A Scalable Wavelet-Based Video Distortion Metric and Applications

Mark Masry Member, IEEE, Sheila S. Hemami Senior Member, IEEE, Yegnaswamy Sermadevi Member, IEEE

Abstract—Video distortion metrics based on models of the human visual system have traditionally used comparisons between the distorted signal and a reference signal to calculate distortions objectively. In video coding applications, this is not prohibitive. In quality monitoring applications, however, access to the reference signal is often limited. This paper presents a computationally efficient video distortion metric that can operate in full- or reduced-reference mode as required. The metric is based on a model of the human visual system implemented using the wavelet transform and separable filters. The visual model is parameterized using a set of video frames and the associated quality scores. The visual model’s hierarchical structure, as well as the limited impact of fine scale distortions on quality judgments of severely impaired video, are exploited to build a framework for scaling the bitrate required to represent the reference signal. Two applications of the metric are also presented. In the first, the metric is used as the distortion measure in a rate-distortion optimized rate control algorithm for MPEG-2 video compression. The resulting compressed video sequences demonstrate significant improvements in visual quality over compressed sequences with allocations determined by the TM5 rate control algorithm operating with MPEG-2 at the same rate. In the second, the metric is used to estimate time series of objective quality scores for distorted video sequences using reference bitrates as low as 10 kbps. The resulting quality scores more accurately model subjective quality recordings than do those estimated using the Mean Squared Error as a distortion metric while requiring a fraction of the bitrate used to represent the reference signal. The reduced-reference metric’s performance is comparable to that of the full-reference metrics tested in the first Video Quality Experts Group evaluation.

I. INTRODUCTION

An objective video distortion metric is an algorithm that measures the distortion in a video sequence without human intervention. It can be used as part of a quality of service monitoring application that can identify changes in video quality over time, or as part of a rate-distortion framework that seeks to optimize the quality of compressed video by minimizing the perceived distortion. In either case, it is essential that the objective distortion scores have a high correspondence with subjective distortion scores, and that they be generated with minimal computational overhead.

An objective video distortion metric can be characterized by the way that it uses the reference signal — the source material from which the distorted sequence originates. Full-reference distortion metrics measure the distance between the reference and distorted sequences. Reduced-reference or no-reference distortion metrics use a lower bandwidth representation of the reference sequence to calculate distortion, or do not use it at all, respectively. Distortion metrics also differ in the types of distortion measures that they produce. Many metrics can produce a map of the distortions at each spatial location in a frame, a time series of frame-level distortion scores, or a single-valued sequence-level distortion score, or all three.

The per-frame Mean Squared Error (MSE) is the most commonly used full-reference distortion metric. Though easy to calculate, the MSE has several shortcomings, most notably its low correlation with subjective scores across a broad range of distortion types. In an effort to address this problem, many recently proposed metrics incorporate accurate models of the low level human visual system (HVS) [1], [2], [3], [4]. Like the MSE, most HVS based metrics require access to the reference sequence. Full-reference metrics were investigated in two Video Quality Experts Group (VQEG) performance evaluations [5], in which the sequence-level objective distortion scores of several of the more recent metrics were found to be more highly correlated with subjective scores than the MSE. Most full-reference metrics can also generate spatial error maps for each frame or a time series of frame-level distortion scores, though it is difficult to record subjective measurements for these types of distortions.

Requiring access to the entire reference signal is a disadvantage in quality monitoring applications, in which the quality of a video sequence is often monitored at a remote location where the reference signal is inaccessible. Several no-reference distortion metrics that estimate sequence distortion without access to the reference signal have recently been proposed. These metrics detect classes of coding artifacts — such as blocking and blurring — in the decoded frames [6], [7], [8]. As a result, no-reference algorithms may not be able to measure the perceived distortion induced by unexpected coding artifacts. Reduced-reference metrics such as the one developed in [9], which use a low-bandwidth versions of the reference sequence as a point of comparison with the distorted sequence, retain the generality of full-reference distortion metrics without requiring access to the entire reference sequence. The quality monitoring applications to which reduced- and no-reference metrics are applied generally require a time-series of frame-level distortion scores, rather than spatial distortion maps.

Rather than develop separate metrics for the tasks of rate-distortion optimization and quality monitoring, this paper proposes a scalable metric that operates in either full-reference mode or reduced-reference mode as required by the application. The metric performs a computationally efficient, perceptually motivated multichannel decomposition on both the reference and distorted sequences. The perceptual decomposition’s pa-
rameters are optimized using a range of distorted video frames and the associated subjective quality scores. Sequence representations consist of trees of coefficients rooted at the coarsest spatial scale and proceeding to the finest. The bitrate of the representation may be scaled from full to reduced by reducing the depth and number of such trees. Integer valued coefficients and transforms are used whenever possible to maximize computational efficiency and to allow the precision of the coefficients at each scale to be specified with bitplanes.

In full-reference mode, the metric produces spatial distortion maps suitable for rate-distortion optimized rate control applications. The effectiveness of these maps is verified by experiments with a rate control algorithm that uses them to minimize the perceived distortion of MPEG-2 compressed video. The resulting compressed video sequences exhibit an improvement in perceived quality over compressed video generated using MPEG-2 and the Test Model 5 (TM5) [10] rate control algorithm operating at the same bitrate. In reduced-reference mode, the metric is used to continuously evaluate the quality of a distorted video sequence. At a reference bitrate of 10 kbps, the metric can produce a time series of objective quality scores in real time on a Pentium 2.5Ghz workstation. These outputs are compared to subjective quality scores for distorted video sequences coded at 10, 15 and 30 fps at rates ranging from 40 kbps to 800 kbps. Over half of the estimated quality data points fall within the 95% confidence interval of the subjective quality recordings across the range of tested content.

The paper is structured as follows: Section II describes recent work on quality metrics. Section III describes the video distortion metric and the underlying HVS model. Section IV evaluates the performance of an MPEG-2 rate-control algorithm that minimizes the perceived distortion over spatial distortion maps calculated using the metric in full-reference mode. Section V compares the results of the metric’s output in a reduced-reference framework to subjective quality recordings. Section VI concludes the paper.

II. PREVIOUS WORK

Of the various types of distortion metrics that have been proposed, full-reference metrics based on human visual models have been the subject of the most research. The HVS models are based on multichannel decompositions in which an input signal is progressively decomposed into multiple bandlimited frequency channels. The outline of a typical perceptual decomposition is given in Figure 1. The decomposition generally begins with conversion of the signal to a mono- or trichromatic color space, followed by spatio-temporal filtering using a non-separable multistage transform (e.g. [11]). The now band-limited input signal is converted to units of contrast. The constructive and destructive interactions of the resulting band-limited contrast components are then modeled in a contrast gain-control stage [12], [13], [14].

Several recent metrics have calculated video distortion using sophisticated multichannel decompositions that model the low-level HVS [1], [2], [3], [15], [16]. The metrics follow a common structure: Reference and distorted sequences each undergo perceptual decompositions and the distances between corresponding elements in the two decompositions are measured and collapsed into a distortion score using successive power summations [17], [18]. The first summation generally collapses the distances at the same spatial location across all frequency channels to generate a spatial map of distortions for each frame. A second summation collapses the spatial distortion map into a single distortion score for each frame. These frame-level distortion scores can then be collapsed into a single sequence-level distortion score. Though the summations need not occur in this order, all of the above-mentioned HVS metrics sum distances across channels in this way and can therefore calculate spatial error maps, frame-level distortion scores and sequence-level distortion scores as needed.

While their ability to generate different types of distortion scores makes them very flexible, metrics based on complete HVS models are generally computationally expensive; they process a large number of channels (as many as 96) and use spatio-temporal decompositions based on non-separable filters (e.g. [19], [11]). Several works have reduced the computational complexity by reducing the number of channels [20], [21], or by constructing metrics using rapid spatio-temporal decompositions such as a blocked DCT [4] applied at the frame-level, separable bandpass filters [9], or the separable wavelet transform [22].

The performance of many full-reference metrics – those metrics requiring access to the entire reference signal – including several of those mentioned above, was investigated in the two recent VQEG evaluations [5]. A test set of distorted video sequences was constructed by coding a database of video sequences with several different coders at rates ranging from 100 kbps to 2 Mbps. The performance of each metric was evaluated by measuring the Pearson and rank-order correlation of its single valued, sequence-level objective distortion scores with subjective scores. Several metrics produced distortion scores that were found to be more highly correlated with subjective scores than those of the MSE.

In quality monitoring applications, access to the reference signal is often limited or nonexistent. Reduced-reference metrics, which require limited access to the reference signal, and no-reference metrics, which do not require the reference signal at all, are favored over full-reference metrics. Many no-reference metrics are generally tuned for the detection of a particular class of artifact such as blocking [6], [7], blurring [8] or sharpness [23]. In [7], the spectral energy of spatial and vertical frequency harmonics whose fundamentals have periods of 8 pixels is used to estimate the presence of edges that occur on 8 × 8 block boundaries. In [8], the first and second derivatives of the intensity values along horizontal scanlines are used to estimate the occurrence of blurring artifacts that result from

![Fig. 1. Block diagram of a typical multichannel perceptual decomposition](image-url)
The metric follows the structure of a multichannel perceptual decomposition. The luminance channel of the reference sequence is extracted and then filtered temporally and spatially, after which subsets of coefficients are chosen for transmission to the client. At the client, the distorted sequence’s luminance channel is also filtered and the subset of coefficients corresponding to those in the reference subset are extracted. Each subset is then converted to units of contrast and passed through a nonlinearity that models spatial masking. The distances between the elements in the reference and distorted subsets are then collapsed into spatial distortion maps or frame-level distortion scores. A block diagram of the metric is given in Figure 2. The individual components are described below.

A. Color Space Conversion

The effects of various color spaces on the performance of a distortion metric were investigated in [20]; the perceptually linear LAB color space provided the best performance. Since the chromatic channels contribute little to spatial resolution processing them requires significant computational overhead, many metrics, including [9], [20], [22] consider only the luminance channel, and discard the chromatic channels. In [20], ignoring the chromatic channels in the YUV color space only reduced the correlation of the objective and subjective scores from 0.78 to 0.76.

Only the luminance channel, termed $Y$, is used in this work. It is calculated using the nonlinear $R’ B’ G’$ pixel intensity values.

$$Y = \frac{54R’ + 179G’ + 22B’}{255}$$ (1)

to yield a luminance component with a dynamic range of 8 bits. Conversion of the $R’ G’ B’$ values to units of monitor intensity is not performed since these coordinates are not perceptually linear, while the $Y$ channel closely approximates the $L$ channel of the LAB space for all but very low intensity values. Note, however, that processing only the luminance channel assumes that errors in the luminance and chrominance channels are highly correlated. This assumption is reasonable for distortions induced by video compression, and held for the compressed sequences evaluated in this work.

B. Temporal Filtering

The visual system is thought to separate visual information into two temporal channels [24], [25], [26]: a lowpass channel with an approximate cutoff around 10 Hz and a bandpass channel with a peak frequency of 15 Hz and an approximate bandwidth of 10 Hz. Since the bulk of the temporal frequency content in video signals is concentrated in frequencies below 10 Hz, several models have used a single lowpass temporal filter to isolate sustained visual information across several frames either before [4], [20], [22], or after [9] spatial filtering. Prior to spatial filtering, the luminance channels are therefore filtered with a lowpass finite impulse response (FIR) filter with an effective cutoff at approximately 10 Hz. Infinite impulse response (IIR) implementations are more commonly used, but the FIR implementation allows small subsets of frames to be processed independently, while requiring access to only a limited number of prior frames.
C. Spatial Filtering

Simulation of the spatial mechanisms of the HVS has been a topic of much recent research. Among the properties which are desirable in a spatial frequency decomposition designed to model the HVS are: invertibility, logarithmically spaced frequency channels, four orientation channels at $0^\circ$, $45^\circ$, $90^\circ$ and $135^\circ$, spatial locality, shift invariance and minimal aliasing between frequency bands. Several different transforms satisfying many of these properties have been used to implement visual models. The spatial decompositions performed in [1], [2], [3], [16] are implemented using transforms based on 2D non-separable filters, including [27], [11]; the computationally demands are therefore high. While a DCT is used in [4] for reasons of computational efficiency, the block based formulation makes it difficult to apply recent masking models as given in [12], [13], [14].

Because it satisfies several (though clearly not all) of these properties, the wavelet transform has been widely used in image processing applications (e.g. [28]). It has not been favored in image or video quality evaluation systems, however, largely because it combines spatial frequencies at $45^\circ$ and $135^\circ$ into a single subband and because it is critically sampled, inducing aliasing in the subbands. The wavelet transform does, however, have the major advantage of being based on separable filters, which allows computationally efficient implementations. Because of this, several recent image quality measures [29], [30], [31] have judged the above shortcomings to be acceptable and used the wavelet transform to perform spatial frequency decompositions.

The spatial frequency decomposition is performed with a four level separable wavelet transform. Critically sampled wavelet transforms downsample the coefficients in the HH, LH and HL bands. If a spatial distortion map with a high degree of local accuracy is desirable (e.g. in the rate control application described in Section IV), the computational advantages of critical sampling are offset by the corresponding increase in coarseness. In these cases, the HH, LH and HL bands are not downsampled at the cost of making the transform overcomplete by a factor of 4. At the frame-level, distortion scores calculated when the DWT is critically sampled were not significantly different from those calculated when the DWT is not. This result is discussed in further detail in Section III-D.

Though the Haar filter is not used in many applications because of its large passband, it is used in this case because its short filter length reduces computation and minimizes edge effects at image borders. Furthermore, an integer implementation of this filter can be trivially obtained by setting the filter taps to have magnitude 1. Integer coefficients speed computation of the exponential functions in the masking non-linearity by allowing look-up tables to be trivially implemented and simplify the construction of a reduced-reference model by eliminating the need for a floating point quantizer.

D. Selection of Coefficient Subsets for Reduced Reference Operation

The bandwidth of the reference signal can be reduced by noting that the distortions produced by typical video compression algorithms are correlated across scale, space and time; this is particularly true for distortions above the threshold of detection. The reference bandwidth is reduced by not-considering the correlation of the multi-resolution structure of the perceptually motivated decomposition implemented in Section III-C: information at the coarsest scales is represented with fewer coefficients than information at the finest scales. Furthermore, the correlation of distortions across space and time means that not all frames, or spatial locations within a given frame, need to be considered.

Using only the coarsest scales to calculate objective distortion scores automatically provides a reduction in the reference signal, since the multiresolution transform uses far fewer coefficients to represent information at the coarsest scales than information at the finest. In many cases, evaluating the distortion in only a subset of frequency bands can yield an improvement in the correlation of the objective distortion scores with subjective scores. The distortion metric described in [9] achieved the best
correlation with subjective scores when its spatial filters were tuned to attenuate frequencies higher than 6.3 cycles per degree (cpd), assuming an image size of 352 x 240 and a viewing distance of 6 picture heights. The multiscale distortion metric in [21] performed best when distortions at the highest spatial frequency scale were not considered (i.e. frequencies greater than 6.3 cpd were ignored, assuming equivalent parameters); similar performance was obtained when the two highest spatial frequency scales were disregarded, effectively attenuating spatial frequencies above 3.2 cpd.

After color space conversion, temporal and spatial filtering, the number of coefficients required to represent each frame of the reference signal is reduced as follows:

1) The subbands of the reference decomposition are downsampled horizontally and vertically by a factor of $M$.
2) A subset $\{x_i\}$ of $1/C$ of the coefficient locations at the coarsest scale are chosen to be the roots of trees of coefficients across scale. These can be chosen randomly or according to a regular pattern.
3) The HH, LH and HL coefficients at each spatial location $x_i$ at the coarsest scale are inserted into the tree rooted at $x_i$.
4) At each of the $S - 1$ higher scales the coefficients whose spatial locations are subtended by $x_i$ are inserted into the tree. The number of coefficients inserted into the tree at each scale is 4 times larger than at the next coarsest scale, as a result of the downsampling operations performed by the wavelet transform.

Note that given a four level decomposition $C = 1, S = 4$ and $M = 1$ selects the entire reference decomposition.

### E. Contrast Computation

There are several definitions by which contrast is measured analytically. The simplest, the Weber fraction, measures the contrast of a simple target on a background of constant (and analytically. The simplest, the Weber fraction, measures the contrast of a simple target on a background of constant (and

$\Delta f = \frac{\Delta L}{L}$ where $\Delta L$ is the change in luminance and $L$ is the background luminance. The Michelson contrast, given by $C_{\text{Michelson}} = \frac{L_{\text{max}} - L_{\text{min}}}{L_{\text{max}} + L_{\text{min}}}$ where $L_{\text{max}}$ and $L_{\text{min}}$ are the maximum and minimum luminances, respectively, is generally used for periodic waveforms. Neither of these takes spatial frequency into account. To correct for this shortcoming, Peli [32] defined a Local Bandlimited Contrast measure (LBC) given by

$$f_k(x, y) = \frac{a_k(x, y)}{l_{k-1}(x, y)} \quad (2)$$

where $f_k(x, y)$, $a_k(x, y)$ and $l_k(x, y)$ are the local contrast, bandpass response and lowpass response, respectively, at a scale $k$ and spatial location $(x, y)$. Division by $l_{k-1}(x, y)$ rather than $l_k(x, y)$ was suggested in [33]. In psychovisual tests using sine-wave gratings and Gabor patches, this corrected measure was better correlated with the observed contrast than several other contrast measures [34], [35]. This formulation has the further advantage that the contrast response is normalized across scale, which eliminates the effects of non-unity filter gains.

Several distortion metrics [36], [31] have used the bandpass response terms $a_k(x, y)$ directly as an approximation to the contrast at a given spatial frequency under the assumption that mean luminance is constant over the image. This formulation does not allow the metric to compensate for variations in overall luminance across a range of frames, as can occur in a video sequence of reasonable length, and eliminates the normalization property of Equation 2.

The local contrast at a frame $n$, spatial location $(x, y)$, scale $k$ and orientation $\theta$ is therefore defined by

$$f_{k,\theta}(x, y, n) = \frac{a_{k,\theta}(x, y, n)}{l_{k-2}} \quad (3)$$

where $\bar{l}_k$ is the mean of the anisotropic lowpass response at scale $k$ computed from the lowpass subbands constructed at each stage of the wavelet transform (this can also be computed directly from the product of the filter mean and the image mean). Following computation of local contrast, the contrast responses at each scale are weighted by the contrast sensitivity function (CSF) for the peak frequency of that scale. Contrast sensitivity is a function of both spatial frequency and mean luminance. Contrast sensitivity is calculated using the CSF model in [37] since it is a function of both spatial frequency (where viewing distance is here assumed to be 6 picture heights) and mean luminance in candelas per square meter ($\frac{cd}{m^2}$). The mean luminance of each frame is calculated using the mean of the intensity plane, assuming a monitor gamma of 2.6 and white point of 98 $\frac{cd}{m^2}$.

### F. Masking

The coefficients generated by the previous stage separate the input signal into multiple scales and orientations. To this point, neither inter- nor intra-channel interactions have been considered. Furthermore, coefficient magnitudes by themselves do not accurately simulate neural responses. Most importantly, the visual phenomenon in which the responses in a particular channel interfere with those in another, known as masking, cannot be simulated without considering inter-channel effects. Several works [13], [14], [36] have proposed models of channel interaction based on the observed behavior of neurons in the visual cortex [12].

The metric implements a masking model based on the generalization gain control formulation in [14]. The channel response at a frame $n$, spatial location $(x, y)$, scale $k$ and orientation $\theta$ is given by

$$r_{k,\theta}(x, y, n) = \frac{A e^{p} f_{k,\theta}^p(x, y, n)}{b + c_k \sum_{\theta} f_{k,\theta}^p(x, y, n)} \quad (4)$$

where $A$ is a scaling constant, $c_k$ is the contrast sensitivity at scale $k$ and $f_{k,\theta}(x, y, n)$ is the contrast as specified in Equation 3. The inhibitory power $q$ is fixed at 2; the excitatory power $p$ the divisory constant $b$ and the scaling constant $A$ are chosen to fit model data, as described in Section III-I. Inhibitory summation only included subbands within the same level, but summed across all orientations.
G. Summation

As in many distortion metrics, distortion is given by the distance between the decompositions for reference and distorted sequences \( r_{k,\rho}(x, y, n) \) and \( r_{k,\rho}^d(x, y, n) \) respectively according to the rules of probability summation [17], [18]. A distortion map is given by

\[
d(x, y, n) = \left( \bigoplus_{k,\theta} |r_{k,\rho}(x, y, n) - r_{k,\rho}^d(x, y, n)|^\beta_f \right)^{1/\beta_f}
\]

where the \( \bigoplus \) is a summation operator that first maps the values at each scale to the pixel level using replication and \( \beta_f \) is the summation exponent for frequency summation. As the value of \( \beta_f \) increases, this summation emphasizes the largest summation terms, until \( \beta_f = \infty \) returns the supremum of the summation terms. A single distortion score for each frame is given by

\[
d(n) = \left( \sum_{x,y} d(x, y, n)^\beta_s \right)^{1/\beta_s}
\]

where \( \beta_s \) is the summation exponent for spatial summation.

Frames may be dropped or replicated during playback of the video sequence. The distortion of each frame is therefore given by the minimum distortion between the reference subset at time \( n \) and the features extracted from any distorted frames falling in the search window \( n - N - 2 \) to \( n + N + 2 \). The value of \( N \) is initially set to 0. If the distorted frame \( n - N \) does not produce the minimum distortion, a frame drop or replication is assumed to have occurred, and the value of \( N \) is adjusted accordingly to track the relative shift.

H. Mapping Distortion to Quality

Because quality, rather than distortion, is generally used as a performance measure, the distortions \( d(n) \) are passed through a two parameter, non-symmetric mapping [38] onto a quality scale from 0 to 100 to yield an objective quality score \( q(n) \) for frame \( n \). Though \( q(n) \) often varies significantly across adjacent frames, recordings of subjective quality made continuously as distorted video is shown demonstrate far less variability [39]. This is perhaps due to viewer inability to make rapid subjective judgments, or to the mechanical properties of the slider devices used to record the subjective scores, or to a combination of the two factors.

To more accurately model these subjective quality recordings, the quality score as frame \( n \) is shown, \( q(n) \), is given by a causal, lowpass function of the frame-level objective quality scores \( q(n) \) [21]:

\[
q(n) = \begin{cases} 
q(n-1) + \alpha_- \Delta(n) & \text{if } \Delta(n) \leq 0 \\
q(n-1) + \alpha_+ \Delta(n) & \text{if } \Delta(n) > 0 
\end{cases}
\]

where \( \Delta(n) = q(n) - q(n-1) \). The values of \( \alpha_- \) and \( \alpha_+ \) weight the dependence of quality scores on positive and negative trends in \( q(n) \) and are determined by optimizing the fit of \( q(n) \) to a set of subjective quality recordings. Different weights are used to measure the effects of sustained increases and decreases in frame-level quality over time because it has not been established that viewers react with equal magnitude to either.

I. Parameterization

Parameterization occurs in two separate stages. Since the distortions under consideration here are generally supra-threshold, and are not masked by simple maskers such as Gabor patches, or sine gratings, the parameters of Equation 4 and the summation exponents \( \beta_f \) and \( \beta_s \) are optimized using a set of distorted video frames. The two independent parameters of the non-symmetric function mapping distortion to quality, and the weights \( \alpha_- \) and \( \alpha_+ \), are then determined using a set of subjective quality recordings.

The parameters of Equation 4 were first initialized using the data in [13]. The parameter \( p \) and the summation exponents \( \beta_s \) and \( \beta_f \) were determined using training sets of video frames and the associated subjective quality scores. The training set consisted of eight groups of 10 frames, each taken from a different distorted video sequence. The distorted video sequences are described in Section V-A. Each distorted video sequence was created by compressing one of four 4:2:0 YUV non interlaced 352 × 240 format reference sequences with one of three video coders. Bitrates varied from 200 kbps to 800 kbps. Because the characteristics of the optimization landscape are uncertain, the optimization was performed using an exhaustive search over integer values of \( \beta_s \) and \( \beta_f \), and values of \( p \) quantized with a step size of 0.02. At each iteration of the search, the mean distortion score across each training set was calculated using the trial values of \( p, \beta_s \) and \( \beta_f \). The optimal values were those that maximized the correlation between the objective distortion scores and the inverse of the subjective quality scores.

The parameter \( p \) determines the degree to which a target is masked by a background that is above its own threshold contrast; lower values of \( p \) correspond to larger masking effects. For the threshold data in [13], \( p = 2.42 \) was optimal, which is close to the results noted in [14]. For the training set of video frames, \( p = 2.1 \) produced the highest correlation with subjective distortion scores. This is due to the varying form of the target distortions in the training set, as well as the fact that the subjective distortion scores were collected in a single-stimulus test, rather than in a double-stimulus comparison as in [13], which increased viewer uncertainty about both target and masker.

The optimal value of \( \beta_s \) was found to be 2. The optimal value of \( \beta_f \) was found to be \( 2 \); the largest distortion measured across frequency channels has a disproportionate effect on the overall distortion. The value of \( \beta_f \) was found to have less effect on the goodness of fit than did the values of the parameters \( p \) and \( \beta_s \), however, so this effect is not pronounced.

The two parameters of the non-symmetric function that maps distortion scores onto a quality scale from 0 to 100, and the smoothing weights \( \alpha_- \) and \( \alpha_+ \), were determined by fitting the distortion metric’s outputs to a set of subjective quality recordings chosen from those described in Section V-A. Optimization was performed simultaneously over all 4 parameters by minimizing the squared error between a subset of subjective quality measurements chosen each recording and the quality calculated from the associated values of \( d(n) \) using a Nelder-Mead simplex [40]. These parameters were determined independently of the model parameters determined above. The resulting values of \( \alpha_- \) and \( \alpha_+ \) are 0.04 and 0.50, respectively, suggest-
ing that viewers of the test sequences in V-A reacted much more strongly to decreases in the frame-level quality than to increases.

IV. APPLICATION I: RATE CONTROL WITH A PERCEPTUAL DISTORTION METRIC

Video compression algorithms reduce the bandwidth of a video signal to a specified rate by reducing redundancies in the video signal, a process which induces distortions in the coded signal. Rate control algorithms assign bits to each frame with the goal of minimizing the visual distortion – as measured by the MSE or another distortion metric such as the one introduced in this work – while meeting target bitrate and buffer constraints. In block-based, predictive video compression algorithms such as MPEG-2, frames are coded as one of three types: I-frames, which are coded independently, and P- and B-frames, which are constructed using 16 × 16 blocks of pixels, termed macroblocks, predicted from other frames in the video sequence. The assignment of bits to each frame is a function of the type of frame being coded and the video content itself, as well as the distortion metric.

A. A Rate Control Framework

The Linear Programming-based (LP) rate control algorithm developed in [41] determines a bit allocation for a given video sequence by minimizing the maximum frame-level distortion score across all frames in the sequence, subject to rate and buffer constraints. This process requires the construction of rate-distortion curves at the frame and macroblock levels using a particular video coder and distortion metric. The visual quality of the coded sequence is thus determined by the chosen metric’s ability to produce objective distortion scores consistent with those that might be produced during subjective evaluations.

The metric described in Section III with \( M = 1 \), \( C = 1 \), \( S = 3 \) is used to generate spatial distortion maps that are then summed into macroblock- and frame-level distortion scores in order to optimize the frame rate allocations for an MPEG-2 video coder. The distortion of macroblock \( b_i \) in frame \( n \) is given by

\[
d(b_i, n) = \left( \sum_{x,y \in b_i} d(x, y, n)^{\beta_s} \right)^{1/\beta_s}
\]  

(8)

The frame-level distortion score is then given by a power summation with power \( \beta_s \) of the resulting set of macroblock-level distortion scores, which is equivalent to determining it directly as in Equation 6.

The LP rate control algorithm begins with an initial allocation, which may be constructed using an existing rate-control algorithm. The algorithm then constructs piecewise linear approximations for each frame-level rate-distortion curve, assuming that the allocation for the previous frame is specified by the initial allocation. These approximations form the basis for a linear programming optimization that determines the optimal allocation for each frame while satisfying rate and buffer constraints. This new allocation is then used as the initial allocation for another iteration.

B. Results

The LP rate control algorithm with distortion maps generated by the proposed metric was used to determine MPEG-2 allocations for several standard test sequences, including football, mobile, tennis, garden and foreman at bitrates ranging from 500 to 800 kbps and a frame rate of 30 fps. The results were compared to MPEG-2 coded sequences whose bit allocations were generated by the MPEG-2 Test Model 5 [10] rate control algorithm.

Test Model 5 (TM5) is a single pass rate control algorithm that produces a bit allocation for each macroblock in a video sequence, given a target bitrate and a fixed GOP size. It does not incorporate an explicit psychovisual model, nor does it consider an entire video sequence before determining the allocation for each macroblock. The final LP bit allocations were determined by iteratively optimizing the TM5 allocation; no more than 4 iterations were required for convergence.

Examination of the coded sequences produced by the LP algorithm reveals a reduction in the number of visible macroblock edges (blockiness), as well increased sharpness of textured areas, compared with the analogous TM5-coded sequences over the range of tested sequences and rates. The LP-coded sequences also exhibit fewer fluctuations in visual quality over their length. The foreman encodings are most improved, largely because blocking effects in the TM5-coded sequences are very noticeable; mobile encodings are least improved, perhaps because this sequence is very difficult to code. Example frames from the LP and TM5 allocations for the football sequence coded at 800 kbps and the garden sequence coded at 500 kbps are shown in Figure 3.

Plots of the frame-level MSE and objective distortion scores \( d(n) \) given by the metric for the football sequence coded at 800 kbps are given in Figure 4. The frame-level distortion scores for the LP allocation are considerably flatter than those of the TM5 allocation, corresponding to the more consistent visual quality across frames. Furthermore, the values of \( d(n) \) for the final allocation are considerably lower than those of the TM5 allocation, particularly for frames 15-29. This is consistent with a visual inspection of these frames. In spite of the improvement in visual quality, the MSE scores for the final allocation for frames 15-29 are very similar to those of the initial allocation.

One shortcoming of the proposed framework is that, in some cases, a particular macroblock’s distortion score does not sufficiently affect the frame-level distortion score, resulting in over-quantization of that block. This may be addressed by lower-bounding the allowable block distortion scores when performing the allocation. The computational performance of the algorithm is another issue. Construction of the rate-distortion curves requires a great deal of computation since the MPEG coder and the distortion metric must be executed multiple times for each frame. This overhead may greatly be reduced by using approximations to the rate-distortion curves.

V. APPLICATION II: CONTINUOUS QUALITY EVALUATION

The suitability of the metric for continuous quality evaluation was evaluated by varying the parameters \( M, S, C \) and \( N \) to construct reduced-reference metrics that require limited
A SCALABLE WAVELET-BASED VIDEO DISTORTION METRIC AND APPLICATIONS

bitrates to transmit a representation of the reference sequence from server to client; the objective quality scores estimated using these metrics were compared to subjective quality scores for a wide range of distorted video sequences. The information transmitted from the server to the client consists of the set of locations \( \{ x_i \} \), the coefficient trees rooted at \( \{ x_i \} \) and the mean intensity values \( \{ l_i \} \).

A. Subjective Quality Recordings

A Single Stimulus Continuous Quality Evaluation (SSCQE) [38] subjective test was performed to generate continuous recordings of perceived quality with the goal of measuring continuous responses to artifacts that are found in compressed video at low rates [39]. A single stimulus format was chosen because the sequences were relatively long and since, in general, viewers of compressed video sequences do not have access to the original. In order to evaluate coder performances over a broad range of video content, eight reference sequences were chosen from a variety of different motion and content types. Each sequence was approximately 30 seconds long in 4:2:0 YUV non-interlaced 352 × 240 format and was recorded at 30 fps.

The sequence identifiers and the five coding conditions with which each was compressed are given in Table I. High, medium and low coding bitrates were chosen according to sequence complexity. Note that coding conditions med / 30, med / 15, med / 10 use the medium bitrate, but vary the framerate in order to examine the effects of frame rate on visual quality. Three coders H.263+ [42], Wavelet [43] and VQ [44] were used for each coding condition, which produced artifacts of quite different types, ranging from predominately blocky to predominately blurry. Compressing the eight reference sequences with each of the coder and condition combinations resulted in 120 distorted sequences. Nineteen viewers viewed and rated each sequence. Their responses were recorded with a continuous slider device. The resulting recordings were processed as per Annex 3 of ITU-BT.500-10 to generate mean recordings for each compressed sequence [38]. Ninety-five percent confidence intervals for the subjective quality recordings were normally distributed with a mean of 8.8 and a standard deviation of 4.

![Fig. 3. Subregions of frame 16 of the football sequence coded at 800 kbps using MPEG-2 with (a) the TM5 rate control algorithm (b) the LP rate control algorithm. (c) and (d) are subregions for analogous output frames for frame 25 of the garden sequence coded at 500 kbps.](image-url)
B. Time Series Comparisons

The metric’s performance was evaluated by measuring the distance between the time series of objective quality scores \( \hat{q}(n) \) given by Equation 7 and the subjective quality recordings described in Section V-A. This work makes use of a time series similarity measure based on Piecewise linear representations (PLRs), which are among the most efficient and frequently used representations for time series data. The bottom-up algorithm as described in [45] was shown to be fast and among the most accurate of the piecewise linear approximation algorithms.

The algorithm constructs the initial approximation to the original length \( N \) time series using \( N/2 \) segments. The algorithm then merges these recursively until the error resulting from merging two segments exceeds a user defined tolerance. The tolerance of a potential merge is defined herein as the vertical distance of any endpoint in the two segments to be merged from the vertical mean of the resulting merged segment. It is assumed that both time series have magnitudes normalized from 0 to 100, as is true in the case of comparisons between the data from [39] to the output of the reduced-reference quality monitoring algorithm. The PLR of a time series \( A \) thus consists of a set of \( M \) contiguous line segments that are defined by the \( M+1 \) endpoints

\[
\{(x_i^A, y_i^A), 0 \leq i \leq M\}
\]  

(9)

A measure of the relative vertical offset between each of the two series, \( Offset(A, B) \), is calculated by comparing the vertical offsets of the corresponding segments in each PLR.

\[
Offset(A, B) = \sum_{i=0}^{M-1} w_i \left| \frac{y_{i+1}^A + y_i^A}{2} - \frac{y_{i+1}^B + y_i^B}{2} \right|
\]  

(10)

The weighting term \( w_i \) is used to weight segments according to their relative length compared to the overall length of the time series to be compared

\[
w_i = \frac{x_{i+1}^A - x_i^A}{x_M^A - x_0^A}
\]  

(11)

Three implementations of the metric were used to test the effects of varying the bandwidth of the reduced-reference representation. The first, termed the “Full-bandwidth metric” was given by setting \( M = 1, S = 2 \) and \( C = 1 \) and \( N = 1 \). All coefficients were represented with full precision. The second metric used a reference bandwidth of approximately 40 kbps, and was given by setting \( M = 2, S = 2 \) and \( C = 6 \) and \( N = 5 \). The third metric used a reference bandwidth of approximately 10 kbps, and was given by setting \( M = 2, S = 1 \) and \( C = 6 \) and \( N = 5 \). In both the second and third cases, coefficients at the coarsest scale were represented with 10 bits.
A SCALABLE WAVELET-BASED VIDEO DISTORTION METRIC AND APPLICATIONS

10

of precision while coefficients at the next coarsest scale were represented with 8 bits of precision.

Each metric’s outputs were compared to subjective quality recordings using Equation 10 and quality scores that fall within the 95% confidence interval of the subjective quality recording. The results are analyzed in Table II over sets of distorted sequences with common reference sequence, coding condition, coder or frame rate. The resulting quality scores generally fell within ±8 quality units of subjective scores on a quality scale of 0-100 for all but the lowest bandwidth reduced-reference metric. In general, the full-reference and 40kbps metrics placed more than 50% of all frames within the 95% confidence interval, whereas the 10kbps metric placed slightly less than 50% of all frames within the confidence interval. These figures also include transient periods where quality changes rapidly from high to low; in these periods the quality scores tended to fall outside the confidence interval more often than in periods where the subjective quality varied little over time.

For the purpose of comparison, Table II also includes an identical analysis for a quality metric that uses the logarithm of the frame-level mean squared error to estimate frame-level distortion scores, which are then temporally filtered using an 8 second sliding window and mapped to a quality scale using a non-symmetric function. The MSE-based metric exhibited severe failures on sequences coded with the commercial coder. These sequences were not included in the analysis to avoid biasing the results. With very few exceptions, notably the video-conference sequences, the proposed metrics outperformed the MSE-based metric across the range of content.

As expected, the full-bandwidth metric exhibited the best performance across the range of distorted sequences, followed by the 40 kbps metric and then the 10 kbps metric. There was little difference between the outputs of the full-bandwidth and 40 kbps metrics, overall, suggesting that reductions in the reference bandwidth do not cause substantial degradations in performance. The performance of the 10 kbps metric was not severely degraded, with the exception of distorted versions of the video-conference and sports reference sequences. Differences in performance between the two reduced-reference metrics are entirely explained by distortions measured at the second-coarsest scale.

The mean offset between the full-bandwidth metric’s output and the subjective quality recordings for all possible intersections of two subsets provides a way to characterize outlier sequences. The 5 worst cases are given in Table III. Quality estimates for the car chase sequence coded at the highest bitrate exhibited the highest difference from subjective scores. This sequence exhibited a range of extremely jerky camera motions which may have masked the coding artifacts that were still present at higher rates. Four of the worst intersections include the med / 10 condition, suggesting that the metric had some difficulty accounting for frame rate preferences in certain sequences. These issues can be addressed by improved temporal processing components; separate instances of the metric can also be parameterized to handle sequences with a great deal of temporal masking, or sequences where frame rate effects play a significant role. Table III also shows the best 5 such cases. The subjective quality changes little over the length of these coded sequences; discrepancies between the metric’s output and the subjective quality are largest in regions where the subjective quality changes rapidly, so the appearance of relatively uniform quality sequences in the best 5 sets is not surprising.

Scatter plots of the subjective vs. objective frame qualities scores for the 40 kbps metric and the MSE-based metric are given in Figure 5. The $R^2$ values for a linear regression fit of objective to subjective quality scores are 0.81 and 0.48, respectively. Four example time series of objective quality scores calculated by the 40 kbps metric are given in Figure 6. Time series 95% confidence intervals are given by the shaded region around the subjective quality recordings. Notably, many of the quality estimates that fell outside the confidence interval were in regions were the subjective quality scores were changing very rapidly over short periods of time; though this may not be a concern in many quality of service monitoring applications, a more sophisticated temporal smoothing model can address this issue.

C. Comparison with Other Metrics

Though few performance results have yet been published for the class of reduced or no-reference video quality metrics using a standardized data set, a comprehensive evaluation of several full-reference metrics using a standardized data set was made available by the Video Quality Experts Group [5], the results of which provide a basis for comparison this work. Objective distortion scores for VQEG 60Hz source sequences 13 through 19 and three of the low quality VQEG coding conditions – HRC12, HRC14 and HRC16 – were estimated using the 40 and 10 kbps reduced-reference metrics. Source sequence 19 coded with HRC16 contained time-varying spatial shifts, so this sequence was discarded. The portion of each frame outside the central 352 × 240 area was cropped. Since the VQEG results are given by single-valued Distortion Mean Opinion Scores (DMOS), rather than a time series of values, a distortion score for the entire sequence was calculated for each of the remaining 17 coded sequences using a power sum of the form $D = (\sum_{n=1}^{N} d(n)^{\beta})^{1/\beta}$. This summation was adopted for simplicity though, as explained in [47], it is not an ideal method of accumulating frame-level distortion scores; performance may be improved with more sophisticated technique.

The resulting sequence-level distortion scores and the PSNR were each mapped onto a distortion scale using 3 parameter logistic functions prior to being compared with the subjective Distortion Mean Opinion Scores (DMOS) for the distorted sequences, taken from the VQEG full-reference evaluation. The parameters of the logistic function were determined using coded sequences based on source sequences 17, 18 and 19 as a training set. Coded sequences based on source sequences 13 through 16 were used as a test set. A temporal summation exponent of $\beta = 0.210$ was chosen by exhaustively searching the range $0.01 \leq \beta \leq 10$ in increments of 0.01 for the value that minimized the correlation between subjective and objective scores.

1 Coding conditions med / 30, med / 15, med / 10 use the same bitrate, but different frame rates. Frame rate preferences across 10, 15 and 30 fps using these coding conditions were investigated in [46]
This evaluation revealed that the proposed 40 kbps and 10 kbps reduced-reference metrics offer performance comparable to the full-reference metrics evaluated in the VQEG evaluation. The correlations of the 40 kbps and 10 kbps metrics with subjective scores for the sequences in the test set were 0.896 and 0.887, respectively. The correlation of PSNR over the same range of sequences was 0.785. Of the metrics tested in the VQEG evaluation, the highest correlation over a comparable set of sequences and 8 low quality coding conditions, including the three considered here, was 0.891 by Proponent 5 [20]. Both Proponent 5 and the PSNR require full access to the reference sequence.

The proposed metric was implemented using a combination of Matlab and C++. Since the metric is based on a recursive transform and uses separable filters, it is computationally efficient, particularly in reduced-reference mode. The 10 kbps metric operates in real-time on a Pentium 4 2.53GHz machine. The 40 kbps metric operates at approximately half real-time speed. The full-reference metrics process several frames per second. More efficient software implementations can easily be constructed.

VI. CONCLUSIONS

The distortion metric presented in this paper is based on an efficient perceptually motivated multichannel decomposition implemented using the wavelet transform. The parameters of the masking model used in this decomposition are determined using a training set of natural video frames and the associated subjective quality scores. The structure of the metric is exploited by a coefficient selection mechanism that allows the bitrate of the reference sequence decomposition to be scaled. The metric can produce both spatial distortion maps and a time series of distortion scores.

Video sequences compressed by a rate-control algorithm that uses the metric as its spatial distortion measure show improvements in perceived quality over sequences compressed using the TM5 rate control algorithm and MPEG-2 at the same rates. The improvement in visual quality is most pronounced when blocking effects in the TM5 sequences are especially noticeable, and when the video sequence contains large textured regions. The time series of objective quality scores generated by the metric operating with several different reference bitrates as low as 10 kbps are similar to subjective quality scores, with approximately 50% of all objective frame-level quality scores falling within the 95% confidence interval for the subjective scores. Though performance degrades somewhat as the reference bitrate decreases, the results for all reference bitrates are an improvement over those generated with a MSE-based distortion metric. Further reductions in the reference bitrate are possible through careful selection of the metric’s parameters.

REFERENCES

A SCALABLE WAVELET-BASED VIDEO DISTORTION METRIC AND APPLICATIONS

<table>
<thead>
<tr>
<th>Subset of Encodings</th>
<th>full-bandwidth metric</th>
<th>40 kbps metric</th>
<th>10 kbps metric</th>
<th>MSE metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Offset % in Ci</td>
<td>Offset % in Ci</td>
<td>Offset % in Ci</td>
<td>Offset % in Ci</td>
</tr>
<tr>
<td>{News}</td>
<td>7.05</td>
<td>70.81</td>
<td>6.79</td>
<td>67.63</td>
</tr>
<tr>
<td>{Videoconference}</td>
<td>7.73</td>
<td>65.39</td>
<td>8.48</td>
<td>56.81</td>
</tr>
<tr>
<td>{Computer animation}</td>
<td>7.52</td>
<td>50.66</td>
<td>4.60</td>
<td>48.22</td>
</tr>
<tr>
<td>{Crowd scene}</td>
<td>7.52</td>
<td>58.67</td>
<td>6.19</td>
<td>70.54</td>
</tr>
<tr>
<td>{Moving people}</td>
<td>5.06</td>
<td>61.40</td>
<td>5.10</td>
<td>55.72</td>
</tr>
<tr>
<td>{Martial arts}</td>
<td>7.98</td>
<td>44.84</td>
<td>8.54</td>
<td>43.78</td>
</tr>
<tr>
<td>{Sports}</td>
<td>8.27</td>
<td>54.65</td>
<td>9.97</td>
<td>44.42</td>
</tr>
<tr>
<td>{Car chase}</td>
<td>6.68</td>
<td>62.74</td>
<td>6.86</td>
<td>64.19</td>
</tr>
</tbody>
</table>

Table II
RESULTS FOR SUBSETS OF DISTORTED SEQUENCES THAT ARE HOMOGENEOUS IN ONE CHARACTERISTIC IN TERMS OF MEAN Offset VALUES, AND PERCENTAGE OF OBJECTIVE QUALITY TIME SERIES VALUES THAT FELL WITHIN THE 95% CONFIDENCE INTERVAL, AS MEASURED ACROSS ALL DISTORTED SEQUENCES WITHIN THE SUBSET

<table>
<thead>
<tr>
<th>Subset of Encodings</th>
<th>Offset % in Ci</th>
</tr>
</thead>
<tbody>
<tr>
<td>{H.263+}</td>
<td>12.66</td>
</tr>
<tr>
<td>{med / 10 fps}</td>
<td>12.69</td>
</tr>
<tr>
<td>{med / 10 fps}</td>
<td>12.77</td>
</tr>
<tr>
<td>{Martial arts}</td>
<td>13.07</td>
</tr>
<tr>
<td>{Car chase}</td>
<td>16.95</td>
</tr>
<tr>
<td>{Car chase}</td>
<td>16.95</td>
</tr>
</tbody>
</table>

Table III
WORST FIVE (LEFT) AND BEST FIVE (RIGHT) INTERSECTIONS OF SETS OF DISTORTED SEQUENCES BY MEAN Offset OF ALL SEQUENCES IN INTERSECTION

Fig. 5. Scatter plot of objective vs. subjective quality scores for a reduced number of frames from every encoding for (a) the proposed 40 kbps metric (b) a metric based on the MSE. The $R^2$ values for a linear regression fit are 0.81 and 0.48, respectively.
Fig. 6. Metric outputs for a) computer animation coded at 10 fps and 200 kbps with the Wavelet coder b) Martial arts coded at 15 fps and 200 kbps with the Quicktime coder and c) sports coded at 10 fps and 150 kbps with the Wavelet coder. d) crowd scene coded at 30 fps and 600 kbps with the Quicktime coder.

