display V, vol. 2179, pp. 127–141 (1994)
66. Verscheure, O., Frossard, P., Hamdi, M.: User-Oriented QoS Analysis in MPEG-2 Video Delivery. Real-Time Imaging 5(5), 305–314 (1999)
67. VQEG, Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment II. Video Quality Expert Group (2003), `http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII/downloads/VQEGII_Final_Report.pdf` (cited August 5, 2009)
68. VQEG, RRNR-TV Group Test Plan, Version 2.0. Video Quality Expert Group (2007), `ftp://vqeg.its.bldrdoc.gov/Documents/Projects/rrnr-tv/RRNR-tv_draft_2.0_changes_accepted.doc` (cited August 5, 2009)
69. VQEG, Test Plan for Evaluation of Video Quality Models for Use with High Definition TV Content, Draft Version 3.0.Video Quality Expert Group (2009), `ftp://vqeg.its.bldrdoc.gov/Documents/Projects/hdtv/VQEG_HDTV_testplan_v3.doc` (cited August 5, 2009)
70. VQEG, Hybrid Perceptual/Bitstream Group Test Plan, Version 1.3. Video Quality Expert Group (2009), `ftp://vqeg.its.bldrdoc.gov/Documents/Projects/hybrid/VQEG_hybrid_testplan_v1_3_changes_highlighted.doc` (cited August 5, 2009)
71. Vlachos, T.: Detection of blocking artifacts in compressed video. Electronics Letters 36(13), 1106–1108 (2000)
72. Wang, X.F., Zhao, D.B.: Performance comparison of AVS and H. 264/AVC video coding standards. Journal of Computer Science and Technology 21(3), 310–314 (2006)
73. Wang, Z., Bovik, A.C.: A universal image quality index. IEEE Signal Processing Letters 9(3), 81–84 (2002)
74. Wang, Z., Shang, X.L.: Spatial pooling strategies for perceptual image quality assessment. In: Proceedings of the International Conference on Image Processing, October7-10, vol. 1, pp. 2945–2948 (2006)
75. Wang, Z., Simoncelli, E.P.: Local phase coherence and the perception of blur. Advances in Neural Information Processing Systems 16, 1435–1442 (2004)
76. Wang, Z., Bovik, A.C., Evans, B.L.: Blind measurement of blocking artifacts in images. In: Proceedings of the International Conference on Image Processing, vol. 3, pp. 981–984 (2000)
77. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004)

78. Wang, Z., Lu, L., Bovik, A.C.: Video quality assessment based on structural distortion measurement. Signal Processing: Image Communication 19(2), 121–132 (2004)
79. Watson, A.B.: The Cortex Transform - Rapid Computation of Simulated Neural Images. In: Computer Vision Graphics and Image Processing, vol. 39(3), pp. 311–327 (1987)
80. Watson, A.B.: DCTune: A technique for visual optimization of DCT quantization matrices for individual images. In: Proc. Soc. Information Display Dig. Tech. Papers XXIV, pp. 946–949 (1993)
81. Watson, A.B., Solomon, J.A.: Model of visual contrast gain control and pattern masking. Journal of the Optical Society of America a-Optics Image Science and Vision 14(9), 2379–2391 (1997)
82. Watson, A.B., Borthwick, R., Taylor, M.: Image quality and entropy masking. In: Proc. SPIE (1997), doi:10.1117/12.274501
83. Watson, A.B., Hu, J., McGowan, J.F.: Digital video quality metric based on human vision. Journal of Electronic Imaging 10(1), 20–29 (2001)
84. Winkler, S.: A perceptual distortion metric for digital color video. In: Human Vision and Electronic Imaging IV, vol. 3644, pp. 175–184 (1999)
85. Winkler, S.: Issues in vision modeling for perceptual video quality assessment. Signal Processing 78(2), 231–252 (1999)
86. Winkler, S.: Metric evaluation. In: Winkler, S. (ed.) Digital video quality: vision models and metrics. Wiley, New York (2005)
87. Winkler, S.: Vision. In: Winkler, S. (ed.) Digital video quality: vision models and metrics. Wiley, New York (2005)
88. Winkler, S.: Digital video quality: vision models and metrics. Wiley, New York (2005)
89. Winkler, S.: Perceptual video quality metrics - a review. In: Wu, H.R. (ed.) Digital video image quality and perceptual coding. CRC Press, Boca Raton (2005)
90. Winkler, S., Mohandas, P.: The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics. IEEE Transactions on Broadcasting 54(3), 660–668 (2008)
91. Wu, H.R., Yuen, M.: A generalized block-edge impairment metric for video coding. IEEE Signal Processing Letters 4(11), 317–320 (1997)
92. Yang, K.C., et al.: Perceptual temporal quality metric for compressed video. IEEE Transactions on Multimedia 9(7), 1528–1535 (2007)
93. Yu, L., et al.: Overview of AVS-Video: Tools, performance and complexity. In: Proceedings of the SPIE Visual Communications and Image Processing, vol. 5960, pp. 679–690 (2005)
94. Yu, Z.H., et al.: Vision-model-based impairment metric to evaluate blocking artifacts in digital video. Proceedings of the IEEE 90(1), 154–169 (2002)
95. Zhai, G.T., et al.: No-reference noticeable blockiness estimation in images. Signal Processing-Image Communication 23(6), 417–432 (2008)