Available online at www.sciencedirect.com
SCIENCE DIRECT®
Signal Processing: *Image Communication* 19 (2004) 121–132
SIGNAL PROCESSING: IMAGE COMMUNICATION
www.elsevier.com/locate/image

Video quality assessment based on structural distortion measurement

Zhou Wang[a,b,*], Ligang Lu[c], Alan C. Bovik[d]

[a] *Laboratory for Computational Vision (LCV), Center for Neural Science and Courant Institute of Mathematical Sciences, New York University, New York, USA*
[b] *Howard Hughes Medical Institute, New York, USA*
[c] *Multimedia Technologies, IBM T.J. Watson Research Center, Yorktown Heights, New York, USA*
[d] *Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, The University of Texas, Austin, TX, USA*

Abstract

Objective image and video quality measures play important roles in a variety of image and video processing applications, such as compression, communication, printing, analysis, registration, restoration, enhancement and watermarking. Most proposed quality assessment approaches in the literature are error sensitivity-based methods. In this paper, we follow a new philosophy in designing image and video quality metrics, which uses structural distortion as an estimate of perceived visual distortion. A computationally efficient approach is developed for full-reference (FR) video quality assessment. The algorithm is tested on the video quality experts group Phase I FR-TV test data set.

Keywords: Image quality assessment; Video quality assessment; Human visual system; Error sensitivity; Structural distortion; Video quality experts group (VQEG)

1. Introduction

There has been an increasing need recently to develop objective quality measurement techniques that can predict perceived image and video quality automatically. These methods are useful in a variety of image and video processing applications, such as compression, communication, printing, displaying, analysis, registration, restoration, enhancement and watermarking. Generally speaking, these methods can be employed in three ways. First, they can be used to monitor image/video quality for quality control systems. Second, they can be employed to benchmark image/video processing systems and algorithms. Third, they can also be embedded into image/video processing systems to optimize algorithms and parameter settings.

Currently, the most commonly used full-reference (FR) objective image and video distortion/quality metrics are mean squared error (MSE) and peak signal-to-noise ratio (PSNR). MSE and PSNR are widely used because they are simple to

*Corresponding author. 4 Washington Place Room 809, New York 10003, USA. Tel.: +1-212-992-8750.
E-mail addresses: zhouwang@ieee.org (Z. Wang), lul@us.ibm.com (L. Lu), bovik@ece.utexas.edu (A.C. Bovik).

0923-5965/$ - see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/S0923-5965(03)00076-6

calculate, have clear physical meanings, and are mathematically easy to deal with for optimization purposes. However, they have been widely criticized as well for not correlating well with perceived quality measurement [8,10,19,22,23,25,26,28]. In the last three decades, a great deal of effort has been made to develop objective image and video quality assessment methods, which incorporate perceptual quality measures by considering human visual system (HVS) characteristics. Some of the developed models are commercially available. The video quality experts group (VQEG) was formed to develop, validate and standardize new objective measurement methods for video quality. Although the Phase I test [4,20] for FR television video quality assessment only achieved limited success, VQEG continues its work on Phase II test for FR quality assessment for television, and reduced-reference (RR) and no-reference (NR) quality assessment for television and multimedia.

It is worth noting that many of the proposed objective image/video quality assessment approaches in the literature share a common error sensitivity-based philosophy [22,26,28], which is motivated from psychophysical and physiological vision research. The basic principle is to think of the distorted signal being evaluated as the sum of a perfect quality reference signal and an error signal. The task of perceptual image quality assessment is then to evaluate how strong the error signal is perceived by the HVS according to the characteristics of human visual error sensitivity.

A general framework following error sensitivity-based philosophy is shown in Fig. 1 [28]. First, the original and distorted image/video signals are subject to preprocessing procedures, possibly including alignment, transformations of color spaces, calibration for display devices, point spread function (PSF) filtering that simulates the eye optics, and light adaptation. Next, a contrast sensitivity function (CSF) filtering procedure may be applied, where the CSF models the variation in the sensitivity of the HVS to different spatial and temporal frequencies. The CSF feature may be implemented before the channel decomposition module using linear filters that approximate the frequency responses of the CSF. It may also be considered as a normalization factor between channels after channel decomposition. The channel decomposition process transforms the signals into different spatial and temporal frequency as well as orientation selective subbands. A number of channel decomposition methods that attempt to model the neuron responses in the primary visual cortex have been used [5,11,13,18,19,29]. Some quality assessment metrics use much simpler transforms such as the discrete cosine transform (DCT) [30,32] and the separable wavelet transforms [1,12,34], and still achieved comparable results. Channel decompositions tuned to various temporal frequencies have also been reported [2,35]. The errors calculated in each channel are adjusted according to the "base sensitivity" for each channel (related to the CSF) as a normalization process. They are also adjusted according to a spatially varying masking factor, which refers to the fact that the presence of one image component reduces the visibility of another image component spatially and temporally proximate. Both intra- and inter-channel masking effects may be considered. Finally, the error pooling module combines the error signals in different channels into a single distortion/quality value, where most quality assessment methods take the form of the Minkowski error metric [15]. The overall framework covers MSE as the simplest special case (with identity preprocessing, no CSF filtering, identity transform, constant error adjustment

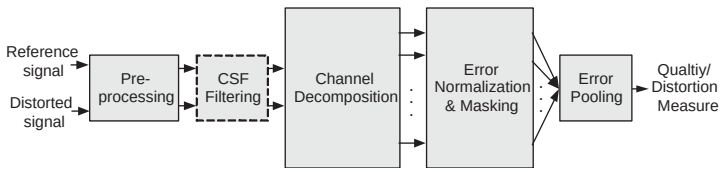

Fig. 1. Framework of error sensitivity-based quality assessment system. Note: the CSF feature can be implemented either as "CSF Filtering" or within "Error Normalization".

and L_2 Minkowski error pooling). A *perceptual* image quality metric may implement one or more perceptually meaningful components of the system. Reviews on perceptual error sensitivity-based models can be found in [6,15,28].

The visual error sensitivity-based algorithms attempt to predict perceived errors by simulating the perceptual quality-related functional components of the HVS. However, the HVS is an extremely complicated system and the current understanding of the HVS is limited. Currently, several issues that are critical for justifying the general framework are still under investigations [28].

The "suprathreshold problem" is one issue that has not been well understood: note that most psychophysical subjective experiments are conducted near the threshold of visibility. These measured threshold values are then used to define visual error sensitivity models, such as the CSF and the various masking effect models. However, current psychophysical studies are still not sufficient to determine whether such near-threshold models can be generalized to characterize perceptual distortions significantly larger than threshold levels, as is the case in a majority of image processing situations. Another question is: when the errors are much larger than the thresholds, can the relative errors between different channels be normalized using the visibility thresholds? Recent efforts have been made to incorporate suprathreshold psychophysics research for analyzing image distortions (e.g., [3,9,16,33,36]). It remains to be seen how much these efforts can improve the performance of the current quality assessment algorithms.

The "natural image complexity problem" is another important issue: the CSF and various masking models are established based on psychovisual experiments conducted using one or a few relatively simple patterns, such as bars, sinusoidal gratings and Gabor patches. But all such patterns are much simpler than real world images, which can be thought of as a superposition of a much larger number of simple patterns. Can we generalize the model for the interactions between a few simple patterns to model the interactions between tens or hundreds of patterns? Are these simple-pattern experiments sufficient to build a model that can predict the quality of complex-structured natural images? Although the answers to these questions are currently not known, the recently established Modelfest dataset [31] includes both simple and complicated patterns, and should facilitate future studies.

Motivated from a substantially different philosophy, a simple structural distortion-based method is proposed for still image quality assessment in [22,23,25]. In this paper, an improved version of the algorithm is employed for video quality assessment. In Section 2, the general philosophy and a specific implementation of the structural distortion-based method are presented. A new video quality assessment system is introduced in Section 3. The system is tested on the VQEG Phase I FR-TV video dataset. Finally, Section 4 draws conclusions and provides further discussions.

2. Structural distortion-based method

2.1. The new philosophy

Natural image signals are highly structured. By "structured signal", we mean that the samples of the signals have strong dependencies between each other, especially when they are close in space. However, the Minkowski error pooling formula used in the error-sensitivity-based method is in the form of pointwise signal differencing, which is independent of the signal structure. Furthermore, decomposing the signals using linear transformations still cannot remove the strong dependencies between the samples of the signals. Therefore, a significant amount of signal structural changes still cannot be captured with the Minkowski metric. An interesting recent trend is to design optimal transformation and masking models that can reduce both statistical and perceptual dependencies (e.g., [7,14]). However, these models significantly complicate the system.

The motivation of our new approach is to find a more direct way to compare the structures of the reference and the distorted signals. In [22,26], a new philosophy in designing image and video

quality metrics was introduced:

The main function of the human visual system is to extract structural information from the viewing field, and the human visual system is highly adapted for this purpose. Therefore, a measurement of structural distortion should be a good approximation of perceived image distortion.

A major difference of the new philosophy from the error-sensitivity-based philosophy is that image degradations are considered as *perceived structural information loss* instead of *perceived errors*. A motivating example is shown in Fig. 2, where the original "Goldhill" image is distorted with global contrast suppression, JPEG compression, and blurring. We tuned all the distorted images to yield a similar MSE relative to the original one. Interestingly, the distorted images exhibit drastically different visual qualities. This can be easily understood with the new philosophy by examining how the image structures are preserved in the distorted images. In the JPEG compressed and blurred images, hardly any detailed structures of the original image can be observed. By contrast, almost all the image structures of the reference image are very well preserved in the contrast-suppressed image. In fact, the original information can be nearly fully recovered via a simple pointwise inverse linear intensity transformation.

Another important difference is that the new philosophy is a *top-down* approach—simulating the hypothesized functionality of the overall HVS,

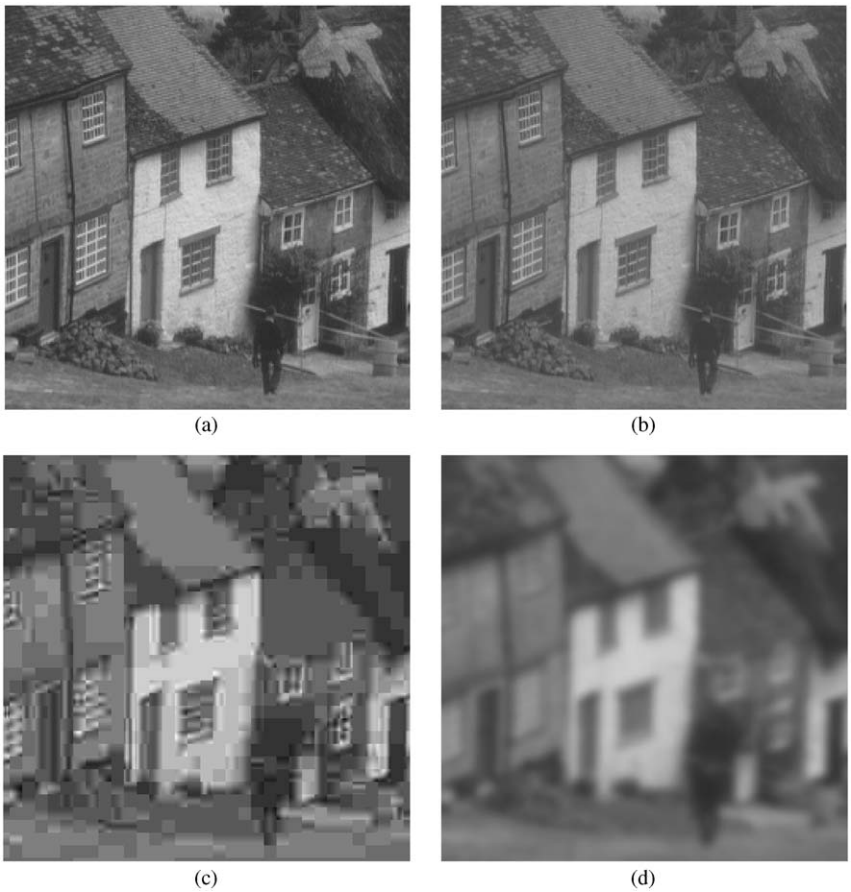

Fig. 2. (a) Original "Goldhill" image (cropped from 512×512 to 256×256 for visibility); (b) contrast suppressed image; (c) JPEG compressed image; (d) blurred image.

while the error-sensitivity-based philosophy uses a *bottom-up* approach—simulating the function of each relevant component in the HVS and combine them together.

It needs to be mentioned that the new philosophy does not intend to solve the problems of the error-sensitivity-based paradigm (e.g., the "suprathreshold" problem and the "natural image complexity" problem mentioned above). Instead, we consider it more as an alternative to avoid the problems (though might not be completely). For example, it suggests not to predict image quality by accumulating simple pattern differences, thus somehow avoids the "natural image complexity" problem. Also, since the quantization of perceived distortions does not rely on threshold psychophysics, the "suprathreshold" problem is avoided.

2.2. The structural similarity (SSIM) index

There may be different implementations of the new philosophy, depending on how the concepts of "structural information" and "structural distortion" are interpreted and quantified. Here, from an image formation point of view, we consider the "structural information" in an image as those attributes that reflect the structure of the objects in the scene, which is independent of the average luminance and contrast of the image. This leads to an image quality assessment approach that separates the measurement of luminance, contrast and structural distortions.

In [22,25], a simple image similarity indexing algorithm was proposed. Let $\mathbf{x}$ and $\mathbf{y}$ be two nonnegative signals that have been aligned with each other (e.g., two image patches extracted from the same spatial location from two images being compared, respectively), and let μ_x, μ_y, σ_x^2, σ_y^2 and σ_{xy} be the mean of $\mathbf{x}$, the mean of $\mathbf{y}$, the variance of $\mathbf{x}$, the variance of $\mathbf{y}$, and the covariance of $\mathbf{x}$ and $\mathbf{y}$, respectively. Here the mean and the standard deviation (square root of the variance) of a signal are roughly considered as estimates of the luminance and the contrast of the signal. The covariance (normalized by the variance) can be thought of as a measurement of how much one signal is changed nonlinearly to the other signal being compared. We define the luminance, contrast and structure comparison measures as follows:

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

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

$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy}}{\sigma_x\sigma_y}. \tag{1}$$

Notice that these terms are conceptually independent in the sense that the first two terms only depend on the luminance and the contrast of the two images being compared, respectively, and purely changing the luminance or the contrast of either image has no impact on the third term. Geometrically, $s(\mathbf{x},\mathbf{y})$ corresponds to the cosine of the angle between the vectors $\mathbf{x}-\mu_x$ and $\mathbf{y}-\mu_y$, independent of the lengths of these vectors. Although $s(\mathbf{x},\mathbf{y})$ does not use a direct descriptive representation of the image structures, it reflects the similarity between two image structures—it equals one if and only if the structures of the two image signals being compared are exactly the same (recall that we consider structural information as those image attributes other than the luminance and contrast information).

When $(\mu_x^2+\mu_y^2)(\sigma_x^2+\sigma_y^2)\neq 0$, the similarity index measure between $\mathbf{x}$ and $\mathbf{y}$ given in [22,25] corresponds to

$$\begin{aligned} S(\mathbf{x},\mathbf{y}) &= l(\mathbf{x},\mathbf{y})\cdot c(\mathbf{x},\mathbf{y})\cdot s(\mathbf{x},\mathbf{y}) \\ &= \frac{4\mu_x\mu_y\sigma_{xy}}{(\mu_x^2+\mu_y^2)(\sigma_x^2+\sigma_y^2)}. \end{aligned} \tag{2}$$

If the two signals are represented discretely as $\mathbf{x}=\{x_i|i=1,2,\ldots,N\}$ and $\mathbf{y}=\{y_i|i=1,2,\ldots,N\}$, then the statistical features can be estimated as follows:

$$\mu_x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \mu_y = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \tag{3}$$

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2,$$

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\bar{y})^2, \tag{4}$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}). \tag{5}$$

One problem with (2) is that when $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ is close to 0, the resulting measurement is unstable. This effect has been frequently observed in our experiments, especially over flat regions in images. In order to avoid this problem, we have modified Eq. (2). The resulting new measure is named the Structural SIMilarity (SSIM) index between signals $\mathbf{x}$ and $\mathbf{y}$:

$$\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \quad (6)$$

Two constants, C_1 and C_2, are added which are given by

$$C_1 = (K_1 L)^2 \quad \text{and} \quad C_2 = (K_2 L)^2, \quad (7)$$

where L is the dynamic range of the pixel values (for 8 bits/pixel gray scale images, $L = 255$), and K_1 and K_2 are two constants whose values must be small such that C_1 or C_2 will take effect only when $(\mu_x^2 + \mu_y^2)$ or $(\sigma_x^2 + \sigma_y^2)$ is small. Throughout the experiments in this paper, we set $K_1 = 0.01$ and $K_2 = 0.03$, respectively.

The SSIM index satisfies the following conditions:

1. $\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = \mathrm{SSIM}(\mathbf{y},\mathbf{x})$;
2. $\mathrm{SSIM}(\mathbf{x},\mathbf{y}) \leqslant 1$;
3. $\mathrm{SSIM}(\mathbf{x},\mathbf{y}) = 1$ if and only if $\mathbf{x} = \mathbf{y}$ (in discrete representations, $x_i = y_i$ for all $i = 1, 2, \ldots, N$).

Based on the philosophy described before, if we consider one of the image signals being compared to have perfect quality, then the SSIM index provides a quantitative measurement of the quality of the other image signal.

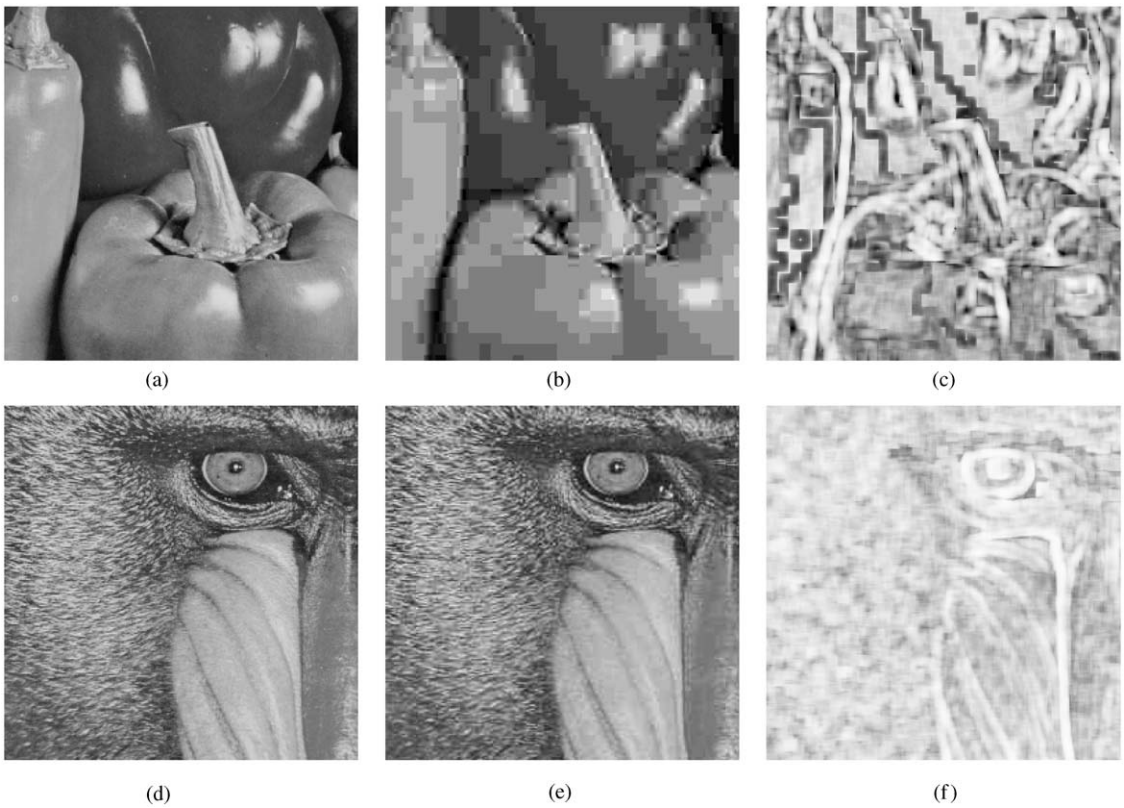

Fig. 3. (a) and (d): Original "Peppers" and "Mandrill" images (cropped from 512×512 to 256×256 for visibility); (b) JPEG compressed "Peppers" image, MSE = 160, MSSIM = 0.6836; (c) SSIM index map of the JPEG compressed "Peppers" image; (e) JPEG compressed "Mandrill" image, MSE = 159, MSSIM = 0.8477; (f) SSIM index map of the JPEG compressed "Mandrill" image. In (c) and (f), brightness indicates the magnitude of the local SSIM index value.

The SSIM indexing algorithm is applied for quality assessment of still images using a sliding window approach. The window size is fixed to be 8×8 in this paper. The SSIM indices are calculated within the sliding window, which moves pixel-by-pixel from the top-left to the bottom-right corner of the image. This results in a SSIM index map of the image, which is also considered as the quality map of the distorted image being evaluated. The overall quality value is defined as the average of the quality map, or, equivalently, the mean SSIM (MSSIM) index. A Matlab implementation of the SSIM index algorithm is available online at [24]. The SSIM index maps of two JPEG compressed images are shown in Fig. 3. The MSSIM values of the test images Fig. 2(b), (c) and (d) are 0.9622, 0.6451 and 0.6430, respectively, which appear to have better consistency with perceived image quality than MSE. While Fig. 2 is a good example for testing the cross-distortion capability of image quality assessment methods, Fig. 3 is a simple test for the cross-image capability of image quality assessment algorithms. Again, the images with similar MSE have significantly different visual quality, and MSSIM delivers better consistency with perceptual evaluations. More demonstrative images are available online at [23].

3. Video quality assessment

In [27], a hybrid video quality assessment method was developed, where the proposed quality indexing approach (with $C_1 = C_2 = 0$) was combined with blocking and blurring measures as well as a texture classification algorithm. In this paper, we attempt to use a much simpler method, which employs the SSIM index as a single measure for various types of distortions.

3.1. Video quality assessment algorithm

The diagram of the proposed video quality assessment system is shown in Fig. 4. The quality of the distorted video is measured in three levels: the local region level, the frame level, and the sequence level.

First, local sampling areas are extracted from the corresponding frame and spatial locations in the original and the distorted video sequences, respectively. The sampling areas are randomly selected 8×8 windows. This is different from the method used for still images in Section 2, where all possible sampling windows are selected since the sliding window moves pixel-by-pixel across the whole image. Instead, only a proportion of all possible 8×8 windows are selected here. We use the number of sampling windows per video frame (R_s) to represent the sampling density. Our experiments show that a properly selected R_s can largely reduce computational cost while still maintains reasonably robust measurement results. The SSIM indexing approach is then applied to the Y, Cb and Cr color components independently and combined into a local quality measure using a weighted summation. Let $\mathrm{SSIM}_{ij}^{\mathrm{Y}}$, $\mathrm{SSIM}_{ij}^{\mathrm{Cb}}$ and $\mathrm{SSIM}_{ij}^{\mathrm{Cr}}$ denote the SSIM index values of the Y, Cb and Cr components of the jth sampling window in the ith video frame, respectively. The local quality index is given by

$$\mathrm{SSIM}_{ij} = W_{\mathrm{Y}}\mathrm{SSIM}_{ij}^{\mathrm{Y}} + W_{\mathrm{Cb}}\mathrm{SSIM}_{ij}^{\mathrm{Cb}} + W_{\mathrm{Cr}}\mathrm{SSIM}_{ij}^{\mathrm{Cr}}, \quad (8)$$

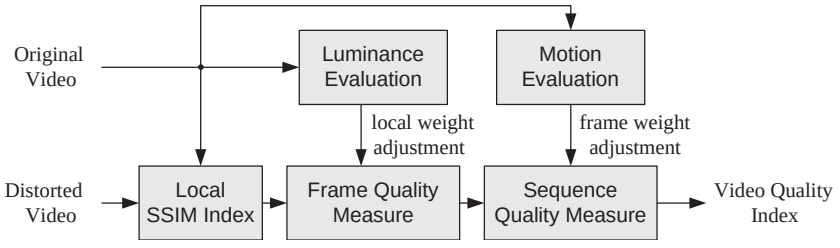

Fig. 4. Proposed video quality assessment system.

where the weights are fixed in our experiments to be $W_Y = 0.8$, $W_{Cb} = 0.1$ and $W_{Cr} = 0.1$, respectively.

In the second level of quality evaluation, the local quality values are combined into a frame-level quality index using

$$Q_i = \frac{\sum_{j=1}^{R_s} w_{ij} SSIM_{ij}}{\sum_{j=1}^{R_s} w_{ij}}, \tag{9}$$

where Q_i denotes the quality index measure of the ith frame in the video sequence, and w_{ij} is the weighting value given to the jth sampling window in the ith frame.

Finally in the third level, the overall quality of the entire video sequence is given by

$$Q = \frac{\sum_{i=1}^{F} W_i Q_i}{\sum_{i=1}^{F} W_i}, \tag{10}$$

where F is the number of frames and W_i is the weighting value assigned to the ith frame.

If all the frames and all the sampling windows in every frame are considered equally then

$$w_{ij} = 1 \quad \text{for all } i,j \quad \text{and}$$

$$W_i = \sum_{j=1}^{R_s} w_{ij} = R_s \quad \text{for all } i. \tag{11}$$

This leads to a quality measure equaling the average SSIM index measurement of all sampling windows in all frames. Such a weighting assignment method may not be optimal because different regions and different frames may be of different importance to the human observers. Optimal weighting assignment is difficult because many psychological aspects are involved, which may depend on the content and context of the video sequence being observed. However, certain appropriate adjustments around the selection of all-equal-weighting may help to improve the prediction accuracy of the quality assessment algorithm.

In this paper, two simple adjustment methods are employed. The first is based on the observation that dark regions usually do not attract fixations, therefore should be assigned smaller weighting values. We use the mean value μ_x (as given in (3)) of the Y component as an estimate of the local luminance, and the local weighting is adjusted as

$$w_{ij} = \begin{cases} 0 & \mu_x \leqslant 40, \\ (\mu_x - 40)/10 & 40 < \mu_x \leqslant 50, \\ 1 & \mu_x > 50. \end{cases} \tag{12}$$

The second adjustment considers the case when very large global motion occurs. Note that some image distortions are perceived differently when the background of the video is moving very fast (usually corresponds to high-speed camera movement). For example, severe blurring is usually perceived as a very unpleasant type of distortion in still images or slowly moving video. However, the same amount of blur may not be as important in a large motion frame, perhaps because large perceptual motion blur occurs at the same time. Such kind of differences cannot be captured by the intra-frame SSIM index, which does not involve any motion information. Our experiments also indicate that the proposed algorithm performs less stable when very large global motion occurs. Therefore, we give smaller weighting to the large motion frames to improve the robustness of the algorithm. First, for each sampling window, we use a block-based motion estimation algorithm to evaluate its motion with respect to its adjacent next frame. Suppose m_{ij} represents the motion vector length of the jth sampling window in the ith frame, then the motion level of the ith frame is estimated as

$$M_i = \frac{(\sum_{j=1}^{R_s} m_{ij})/R_s}{K_M}, \tag{13}$$

where K_M is a constant that serves as a normalization factor of the frame motion level. We use $K_M = 16$ in our experiment. The weighting of frame is then adjusted by

$$W_i = \begin{cases} \sum_{j=1}^{R_s} w_{ij} & M_i \leqslant 0.8, \\ ((1.2 - M_i)/0.4) \sum_{j=1}^{R_s} w_{ij} & 0.8 < M_i \leqslant 1.2, \\ 0 & M_i > 1.2. \end{cases} \tag{14}$$

3.2. Test with VQEG data set

The VQEG Phase I test data set for FR-TV video quality assessment [21] is used to test the

system. We follow the performance evaluation procedures employed in the VQEG Phase I FR-TV test [20] to provide quantitative measures on the performance of the objective quality assessment models. Four metrics are employed. First, logistic functions are used in a fitting procedure to provide a nonlinear mapping between the objective/subjective scores. In [20], Metric 1 is the correlation coefficient between objective/subjective scores after variance-weighted regression analysis. Metric 2 is the correlation coefficient between objective/subjective scores after nonlinear regression analysis. These two metrics combined, provide an evaluation of *prediction accuracy*. The third metric is the Spearman rank-order correlation coefficient between the objective/subjective scores. It is considered as a measure of *prediction monotonicity*. Finally, Metric 4 is the outlier ratio (percentage of the number of predictions outside the range of ± 2 times of the standard deviations) of the predictions after the nonlinear mapping, which is a measure of *prediction consistency*. For more details about these metrics, readers can refer to [4,20].

Figs. 5(a)–(d) show the scatter plots of the subjective/objective comparisons on all test video

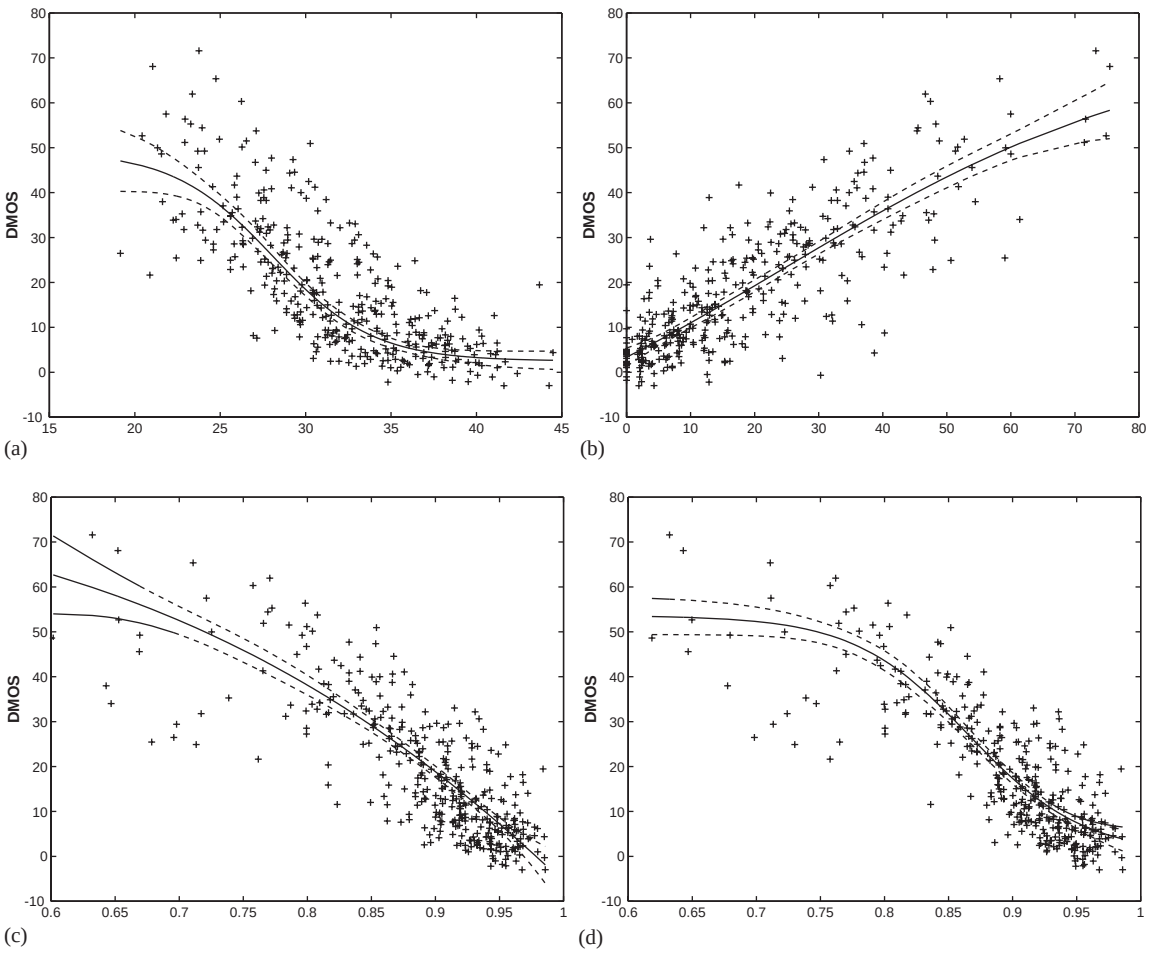

Fig. 5. Scatter plot comparison of different video quality assessment models on VQEG Phase I test dataset (all test video sequences included). Vertical and horizontal axes are for subjective and objective measurements, respectively. Each sample point represents one test video sequence: (a) PSNR, (b) KPN/Swisscom CT model, (c) proposed method (without weighting adjustment) and (d) proposed method (with weighting adjustment).

Table 1
Performance comparison of video quality assessment models on VQEG Phase I test data set (all test video sequences included)

Model	Metric 1	Metric 2	Metric 3	Metric 4
P0 (PSNR)	0.804	0.779	0.786	0.678
P1 (CPqD)	0.777	0.794	0.781	0.650
P2 (Tektronix/Sarnoff)	0.792	0.805	0.792	0.656
P3 (NHK/Mitsubishi)	0.726	0.751	0.718	0.725
P4 (KDD)	0.622	0.624	0.645	0.703
P5 (EPFL)	0.778	0.777	0.784	0.611
P6 (TAPESTRIES)	0.277	0.310	0.248	0.844
P7 (NASA)	0.792	0.770	0.786	0.636
P8 (KPN/Swisscom CT)	0.845	0.827	0.803	0.578
P9 (NTIA)	0.781	0.782	0.775	0.711
Proposed (w/o weighting adjustment)	0.830	0.820	0.788	0.597
Proposed (w/ weighting adjustment)	0.864	0.849	0.812	0.578

Metric 1: variance-weighted regression correlation coefficient; Metric 2: nonlinear regression correlation coefficient; Metric 3: Spearman rank order correlation coefficient; Metric 4: outlier ratio; P0–P9: the VQEG proponents [20]. Data for P0–P9 is from [20].

sequences given by PSNR, the KPN/Swisscom CT model (the best VQEG Phase I proponent in terms of the performance measurement used in the VQEG Phase I test when all test video sequences are included), the proposed method without any weighting adjustment (Eq. (11)), and the proposed method with weighting adjustment (Eqs. (12) and (14)), respectively. In Table 1, we give the comparison results of the four metrics when all the test video sequences are included. Despite its simplicity, the proposed method without any weighting adjustment provides reasonably good results compared with the other approaches. The proposed method with weighting adjustment performs better than all the other models.

Note that the proposed algorithm was developed after the VQEG Phase I test and after the test video sequences and subjective data became available to the public. Many of the proponents that attended the test have improved their algorithms after the test. Therefore, the experimental comparison given in this paper only suggests the potential of the proposed method to compete against the VQEG Phase I proponents. Further and broader comparisons are still needed to draw solid conclusions about the relative merits and demerits of the proposed method against the other models.

It also needs to be mentioned that the parameters of the algorithm were selected empirically, without any optimization process for a specific data base. We believe that careful selection of the parameters via psychophysical studies should be helpful in improving the performance of the algorithm. However, since the proposed algorithm is intended for general purpose video quality assessment, it is not preferable to train the parameters for any specific data base that does not cover a very wide range of image and distortion types. On the other hand, we observe in our experiments that the proposed algorithm is in general insensitive to these parameters. For example, setting the color weighting parameters W_Y to one and W_{Cb} and W_{Cr} to zero in (8) (in other words, only the luminance channel is used for video quality assessment) does not have significant effect on the overall performance of the algorithm on the VQEG data set. For another example, the same set of parameters of K_1 and K_2 in (7) were used to test LIVE image quality database [17] composed of 344 subject-rated JPEG and JPEG2000 compressed images with a wide range of compression bit rates, and the proposed algorithm outperforms PSNR by a clear margin.

4. Conclusions and discussions

We designed a new objective video quality assessment system. The key feature of the proposed method is the use of structural distortion instead

of error-sensitivity-based measurement for quality evaluation. Experiments on VQEG FR-TV Phase I test data set show that it has good correlation with perceived video quality.

One of the most attractive features of the proposed method is perhaps its simplicity. Note that no complicated procedures (such as spatial and temporal filtering, linear transformations, object segmentation, texture classification, blur evaluation, and blockiness estimation) are involved. This implies that the SSIM index is a simple formula that inherently has effective normalization power for various types of image structures and distortions. The simplicity of the algorithm also makes real-time implementation easy. In addition, the speed of the algorithm can be further adjusted by tuning the parameter of frame sampling rate R_s. Our experiments show that reasonably robust performance can be obtained with a relatively small sampling rate (e.g., $R_s < 100$), allowing real-time software implementation on moderate speed computers.

The proposed method has been found to be consistent with many observations of HVS behaviors. For example, the blocking artifact in JPEG compressed images may significantly impair the "structure" in smooth image regions, but is less disturbing in highly textured regions. This is captured very well in the quality maps in Fig. 3. However, there are other HVS characteristics that may not be well understood with the proposed method. For example, vertical distortions may appear more significant than horizontal distortions. It remains a problem that how to systematically connect and adjust the proposed quality index in accordance with psychophysical and physiological HVS studies.

In order to improve the proposed algorithm, many other issues also need further investigations in the future. One important issue is related to motion. The current SSIM index is oriented for comparison of still image structures. Notice that there are several significant outliers in the scatter plots of the proposed algorithms (in the lower-left parts of Figs. 5(c) and (d), where the models give much lower scores than they should supply). In fact, most of these significant outliers corresponds to the video sequences with large global motions (such as SRC5, SRC9 and SRC19 in the VQEG Phase I test data set). So far, no method has been found to naturally incorporate motion information into the SSIM index measure. We have attempted to apply the same SSIM index measure as in (6) for three-dimensional windows (instead of the current intra-frame two-dimensional windows). Unfortunately, no significant improvement has been observed. Another issue is regarding the case of burst-of-error. For example, when most of the frames in a video sequence have high quality, but only a few are damaged and have extremely low quality, the human observers tend to give a lower quality score than averaging all the frames. To solve this problem, a nonlinear pooling method (instead of weighted summation used in this paper) may need to be applied. Furthermore, how to measure and incorporate color distortions also needs more investigations.

Acknowledgements

The authors would like to thank Dr. Eero Simoncelli, Mr. Hamid Sheikh and Dr. Jesus Malo for helpful discussions, and thank Dr. Philip Corriveau and Dr. John Libert for providing the Matlab routines used in VQEG FR-TV Phase I test for the regression analysis of subjective/objective data comparison.

References

[1] A.P. Bradley, A wavelet visible difference predictor, IEEE Trans. Image Process. 5 (May 1999) 717–730.

[2] C.J. van den Branden Lambrecht, O. Verscheure, Perceptual quality measure using a spatio-temporal model of the human visual system, in: Proceedings of the SPIE, Vol. 2668, 1996, pp. 450–461.

[3] D.M. Chandler, S.S. Hemami, Additivity models for suprathreshold distortion in quantized wavelet-coded images, in: Human Vision and Electronic Imaging VII, Proceedings of the SPIE, Vol. 4662, January 2002, pp. 742–753.

[4] P. Corriveau, et al., Video quality experts group: current results and future directions, Proc. SPIE Visual Comm. Image Process. 4067 (June 2000).

[5] S. Daly, The visible difference predictor: an algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), Digital Images and Human Vision, The MIT Press, Cambridge, MA, 1993, pp. 179–206.

[6] M.P. Eckert, A.P. Bradley, Perceptual quality metrics applied to still image compression, Signal Processing 70 (November 1998) 177–200.
[7] I. Epifanio, J. Gutirrez, J. Malo, Linear transform for simultaneous diagonalization of covariance and perceptual metric matrix in image coding, Pattern Recognition 36 (August 2003) 1679–1923.
[8] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Comm. 43 (December 1995) 2959–2965.
[9] D.R. Fuhrmann, J.A. Baro, J.R. Cox Jr., Experimental evaluation of psychophysical distortion metrics for JPEG-encoded images, J. Electron. Imaging 4 (October 1995) 397–406.
[10] B. Girod, What's wrong with mean-squared error, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 207–220.
[11] D.J. Heeger, P.C. Teo, A model of perceptual image fidelity, in: Proceedings of the IEEE International Conference on Image Processing, 1995, pp. 343–345.
[12] Y.K. Lai, C.-C.J. Kuo, A Haar wavelet approach to compressed image quality measurement, J. Visual Comm. Image Representat. 11 (March 2000) 17–40.
[13] J. Lubin, The use of psychophysical data and models in the analysis of display system performance, in: A.B. Watson (Ed.), Digital Images and Human Vision, The MIT Press, Cambridge, MA, 1993, pp. 163–178.
[14] J. Malo, R. Navarro, I. Epifanio, F. Ferri, J.M. Artifas, Non-linear Invertible Representation for Joint Statistical and Perceptual Feature Decorrelation, Lecture Notes in Computer Science, Vol. 1876, Springer, Berlin, 2000, pp. 658–667.
[15] T.N. Pappas, R.J. Safranek, Perceptual criteria for image quality evaluation, in: A. Bovik (Ed.), Handbook of Image and Video Processing, Academic Press, New York, 2000.
[16] J.G. Ramos, S.S. Hemami, Suprathreshold wavelet coefficient quantization in complex stimuli: psychophysical evaluation and analysis, J. Opt. Soc. Am. A 18 (2001) 2385–2397.
[17] H.R. Sheikh, Z. Wang, A.C. Bovik, L.K. Cormack, Image and video quality assessment research at LIVE, http://live.ece.utexas.edu/research/quality/.
[18] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, D.J. Heeger, Shiftable multi-scale transforms, IEEE Trans. Inf. Theory 38 (1992) 587–607.
[19] P.C. Teo, D.J. Heeger, Perceptual image distortion, in: Proceedings of the SPIE, Vol. 2179, 1994, pp. 127–141.
[20] VQEG, Final report from the video quality experts group on the validation of objective models of video quality assessment, March 2000. http://www.vqeg.org/.
[21] VQEG: The Video Quality Experts Group, http://www.vqeg.org/.
[22] Z. Wang, Rate scalable foveated image and video communications, Ph.D. Thesis, Department of ECE, The University of Texas, Austin, December 2001.
[23] Z. Wang, Demo images and free software for 'a universal image quality index', http://anchovy.ece.utexas.edu/~zwang/research/quality_index/demo.html.
[24] Z. Wang, The SSIM index for image quality assessment, http://www.cns.nyu.edu/~zwang/files/research/ssim/.
[25] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (March 2002) 81–84.
[26] Z. Wang, A.C. Bovik, L. Lu, Why is image quality assessment so difficult? in: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, May 2002.
[27] Z. Wang, L. Lu, A.C. Bovik, Video quality assessment using structural distortion measurement, in: Proceedings of the IEEE International Conference on Image Processing, Rochester 3 (September 2002) 55–68.
[28] Z. Wang, H.R. Sheikh, A.C. Bovik, Objective video quality assessment, in: B. Furht, O. Marques (Eds.), The Handbook of Video Databases: Design and Applications, CRC Press, Boca Raton, FL, 2003.
[29] A.B. Watson, The cortex transform: rapid computation of simulated neural images, Comput. Vision Graphics Image Process. 39 (1987) 311–327.
[30] A.B. Watson, DCT quantization matrices visually optimized for individual images, in: Proceedings of the SPIE, Vol. 1913, 1993, pp. 202–216.
[31] A.B. Watson, Visual detection of spatial contrast patterns: evaluation of five simple models, Opt. Express 6 (January 2000) 12–33.
[32] A.B. Watson, J. Hu, J.F. McGowan III, DVQ: a digital video quality metric based on human vision, J. Electron. Imaging 10 (1) (2001) 20–29.
[33] A.B. Watson, L. Kreslake, Measurement of visual impairment scales for digital video, in: Human Vision, Visual Processing, and Digital Display, Proceedings of the SPIE, Vol. 4299, 2001.
[34] A.B. Watson, G.Y. Yang, J.A. Solomon, J. Villasenor, Visibility of wavelet quantization noise, IEEE Trans. Image Process. 6 (August 1997) 1164–1175.
[35] S. Winkler, A perceptual distortion metric for digital color video, Proceedings of the SPIE, Vol. 3644, 1999, pp. 175–184.
[36] J. Xing, An image processing model of contrast perception and discrimination of the human visual system, in: SID Conference, Boston, May 2002.