Title

Quality Evaluation of Motion-Compensated Edge Artifacts in Compressed Video

Journal

IEEE Transactions on Image Processing, 16(4)

ISSN

1057-7149

Authors

Leontaris, Athanasios
Cosman, Pamela C
Reibman, Amy R

Publication Date

2007-04-01

DOI

10.1109/TIP.2007.891778

Peer reviewed

Quality Evaluation of Motion-Compensated Edge Artifacts in Compressed Video

Athanasios Leontaris, *Member, IEEE*, Pamela C. Cosman, *Senior Member, IEEE*, and Amy R. Reibman, *Fellow, IEEE*

***Abstract*—Little attention has been paid to an impairment common in motion-compensated video compression: the addition of high-frequency (HF) energy as motion compensation displaces blocking artifacts off block boundaries. In this paper, we employ an energy-based approach to measure this motion-compensated edge artifact, using both compressed bitstream information and decoded pixels. We evaluate the performance of our proposed metric, along with several blocking and blurring metrics, on compressed video in two ways. First, ordinal scales are evaluated through a series of expectations that a good quality metric should satisfy: the objective evaluation. Then, the best performing metrics are subjectively evaluated. The same subjective data set is finally used to obtain interval scales to gain more insight. Experimental results show that we accurately estimate the percentage of the added HF energy in compressed video.**

***Index Terms*—Added high-frequency energy, blocking artifacts, blur, motion-compensated edge artifact, objective metrics, paired comparison, subjective tests, Thurstone's law, video compression, video quality assessment.**

I. Introduction

STANDARDIZATION bodies such as the Video Quality Experts Group (VQEG) [3] have been coordinating research efforts towards designing an efficient objective video quality metric. The goal of an objective video quality metric is the automatic prediction of perceived quality, useful not only to assess the quality of reconstructed video, but also for fine-tuning and design of video coding systems. Mean-squared error (MSE) and peak signal-to-noise ratio (PSNR) have seen widespread use as video quality metrics due to their implementation simplicity and adequate performance. Unfortunately, they do not take into account the perceptual characteristics of the Human Visual System (HVS). Previous research [4] showed that PSNR cannot perform well in video sequences with significant luminance or spatial masking. Incorporating HVS models into video quality metrics is highly desirable.

Objective quality metrics can be categorized as full-, reduced-, and no-reference (NR) metrics. Full-reference (FR) metrics have access to both the original and the reconstructed video. An example is PSNR. Reduced-reference metrics rely on some features that have been previously extracted from the original video content by some other means. These metrics do not have direct access to the original video. Finally, NR metrics have access only to the reconstructed video sequence and its bitstream. These metrics are universally deployable, since they do not require access to the original sequence.

Video quality has both spatial and temporal dimensions [5]. We treat the spatial component of video quality. Visual quality assessment is a highly nonlinear process that involves the HVS and high-level cognitive functions. Quality depends on the *user*. The user's assessment differs due to variations in sensory capabilities (eyesight), personal expectations, experience, and motivation. Second, quality depends on the *environment*, such as the room lighting, the type of screen (LCD or CRT) used for the evaluation, and the viewing distance. Third, quality is *content* dependent, since the judgment criteria change whether the subject is watching a TV commercial or a sports program.

Furthermore, viewers perceive quality on many axes simultaneously. For example, an image may be blurry, blocky, or have ringing artifacts. Temporally, a video may have jerky motion, or added "mosquito" noise. In this paper, while we are interested in video quality, we assess only the quality of the individual still images (frames) that comprise the video. Video quality is a function of four types of impairments: blocking,[1] blurring, ringing, and motion-compensated edge artifacts.

Blockiness arises from the vertical and horizontal edges along a regular blocking grid that result from the block-based processing in many image and video codecs. Coarse quantization yields more blockiness, while edge-attenuating filters reduce its perceptual effect. We concentrate on blocks of size 8×8 as it is the default DCT block size. *Blurriness* is caused by the removal of high-frequency content from the original image/video signal. Increased blurriness can be caused by coarser quantization, edge-attenuating filters, fractional-pixel motion compensation (MC) or overlapped block motion compensation (OBMC). *Ringing artifacts*, also known as the Gibbs phenomenon, are caused by the absence of high-frequency terms from a Fourier series due to coarse quantization. They are perceived as ripples and overshoots near high-contrast edges, and are most prevalent

Manuscript received December 9, 2005; revised October 26, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jelena Kovačević.

Parts of this work appeared in the IEEE 2005 International Conference on Acoustics, Speech, and Signal Processing and the IEEE 2005 International Conference on Image Processing. This work was completed while A. Leontaris was at the University of California, San Diego (UCSD). This work was supported in part by the National Science Foundation, in part by the Center for Wireless Communications at UCSD, in part by the Office of Naval Research, and in part by the University of California Discovery Grant program of the State of California.

A. Leontaris is with Dolby Laboratories, Inc., Burbank, CA 91505-5300 USA (e-mail: aleon@dolby.com).

P. C. Cosman is with the Information Coding Laboratory, Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093-0407 USA (e-mail: pcosman@code.ucsd.edu).

A. R. Reibman is with AT&T Labs—Research, Florham Park, NJ 07932-0971 USA (e-mail: amy@research.att.com).

Digital Object Identifier 10.1109/TIP.2007.891778

[1]In [1], the term *blocking* referred to the sum of both on-grid blocking, as well as motion-compensated edge artifacts. In this work, blocking refers to the traditional on-grid impairment.

in wavelet coders. *Motion-compensated edge artifacts (MCEA)* are typical in video codecs that use block-based MC prediction. When coarse quantization is combined with MC prediction, blocking artifacts propagate from I-frames into subsequent frames and accumulate, causing structured HF noise that is no longer located at block boundaries (off grid). Fractional-pixel MC and edge-attenuating filters can reduce this artifact. By definition, the MCEA involve HF noise within the blocks, while the blocking impairment involves HF noise along the block boundaries (on grid). These motion-compensated edge artifacts were called "false edges" in [6]. An in-depth discussion of these artifacts can be found in Section III.

The estimation of blocking, blurring, and ringing with the help of HVS models was the scope of [7]. All three can be encountered in both compressed still images and compressed video. Most research on video quality assessment has concentrated on blocking and blurring. These components are not orthogonal; interactions between two artifact types affect the perceived strength of each. Farias *et al.* [8] indicated that when artifacts are perceived to have equal strength, blurriness is more annoying than blockiness. One approach to quality assessment involves the use of HVS principles to determine a single value-index that characterizes the overall video quality [9], [10]. An alternative is to design metrics that assess a single impairment type, such that the impact of multiple impairments can be subsequently combined into a single quality value [11], [12]. In this paper, we concentrate on evaluating single-impairment quality metrics in the context of video compression, without transmission errors.

Reviews of image and video quality metrics include [13]–[17]. A comprehensive overview of monochrome image quality metrics dating from 1974 to 1999 was presented in [13], which focuses on describing HVS models without evaluation. Several monochrome FR image quality metrics were evaluated in [14], using the correlation between the metric output and a subjective evaluation. In [15], four quality metrics were evaluated in an error-prone environment: the SNR, PSNR, ITS [18] (an objective metric based on subjective tests), and MPQM [9] (an objective metric based on HVS). Three still FR image metrics that use HVS criteria were compared in [16] using subjective tests. Three quality metrics [19]–[21] were evaluated for video streaming in [17], which indicated the metrics had difficulty correctly ranking the quality produced by different codecs.

Available blocking metrics were primarily developed for use with image codecs; thus, they ignore motion-compensated edge artifacts. With the help of subjective tests, we show that the perceptual effect of this unexplored visual impairment is significant. We propose a novel quality assessment measure that estimates the DCT energy of the MCEA.

Our goal is to test the ability of available similarity, blocking, and blurring metrics to correctly order the subjective impact of different 1) spatial content, 2) quantization parameters, 3) amounts of filtering, 4) distances from the most recent I-frame, and 5) long-term frame-prediction strategies. Because many metrics *appear* to work well when averaged over an entire video sequence, we explore their performance on a per-frame basis. Our evaluation consists of two parts: the ordinal scale and the interval scale evaluation.

The ordinal scale evaluation is comprised of both an objective and subjective evaluation. In the objective evaluation, expectations are derived to systematically explore the impact of multiple parameters that affect compressed video quality, based on common sense. This objective evaluation exposes several inadequacies in the performance of many metrics. We then describe a subjective evaluation methodology designed to verify the logic and intuition behind our expectations and to further evaluate the metrics. All metrics undergo the objective evaluation; the ones that perform best are further evaluated using more exhaustive subjective tests. Interval scales are obtained from the same subjective data set of the ordinal evaluation with the help of Thurstone's Law to further study these metrics.

The objective evaluation framework is useful to anyone designing quality metrics for compression. In contrast, subjective tests are often constrained with respect to the examined parameters and their conclusions are not easily applicable to cases other than the original experiment. Hence, it makes sense to evaluate metrics both with objective and then with subjective tests.

The paper is organized as follows: Existing evaluated metrics are summarized in Section II. We propose a new metric to calculate the effect of propagating edge artifacts in Section III. We derive our expectations for the ordinal scale objective evaluation in Section IV. We then introduce the subjective evaluation methodology in Section V. The results of the ordinal scale objective evaluation are described in Section VI. The worst performing are weeded out, and the surviving metrics are further studied through the subjective tests. The same metrics are finally studied with the help of subjective test interval scales in Section VII. The paper is concluded in Section VIII.

II. Existing Quality Metrics

In this paper, we consider three similarity, nine blocking, and three blurring metrics. All of the similarity and one of the blocking metrics are FR measures. The rest are NR metrics.

A. Three Similarity Metrics

The similarity metrics all require both the original and degraded images. Similarity metrics are also known as comprehensive quality metrics. *PSNR*: MSE and, equivalently, PSNR use a pixel-by-pixel comparison between two images. *SSIM*: The structural similarity index (SSIM) [22] uses means, variances, and correlations of both images. *PIQE-S*: The psychovisual image quality evaluator (PIQE) [23] is a FR method consisting of two parts: a blockiness component and a similarity component, denoted PIQE-S, which counts the number of edges common to both the original and degraded image.

B. Nine Blocking Metrics

We consider one FR and eight NR blocking metrics. None of the evaluated blocking metrics uses temporal masking, or is parameterized with respect to the viewing distance. *PIQE-B*: The blockiness component PIQE-B of PIQE [23] uses the DCT DC coefficients of the current block and its eight neighboring blocks of both the decoded and the original frame to compute a FR blocking measure. *BD*: The boundary discontinuity (BD) [24] metric was initially proposed to identify frame areas with increased blocking artifacts so that de-blocking algorithms

may be applied. The boundary discontinuity is defined as the amount of the slope increase in the boundary pixels as compared to the slopes of the internal pixels of a block. *MSDS*: This metric was proposed in [25], and shares the same original motivation (de-blocking) as well as the same assumption with the BD metric. *BAM*: Gao *et al.* [26] present a blocking artifact metric (BAM) based on differences of averages of eight-pixel rows (columns) across vertical (horizontal) block edges. *Phase Correlation (PC)*: In this method [19], the denominator of the metric measures interblock similarity while the numerator measures intrablock similarity.

The remaining four blocking metrics all incorporate some form of HVS modeling. *GBIM*: In the generalized block impairment metric (GBIM) [21], HVS masking is incorporated by means of weights derived from localized averages and standard deviations across pixel block boundaries. *Power Spectrum*: In [20], the power spectrum of the 1-D absolute difference signal is calculated using the fast Fourier transform. Luminance and texture masking are exploited to scale the signal prior to the frequency analysis. *DCT-Step*: The DCT-Step metric [27] models blocking artifacts as 2-D step functions, and weighs results using local background luminance and activity masking measures. *PSBIM*: The perceptual block impairment metric [28], which modifies GBIM to include more comprehensive luminance masking, has a similar structure to the phase correlation metric in [19]. In PSBIM, the numerator represents edge strength, while the denominator represents the inner-block spatial similarity.

C. Three Blurring Metrics

All three evaluated blurring metrics are NR metrics. *DCT-Histogram*: The first blurring metric computes a global blur for an image using a histogram of DCT coefficients gathered from the compressed bitstream [29]. *Edge-blur*: The second metric computes the spatial extent of each edge in an image using inflection points in the luminance to mark the start and end of an edge [30]. The spread of edges in an image is estimated by observing the smoothing effect of blur on edges. *Kurtosis*: The third blurring metric computes sharpness, the inverse of blurring, by calculating local edge Kurtosis [31]. Kurtosis is interpreted as the extent to which the signal is non-Gaussian.

III. Motion-Compensated Edge Artifacts

An example of MCEA appears in Fig. 1. This 42 × 37 pixel segment, with its top left corner at the (207 155) pixel of frame 21 of the image sequence "foreman" at CIF resolution (352 × 288), was encoded using the H.263+ encoder with QP set to 22. The original content is shown in Fig. 1(a). The same frame is encoded first as an I-frame, and then as a P-frame with $d = \{1, 6, 14\}$, where d is the distance from the last I-frame. In Fig. 1(b), the regular blocking grid is well perceived as it is a portion of an I-frame. P-frame coding with $d = 1$ in Fig. 1(c) is similar to the previous case with minor spatial displacements of some of the block edges. For Fig. 1(b) and (c), the spatial content within each block has low spatial frequency. As d increases, we observe significant changes in Fig. 1(d)–(e). Not only do block boundaries of the blocking grid dissipate, but high-frequency artifacts, that are not part of the original image content, appear within the block boundaries.

All blocking metrics presented in Section II assume constant blocking boundaries, ignoring propagation of blocking artifacts and are, therefore, not designed to measure these artifacts in P-frames. Pixel-based [21], [28] metrics require exact knowledge of artifact location, which is difficult to achieve due to the combination of fractional-pixel MC and variable-sized blocks. It would also be very difficult to modify frequency-based blocking metrics such as PC and Power Spectrum to measure these artifacts. These methods rely on the periodicity of the blocking grid [see Fig. 1(b)], which vanishes as d increases. Therefore, we develop a new metric to measure the energy of these MC edge artifacts.

A. Scheme Formulation

Our approach involves the calculation of MCEA energy on a block basis. Let the set of all 8 × 8-pixel blocks in a frame be T. The *received* DCT coefficient in the compressed bitstream (transmitted prediction residuals) at location (i, j) of an 8 × 8 block $\tau \in T$ in frame n is $c_\tau^n(i, j)$. We consider the set N of all AC DCT coefficients. The *measured* DCT coefficient obtained from the 8 × 8 DCT transform of the reconstructed frame is $m_\tau^n(i, j)$. Finally, $s_\tau^n(i, j)$ denotes a DCT coefficient from the block in the original *source* frame, and $p_\tau^n(i, j)$ denotes a DCT coefficient from the motion-compensated *prediction* block.

To encode the source, the encoder selects a motion-compensated prediction block because it is the best fit overall according to some criterion (e.g., sum of absolute differences). The encoder compresses and transmits the prediction residuals $s_\tau^n(i, j) - p_\tau^n(i, j)$ with available bits, resulting in quantized residuals $c_\tau^n(i, j)$, that help reduce prediction errors. The transmitted residual DCT energy is

$$C_\tau^n = \sum_{(i,j)\in N} \left(c_\tau^n(i, j)\right)^2 \tag{1}$$

and the measured DCT energy is

$$M_\tau^n = \sum_{(i,j)\in N} \left(m_\tau^n(i, j)\right)^2. \tag{2}$$

We can compute the energy P_τ^n in the prediction block exactly, using the measured DCT coefficients in the reconstruction, $m_\tau^n(i, j)$, and the received coefficients, $c_\tau^n(i, j)$

$$P_\tau^n = \sum_{(i,j)\in N} \left(m_\tau^n(i, j) - c_\tau^n(i, j)\right)^2 \tag{3}$$

P_τ^n, C_τ^n, and M_τ^n are known exactly, given the decoded pixels and the compressed bitstream. We seek to design a NR metric that estimates the percentage of high-frequency energy in M_τ^n that is not part of the original image content.

The coefficient energy sent by the encoder in C_τ^n reduces the visual impact of the error $s_\tau^n(i, j) - p_\tau^n(i, j)$. The AC image distortion is, thus, $(s_\tau^n(i, j) - p_\tau^n(i, j)) - c_\tau^n(i, j)$. Our metric seeks to estimate the HF energy $(P - S)$ added through MC and not removed by the transmitted coefficients with energy C.

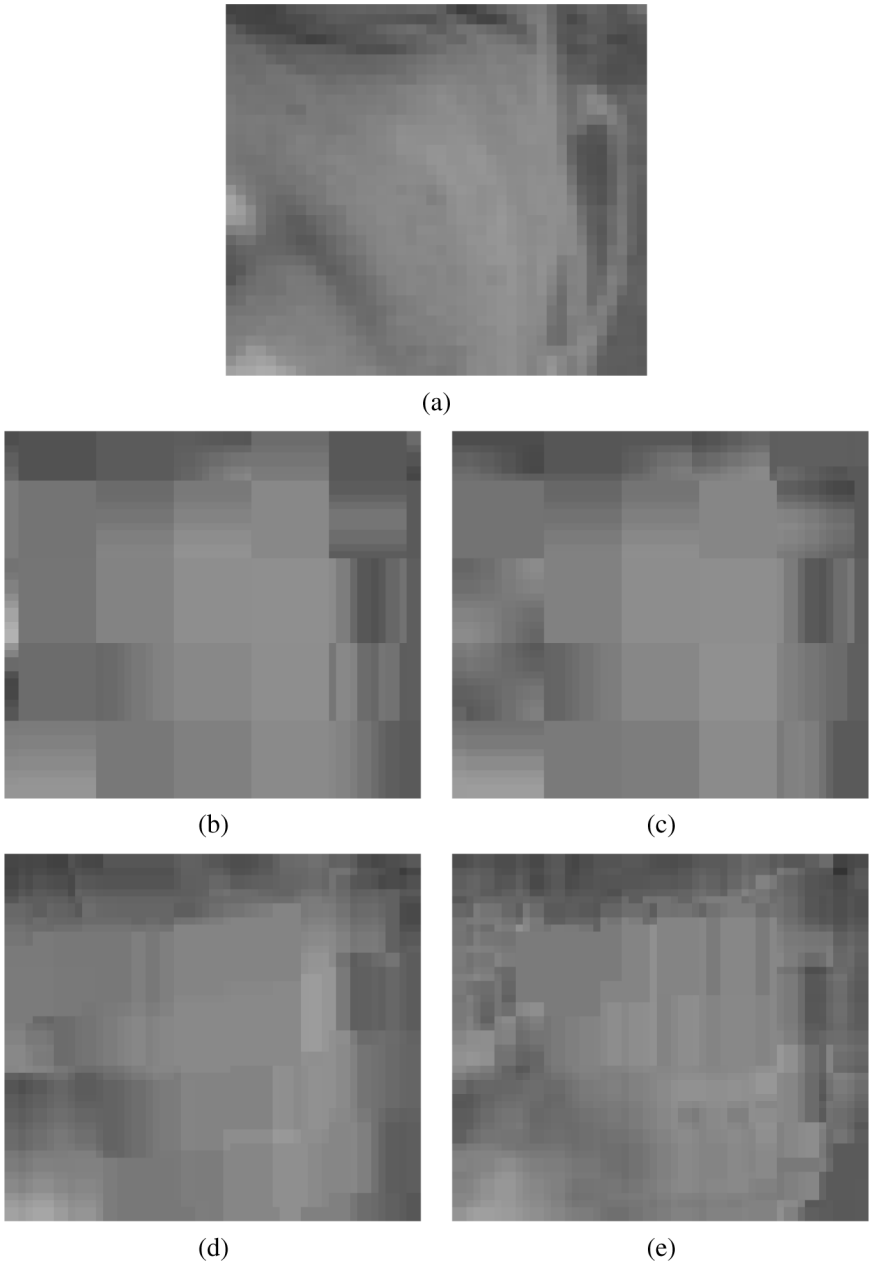


Fig. 1. Visual sample of propagating motion-compensated edge artifacts in "foreman." (a) Original content, (b) I-frame, (c) P-frame $d = 1$, (d) P-frame $d = 6$, and (e) P-frame $d = 14$.

HF energy is added and removed through the AC coefficients, which characterize the frequency content of the signal. As a result, we do not include the DC coefficient in our calculations. Hereafter we refer to our proposed metric as MCEAM, to differentiate it from the MCEA artifacts it measures. We formulate it on an energy basis $(P_\tau^n(i,j) - S_\tau^n(i,j)) - C_\tau^n(i,j)$. However, $S_\tau^n(i,j)$ is unknown at the decoder and must be estimated. We denote this estimate $E_\tau^n(i,j)$. This estimate is close to the energy of the prediction block in the previous frame. The prediction blocks will often overlap the blocking grid, and, thus, their energy content may be contaminated with the energy of blocking artifacts. To alleviate this, the energy of the unknown source block is calculated as the weighted average energy of the four on-grid blocks in the past decoded frame that overlap the prediction block. These blocks carry fewer blocking artifacts within them. We assume that the energy of regular grid blocks in a small neighborhood does not change significantly from one frame to the next.

To estimate the source energy for a given block, τ, we let $\sigma(\tau)$ indicate the set of (up to) four on-grid blocks in frame $n-1$ that are used to predict block τ in frame n. The prediction of τ uses $w(\beta)$ percent of the block $\beta \in \sigma(\tau)$. Then, the energy estimate

$$E_\tau^n = \sum_{\beta \in \sigma(\tau)} w(\beta) \sum_{(i,j) \in N} \left(m_\beta^{n-1}(i,j) \right)^2 \tag{4}$$

approximates the energy content of the source for the block τ in the current frame.

The recursive per-block MCEAM energy metric is obtained

$$\mu_\tau^n = (P_\tau^n - E_\tau^n) - C_\tau^n + \sum_{\beta \in \sigma(\tau)} w(\beta)\mu_\beta^{n-1}. \quad (5)$$

The last term is the propagated MCEA energy from the previous referenced blocks. Note that $(P_\tau^n - E_\tau^n) - C_\tau^n$ can be negative, indicating that the transmitted energy C_τ^n was enough not only to counter the potential new MCEA energy $(P_\tau^n - E_\tau^n)$ but also to reduce previously propagated MCEA energy. Furthermore, μ_τ^n is constrained to be non-negative.

This block-based approach addresses the occurrence of skip and I-blocks with the efficient use of recursion. For example, if the current block is a skip block, the MCEA energy is set equal to that calculated for the co-located block in the previous frame. Similarly, for I-blocks, we merely set it to zero as it is by definition.

In those few cases that the transmitted energy C_τ^n is found to be greater than the HF energy added through MC $(P_\tau^n - E_\tau^n)$ *and* greater than the increase in local energy (measured energy M_τ^n minus the estimated energy E_τ^n), the calculated metric μ_τ^n for the block is set to be zero disregarding previously accumulated energy. Intuitively, if the transmitted DCT energy was enough to offset $(P_\tau^n - E_\tau^n)$, and was again larger than the increase in local energy, we can speculate that it was enough to counterbalance all previously propagated MCEA energy (since the increase in energy can be solely attributed to the image content). The final metric is then scaled to incorporate texture masking

$$\text{MCEAM} = \frac{\sum_{\tau \in T} \mu_\tau^n}{M^n} \quad (6)$$

where $M^n = \sum_{\tau \in T} M_\tau^n$ is the measured energy content of the entire frame n. *MCEAM* is an estimate of the percentage of DCT energy in the reconstructed video frame that is caused by motion-compensated edge artifacts.

IV. Ordinal Expectation Framework

Here, we derive a comprehensive ordinal scale *objective* evaluation framework that systematically changes parameters affecting compressed video quality to sample a significant part of the image quality parameter space.

Typical of many video quality assessment studies, video sequences were encoded at a single rate in [17] and the sole parameter explored was the frame number. Quality perception is affected by many parameters, including the six we consider: 1) QP of the I-frames, Q_I, 2) QP of the P-frames, Q_P, 3) distance d between the current frame and the most recent I-frame, 4) video content: static versus high-motion activity, 5) presence of edge attenuating filtering, and 6) the video codec. We note that filters can be applied before compression (prefiltering), during compression (in-loop filtering), after reconstruction (postprocessing), or implicitly by using OBMC and fractional-pixel MC. For video codecs, we considered H.263+ and H.264/AVC. These differ, in part, in the size of their block partitions (minimum blocksize of 8×8 and 4×4, respectively) and the degree of fractional-pixel MC prediction (half-pixel and quarter-pixel accurate MC, respectively).

To help identify metrics that perform inadequately, we rely on common sense to create a list of expectations that a well-designed metric should satisfy.

Foreman is characterized by large moving uniform areas, while *coastguard* has extensive high-frequency detail that masks artifacts. Common sense dictates that *foreman* will have more visible blocking artifacts compared to *coastguard*. This leads to the first expectation.

A) For the same QP and no filtering, *coastguard* is less blocky and has less MCEA than *foreman*. Also, the quantized *coastguard* is more similar to its original than is the quantized *foreman*. This expectation should hold across all codecs and frame-types.

B) Without filtering, i) blockiness for I-frames increases as QP increases, ii) MCEA in P-frames increases as QP increases, and iii) similarity for both I- and P-frames decreases as QP increases.

C) With edge-attenuating filtering, blurriness increases and similarity decreases as QP increases.

D) For fixed QP and d, more filtering decreases blockiness, MCEA, and similarity but increases blurriness. It should hold across all possible codecs, frame-types, and spatial content.

E) For fixed QP and filtering and for a single reference frame, i) blurriness increases with d, ii) the sum of blockiness and MCEA increases (because artifacts accumulate) with d, and iii) similarity decreases as the distance d from the most recent I-frame increases [1].

F) For fixed QP, filtering, and d, using a long-term (LT) high-quality reference frame improves quality; namely, it increases similarity and reduces blurring, MCEA, and blocking. This expectation is based on observations using *coastguard* with and without LT prediction. The section of *coastguard* we considered particularly benefits from LT prediction because occluded areas were uncovered.

V. Subjective Evaluation

Next, we conduct subjective tests to verify the validity of our ordinal expectations discussed in Section IV, and to yield a ground truth data set against which good metrics can be compared. The method of paired comparisons [32] is used. We derive both ordinal scales as well as interval scales from the obtained data set. The experimental results are analyzed and converted into interval scales using Thurstone's Law of Comparative Judgment [33].

A. Why Paired Comparison?

In a subjective experiment, *ordinal scale* information is easy to obtain. Objects (in our case frames encoded with different coding parameters) are ranked according to the underlying property of interest (e.g., visual quality). The magnitude of the difference between objects is not determined. With an *interval scale* (e.g., Celsius temperature scale), the magnitude of the differences between scale values indicates the extent to which one object will be preferred over another. Finally, with a *ratio scale* the ratios of differences are equal across the range of the scale [34].

Ratio scales (e.g., Kelvin temperature scale) require an "absolute" origin point. In our subjective experiment, we seek to explore both the ordinal and the interval scales.

Three general methods are available to obtain interval scales. First, with the *rank-order* procedure, the test subject orders a small set of objects. With one observer, this produces an ordinal scale, but by combining results of many observers, one can obtain an interval scale [35], [36]. Second, a *single stimulus* test involves the presentation of one stimulus (a motion video [3], or a still video frame as in our case). The test subject evaluates the stimulus by assigning values from a continuous interval scale. This has many weaknesses [34]: scales of test subjects have unequal intervals, scales do not correspond between observers (personal, social, and situational factors can affect the test subject's results), and scales can be inconsistent within one observer over time. Third, *paired comparisons* [32], address most of the above problems. The test subject is presented with two stimuli and asked to select one. All three problems are ameliorated, since only *binary* decisions are made. This robustness is a result of the minimal thought process that a forced-choice comparison involves.

Paired comparison [32] requires presenting all possible pairs of the investigated objects. To evaluate frames encoded with n different coding parameter values, the resulting $n(n-1)/2$ pairs have to be evaluated by at least $n(n-1)/2$ test subjects to ensure sufficient statistical reliability [37].

B. Ordinal Scales

To order the investigated stimuli, the following simple algorithm is employed: Each stimulus is assigned a numerical score equal to the number of times that that particular stimulus was chosen over the others in all trials where it was presented. The stimuli are then ordered increasingly according to their numerical scores to produce the ordinal scales.

C. Interval Scales

Thurstone's Law [33] is used to derive the associated interval scale values, $S_i, 1 \leq i \leq n$. These are estimated as the average of the interval distances $d_{i,j}$ between the stimulus i and *all* other stimuli j, where $j \neq i$ (hence, the need to conduct all pairwise comparisons), as

$$\tilde{S}_i = \frac{1}{n} \sum_{j=1, j \neq i}^{n} d_{i,j}. \tag{7}$$

The interval distances are estimated as follows: Let S_1 and S_2 be interval scale values of two stimuli. When a test subject compares the stimuli pair of 1 and 2, two internal HVS responses s_1 and s_2 are elicited, modeled as gaussian distributed random variables. The S_1 and S_2 interval scale values can be thought of as the "mean values" of the internal HVS responses s_1 and s_2. When the response s_1 is larger than response s_2, stimulus 1 is preferred; otherwise, stimulus 2 is preferred. Thurstone's Law is now written

$$S_1 - S_2 = \chi_{1,2} \sqrt{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{8}$$

where $\chi_{1,2}$ is the normal deviate corresponding to the proportion $p_{1>2}$ of outcomes where 1 is selected over 2, i.e., $s_1 - s_2 > 0$. The standard deviations of random variables s_1 and s_2 are σ_1 and σ_2. The correlation between them is ρ. Assuming that the s_i are uncorrelated ($\rho = 0$) and that $\sigma_1 = \sigma_2$, Thurstone's Law can be simplified as

$$S_1 - S_2 = \chi_{1,2} \sqrt{2}\sigma = d_{1,2}. \tag{9}$$

This is known as Case V of Thurstone's Law. The unit normal deviate $\chi_{1,2}$ is equal to $Z^{-1}(p_{1>2})$, where Z^{-1} is the inverse cumulative normal distribution function (inverse Z-score).

The ordering obtained through intervals is similar but not identical to that obtained with the trivial ordinal scale method. Thurstone's Law obtains the numerical score by summing the inverse Z-score of the preference ratios. The trivial scheme sums the ratios themselves. Due to the nonlinearity of the Z-score, the ordering can be different.

D. Subjective Testing Methodology

The H.263+ video codec from the software library FFMPEG of the video player/encoder package MPlayer-1.0pre4 (available from http://www.mplayerhq.hu/) was used to encode the test sequences. The motion vectors were obtained with half-pixel accuracy (introducing weak blurriness and block edge artifact attenuation). The in-loop filtering of H.263+ was used to provide loop-filtered sequences. A minimum block-size of 8×8 pixels was adopted. INTRA frames were inserted every 15 frames. Two image sequences were examined at CIF 352×288 resolution with YUV 4:2:2 color representation: the medium-low motion video-conferencing sequence "foreman" and the medium-high motion "coastguard" sequence, with camera panning and water ripples. We evaluated the effects of the quantization parameter QP and the distance d from the I-frame.

Thirty experts and nonexperts took part in the experiment, and were compensated for participating. Each paired comparison was repeated thirty times, yielding $T = 30$ trials. Six preliminary paired comparisons with frames other than those used in the test were conducted to train the test subjects. All possible pairwise combinations of the examined parameter space were investigated. Two decoded frames from sequences encoded with different coding parameters were presented to the viewer at the same time. The spatial content was the same (same frame). Before the next pair, a gray image was shown for half a second to eliminate contextual effects. Left/right positioning was random, as was the order in which image pairs were presented. Each pair was viewed only once. The viewer was asked to evaluate the frame pair. Three questions were asked:

1) "Which of the two images is *blockier*?"
2) "Which of the two images is *noisier*?"
3) "Which of the two images is *worse*?"

During the preliminary training comparison, it was made clear to the viewers, both orally and by displaying characteristic examples, that "blockier" refers to traditional on-grid blocking artifacts, and "noisier" refers to off-grid edge artifacts (MCEA). The "worse" attribute refers to the overall image quality. It is not meant to be a sum of "blockier" and "noisier." The "worse"

attribute will be discussed in Section VII. We are interested in correlations of "blockier" and "noisier" with "worse."

Two categories of subjective pair comparisons were conducted. Neither used loop filtering. First, with d set to 0, we showed all possible pairwise combinations of $QP = 10$, 15, 20, 25. In this case, $n = 4$ and $n(n-1)/2 = 6$; hence, $T = 30$ trials are enough. The same was done for $d = 14$. Second, the QP was fixed to three values: 10, 20, and 30. All pairwise combinations of $d = 0$, 1, 6, 10, 14 were shown. In this second case, $n = 5$ and $n(n-1)/2 = 10$; hence, $T = 30$ trials are sufficient. Results are presented in Section VII.

VI. Ordinal Scale Results

In this section, we compare the metrics and identify those performing poorly using the systematic expectation framework introduced in Section IV. The worst-performing metrics are excluded from further analysis. In Section VII, the surviving metrics are compared using the subjective test described in Section V. Ordinal scale information is solely used in this section.

A. Simulation Framework

Three cases are considered. The first, **All I-frames**, consists only of I-frames with varying Q_I. We explore expectations A–D for H.263+ and H.264/AVC with and without loop filtering across different spatial content. The second, **P-frames**, consists of an I-frame followed by multiple P-frames, where we vary the quantizer while keeping $Q_I = Q_P$. For this case, we explore expectations A–E, focusing on three values of $d : d = 1$, 6, 14. The third case, **P-Frames LT**, is designed to examine expectation F, the visual impact of using a high-quality LT prediction frame in H.264/AVC [38]. Here, we fix both $Q_I = 18$ (nearly lossless) and $Q_P = 32$ (medium quality) and vary d across all values from 1 to 14.

To isolate the impact of spatial content, we choose identical frames for the comparison, regardless of the prediction structure. Thus, for frame number 21 to use $d = 1$, we set frame 20 to be an I-frame, while to achieve $d = 14$, frame 7 is an I-frame. We examine two sequences, *foreman* and *coastguard* (frames 21 and 141, respectively). "Foreman" exhibits large uniform moving objects which when coarsely quantized lead to extensive blocking artifacts. "Coastguard" is characterized by continuous motion, but texture masks most of the compression artifacts. We evaluate our proposed MCEAM metric on two additional video sequences: "mother-daughter," a highly static sequence, and "mobile-calendar," a sequence with extreme spatial (texture) masking.

For the tests on **All I-frames** and **P-frames**, we vary the quantization parameter using constant increments starting from nearly lossless (2 for H.263+ and 18 for H.264/AVC) to nearly unwatchable (30 for H.263+ and 45 for H.264/AVC). Our H.263+ test sequences are computed using the H.263+ codec in MPlayer/MEncoder [39]. For H.263+, we considered no filtering, 8×8 block motion vectors and OBMC. The H.264/AVC test sequences, with and without loop filtering, are generated with the JVT reference software version JM 8.2 using CABAC.

B. Auxiliary Comparison Tools

The monotonicity requirements of expectations B and C are evaluated with the help of Kendall's τ. The MCEAM metric is further evaluated with the help of a simple FR metric.

In addition to holding across all possible codecs and spatial content, expectations B, C should also be monotonic. For example, as QP increases, blockiness and MCEA should increase monotonically. To characterize how well a blocking or blurring metric is able to capture this monotonic increase, we use Kendall's tau, τ_a [40], which is an estimate of the probability that a pair of variables is more likely to be correctly ordered than incorrectly ordered. For a set of data $\{x_i\}$, $i = 1, \ldots, K$, which should always increase, Kendall's tau is defined to be $\tau_a = (\gamma - \delta)/(\gamma + \delta - \epsilon)$, where γ is the number of pairs (x_i, x_j), $i < j$ for which $x_i < x_j$ (i.e., the number of pairs correctly ordered), δ is the number of pairs incorrectly ordered, and ϵ is the number of pairs for which $x_i = x_j$. Note that $\gamma + \delta + \epsilon = K(K-1)/2$. For completely monotonic data, $\tau_a = 1$. As pairs become incorrectly ordered, τ_a decreases.

We preferred Kendall's τ over Spearman's rank correlation (r). Spearman's r is satisfactory for testing a null hypothesis of independence between two variables but is difficult to interpret when the null hypothesis is rejected. Kendall's τ improves upon this by reflecting the strength of the dependence between the variables being compared. Kendall's statistic has greater universality and has an intuitively simple interpretation [41], [42].

As a further reference for comparison, we designed a simple FR metric, denoted FR-MCEAM, that calculates the actual added HF energy due to MC. As the original signal $s_\tau^n(i,j)$ is available to an FR metric, we calculate the energy difference on a block basis between the reconstructed video and the original sequence. The FR-MCEAM metric is

$$FR_{\text{MCEAM}} = \sum_{\tau \in T} \left| \sum_{(i,j) \in N} \left(m_\tau^n(i,j)\right)^2 - \sum_{(i,j) \in N} \left(s_\tau^n(i,j)\right)^2 \right|. \tag{10}$$

C. Objective Evaluation Results Discussion

Table I summarizes how well each metric satisfies the expectations. Ability to satisfy expectations A, D, E, F is indicated by "Y," while inability is indicated with "x." Results for expectations B, C show the minimum value of Kendall's τ_a achieved across the set of situations considered for the given video codec (H.263+ or H.264/AVC). Recall that $\tau_a = 1$ means that a metric completely satisfies this expectation (perfect monotonicity), while negative τ_a strongly indicates an inability to satisfy this expectation. For expectation B, we consider I and P frames separately.

1) Similarity Metrics: Results for the FR similarity metrics are shown in Table I(a). Expectation A is satisfied solely by PIQE-S. Expectation B is satisfied by all three metrics exhibiting good Kendall τ monotonicity values. The metrics' performance seems unaffected by the choice of codec (H.263+ or H.264/AVC). Finally, expectations D–F are all satisfied by PSNR and SSIM. However, PIQE-S fails in all of them for reasons explained later in this section.

TABLE I
ABILITY OF EACH METRIC TO SATISFY EXPECTATIONS A–F. A "X" IN COLUMNS A, D, E, F MEANS METRIC FAILED EXPECTATION; "Y" DENOTES SATISFACTION. COLUMNS B, C DENOTE KENDALL'S τ_a CHARACTERIZING MONOTONICITY. (A) SIMILARITY, (B) BLOCKING, AND (C) BLURRING METRICS

method	A	B(All I-frames)		B(P-frames)		C		D	E	F
		H.263+	H.264/AVC	H.263+	H.264/AVC	H.263+	H.264/AVC			
PSNR	x	1.000	0.994	1.000	1.000	1.000	0.994	Y	Y	Y
SSIM [22]	x	1.000	0.994	0.995	1.000	0.965	0.994	Y	Y	Y
PIQE-S [23]	Y	0.955	0.947	0.921	0.968	0.946	0.947	x	x	x

(a)

method	A	B(All I-frames)		B(P-frames)		D	E	F
		H.263+	H.264/AVC	H.263+	H.264/AVC			
PIQE-B [23]	x	0.759	0.984	0.980	0.994	x	Y	x
BD [24]	x	0.940	0.725	0.724	-0.994	Y	x	x
MSDS [25]	x	0.911	0.814	-0.901	-0.994	Y	x	x
BAM [26]	x	1.000	0.963	0.980	0.894	Y	x	x
PC [19]	x	0.606	0.730	0.310	0.089	x	x	x
GBIM [21]	Y	1.000	0.994	0.965	0.740	Y	x	x
Power Spectrum	x	1.000	0.968	0.916	-0.798	Y	x	x
DCT-Step [27]	x	0.970	0.851	0.768	0.650	Y	x	x
PSBIM [28]	Y	0.980	1.000	0.906	0.984	x	x	Y
MCEAM	Y	N/A	N/A	0.847	N/A	Y	Y	N/A

(b)

method	C		D	E	F
	H.263+	H.264/AVC			
DCT-Histogram	0.892	0.977	Y	x	x
Edge-Blur	0.956	0.952	Y	x	Y
Kurtosis [31]	0.591	0.724	Y	x	x

(c)

2) Blocking Metrics: Results are given in Table I(b). We note that qualitatively the GBIM graphs are representative of the performance of the other blocking metrics (omitted due to space constraints).

Expectation A was satisfied by GBIM, PSBIM, and MCEAM. "Foreman" was shown to be blockier than "Coastguard." MCEAM was further evaluated on two additional sequences: "mobile," where MCEA are masked by the abundance of HF image content, and "mother-daughter" where MCEA are practically invisible. It correctly rank-ordered the four sequences. The low motion "mother-daughter" is easy to encode and exhibits zero MCEA, so the MCEAM metric mostly measures image content and quantization noise. It correctly showed the percentage of MCEA energy as extremely low.

Expectation B requires that blockiness and MCEA increase with increasing QP. While most metrics did well for I-frames, PC and PIQE-B were clearly the weakest for H.263+. When H.264/AVC is used to encode the sequence, PC, BD, and MSDS perform the worst. PIQE-B, while unsatisfactory with H.263+, proved better when used on H.264/AVC video. The reason may be the combination of smaller transform size and the use of DC values. Since MC causes off-grid blocking artifacts, most metrics are at a disadvantage. For P-frames, BD, MSDS, and PC are clearly the weakest in H.263+, while for H.264/AVC, Power Spectrum shows a sharp degradation in performance. Still, PIQE-B, PSBIM, BAM, DCT-Step, MCEAM, and GBIM perform reasonably well in this situation.

Expectation D was not satisfied by three metrics: PSBIM, PIQE-B, and PC.

Expectation E is satisfied only by the FR PIQE-B metric and the MCEAM metric.

Expectation F is satisfied by one metric: PSBIM. Blockiness and MCEAM have small values for *coastguard* frame 141.

We then use Kendall τ to evaluate the ordinal scales of our MCEAM method and the FR-MCEAM metric. The result of the FR-MCEAM is taken to be the ground-truth and the output of MCEAM is compared against it by re-ordering the MCEAM output according to the ordering of the FR-MCEAM metric, and

calculating Kendall's τ. For "foreman," "coastguard," "mother-daughter," and "mobile-calendar" the values are 0.8473, 0.8030, −0.4680, and 0.8374. The Kendall τ is negative for "mother-daughter" for reasons given above. Our method behaves similarly to FR-MCEAM, and estimates the energy of MCEA with sufficient accuracy.

3) Blurring Metrics: As shown by the results for blurring metrics in Table I(c), all are effective at satisfying expectation D by showing that filtering increases blurriness, while none was able to satisfy expectation E, that increasing d increases blurriness. Only edge-blur was successful with F, the LT prediction, while Kurtosis was the weakest with regard to expectation C.

D. Reasons for Failure

The blocking metrics share a number of weaknesses for our sequences. First, only four incorporate some form of HVS modeling, but even so DCT-Step and Power Spectrum fail in expectation A. Second, BD and MSDS both assume that the encoding process preserves the inner-block structure and pixel slopes. Unfortunately, the experimental results indicate that this assumption does not hold for the codecs we considered. Third, many metrics appear to discard useful information when they compute blockiness. For example, PIQE-B and BAM use only DC coefficients on a block and row/column level, respectively. Further, BAM employs a very strong cutoff threshold for measuring an edge. In DCT-Step, the simple 2-D step function model discards many coefficients. Similarly, the Phase Correlation method [19] discards vital information during spatial subsampling. Fourth, as pointed out in [17], many blocking metrics assume blocking artifacts appear only on 8×8 block boundaries and are unable to measure MCEA in P-frames. A further problem is the assumption on the periodicity of the blocking grid. Finally, most of these metrics average the blockiness across the image. Thus, a very strong edge will be averaged with weaker edges. On the other hand, humans are likely to perceive blockiness using only the most visible blocking artifact. HVS masking is often used to give more weight to stronger edges, but often this is not enough.

None of the blurring metrics incorporates any HVS modeling. The Kurtosis metric and edge-blur are similar to PIQE-S, in that they are heavily dependent on obtaining reliable edge information. Edge-attenuating filtering during compression reduces the number of available edges in their sample space, which decreases the statistical reliability and consequently the performance of both Kurtosis and PIQE-S. However, the same is not true for the edge-blur metric. Even though fewer edges are detected, the fact that the edge-blur metric measures the extent of those edges and is not highly dependent on the actual number of measured edges, makes it robust to compression. Finally, DCT-Histogram which is based on receiving DCT coefficients to form a histogram, suffers performance degradation for heavily quantized P-frames, as received DCT coefficients become scarce.

E. Conclusions and Insight From Objective Evaluation

The quality metrics we consider all have some weakness in measuring the quality of still frames from compressed video. We derive our expectations using common sense and each metric is unable to correctly rank order images for at least one of our common-sense expectations. Several metrics also prove inadequate when applied on H.264/AVC video, since they are designed for 8×8 DCT blocks. Kurtosis and Power Spectrum are particularly weak for H.264/AVC due to its complex blocking structure.

The most challenging expectations for the metrics to satisfy, as a whole, are to correctly characterize (A) the impact of spatial content, (B) the impact of blocking and MCEA in P-frames, (E) the increased blurriness as the distance d from the most recent I-frame increases, and (E) the increased perceivable blocking and MCEA with increasing d. The inability to characterize the second and fourth of these are due to the fact that these metrics have been designed for images, and are then applied to stills taken from video which may have propagated MCEA. This impairment is what the MCEAM metric was designed to detect, and as these preliminary results show, it accomplishes its goal.

A combination of the best performing blocking metric and the MCEAM metric could yield a high-performance quality metric. Through this first evaluation, we identified the most promising blocking metrics. Four of them will be further evaluated: GBIM, PSBIM, DCT-Step, and BAM. The metric from BAM exhibits excellent τ values but fails expectation A. PSBIM fulfills it but fails the critical expectation D. DCT-Step performs slightly worse than BAM. Finally, based on our expectation framework, GBIM seems the most promising blocking metric. In the next section, we evaluate MCEAM and the two FR metrics: PSNR and SSIM, in addition to the above four blocking metrics. Hereafter, noisiness and added HF energy will be used interchangeably with MCEA.

F. Subjective Results

Ordinal scales of the quantization parameter QP and the distance d are obtained through the process outlined in Section V. Images encoded with various QP (or d) values are ordered according to the perceptual strength of the attribute in question ("block," "noise," and "worse"). The subjective ordinal scales obtained in this way are then compared with the objective orderings. Thus, the ability of each objective metric to correlate well with subjective results is evaluated. Before doing that, we are interested in investigating the interattribute correlations; one would like to know whether the "block" attribute has more similar ordering to the "worse" attribute than the "noise" attribute does, and vice versa. Last, we are motivated to use the subjective ordering information to verify the validity of our expectations described in Section IV. Expectation A is easy to verify visually. Expectations C and D are also valid since the low-pass filtering effects are known. In addition, expectation F is valid since access to a higher quality reference frame can only improve the prediction. Thus, expectations A, C, D, and F need not be verified further. However, we are particularly interested to verify expectations B and E, which characterize a metric's response for varying QP and d, respectively.

The subjective data set is analyzed as discussed in Section V to yield *ordinal scales* for the varying stimuli: $QP = \{10, 20, 30\}$, the QP ordinal scale, and $d = \{0, 1, 6, 10, 14\}$, the d ordinal scale. Since Paired Comparison is used, each stimulus is compared an equal number of times with all other stimuli. The rank ordering can be obtained as the number of times that

TABLE II
SUBJECTIVE TEST ORDINAL SCALE RESULTS FOR THE FOLLOWING DISTANCES d FROM THE PAST INTRA FRAME: $d = \{0, 1, 6, 10, 14\}$. THE LEFTMOST VALUE HAS THE WEAKEST PERCEPTIBILITY, WHILE THE RIGHTMOST VALUE IS THE MOST PERCEPTIBLE ACCORDING TO THE EVALUATED ATTRIBUTE

	ordinal scale				
block	14	10	6	1	0
noise	0	1	6	10	14
worse	6	1	14	0	10

the stimulus is preferred over other stimuli. When examining the ordinal scales, we treat all comparisons equivalently; for example, a comparison between $d = 0$ and $d = 1$ is treated the same as a comparison between $d = 0$ and $d = 14$. The interval scales, considered later in Section VII, will not make this assumption.

We start with the evaluation of the QP ordinal scales. Kendall's τ is calculated on the experimental data set, and yields 1 for all combinations of sequences and distances d. All three attributes, "blockier," "noisier," and "worse" are strictly increasing in perceptibility as QP increases. This verifies the validity of expectation B. Consequently, the discussion in Section VI-C, which examines how well the metrics satisfy this expectation is valid, as well.

We continue next with the case of the d ordinal scales. The rank ordering of d is presented in Table II. We sum for every d the number of times it is preferred compared to all other d's over all sequences and QP combinations. We observe that the "blockier" attribute clearly decreases with increasing d, as is expected. The "noise" attribute clearly increases with d. However, the "worse" attribute behavior is neither increasing nor decreasing. This indicates that the overall impairment, represented by "worse," is affected by both "block" and "noise."

To further study the ordinal scale results and evaluate the metrics, we plot the interattribute correlations and Kendall τ values of the three test attributes with each of the investigated metrics in Fig. 2. Each star represents one combination of sequence and QP. For example, the top-left star in Fig. 2(a) is the correlation coefficient for sequence "foreman" and $\mathrm{QP} = 30$ between the ordinal scale of the "block" attribute, and the ordinal scale of the "noise" attribute.

The correlation coefficients between the attribute orderings are displayed in Fig. 2(a). We begin our discussion with the correlations: "block" has low correlation with "noise," while the same is observed between "noise" and "worse." On the other hand, "block" looks slightly correlated with "worse," but the correlation is still low. Moving now to Kendall τ values between metrics and attributes, we see in Fig. 2(b) that the PSNR, SSIM, and MCEAM metrics exhibit low Kendall τ scores with "block." In contrast, the blocking metrics share the ordering of the "block" attribute. Next, in Fig. 2(c), the PSNR, SSIM, and MCEAM metrics have similar ordering with the "noise" attribute, unlike the blocking metrics. On the other hand, in Fig. 2(d) it is challenging to indicate metrics that clearly share the same ordinal scale with the "worse" attribute. Using only the ordinal scales, we are not able to draw meaningful conclusions regarding the "worse" attribute.

Recall that expectation E stated that the *sum* of blocking and MCEA increases with d. The results from the ordinal scale evaluation prove, however, inconclusive with respect to this expectation. Further evaluation of the subjective data set with the help of interval scales will follow in the next Section to gain more insight on interattribute correlations and the ordering of the evaluated metrics with respect to the "worse" attribute. Still, the ordinal scale evaluation pointed to the validity of expectation B.

VII. INTERVAL SCALE RESULTS

The interval scales obtained from the subjective data set are first used to investigate correlations among the tested attributes: "block," "noise," and "worse." We evaluate the metrics briefly with the QP interval scale results and then extensively with the d interval results. We then seek to learn more on the metrics' performance with respect to the test attributes. The conclusions are used to support the expectations of Section IV, as well as to design a hybrid metric based on our findings.

A. Interattribute Correlations

Thurstone's Law is used to obtain the interval scales for increasing QP. All three attributes: "blockier," "noisier," and "worse" are shown to increase with QP, coinciding with the conclusions of the ordinal scale evaluation.

Conclusions for increasing d follow. The behavior of the interval scales is qualitatively similar to the ordinal scales: With increasing d, "blockier" decreases, while "noisier" increases. The "worse" attribute is neither clearly increasing nor clearly decreasing. It is correlated with both attributes, "blockier" and "noisier." Fig. 3(a) shows the spread of the correlation coefficients (marked with stars) between the three tested attributes for the d interval scales. Each star represents one combination of sequence and QP. Compared with the ordinal scale evaluation which is inconclusive on interattribute correlations, the interval scale evaluation yields valuable information. First, the "block" attribute has low correlation with "noise." Second, the "noise" attribute is more correlated to "worse," than "block" is. Since "worse" represents the overall impairment, and "noise" increases with d, we can conclude that "worse" also increases with d.

This last conclusion supports expectation E, which stated that the *sum* of blocking and MCEA increases with d. The rest of the expectations are self-evident and easily verifiable. We believe these findings show that motion-compensated edge artifacts have been overlooked and comprise a significant dimension of visual impairment. The test subject's response to the "worse" question allows us to investigate, with the help of attribute correlation coefficients, how blocking and MCEA contribute to the overall image impairment. Next, we study the relationship of the individual metrics with the test attributes.

B. Evaluation of Metrics With the QP Intervals

The correlation coefficients between the subjective interval scales and the objective FR and NR metrics results are also calculated. The FR metrics perform similarly with the NR blocking metrics. In general, all metrics exhibit high correlations with the subjective data set for varying QP. Conclusions are identical to those of the ordinal scale evaluation.

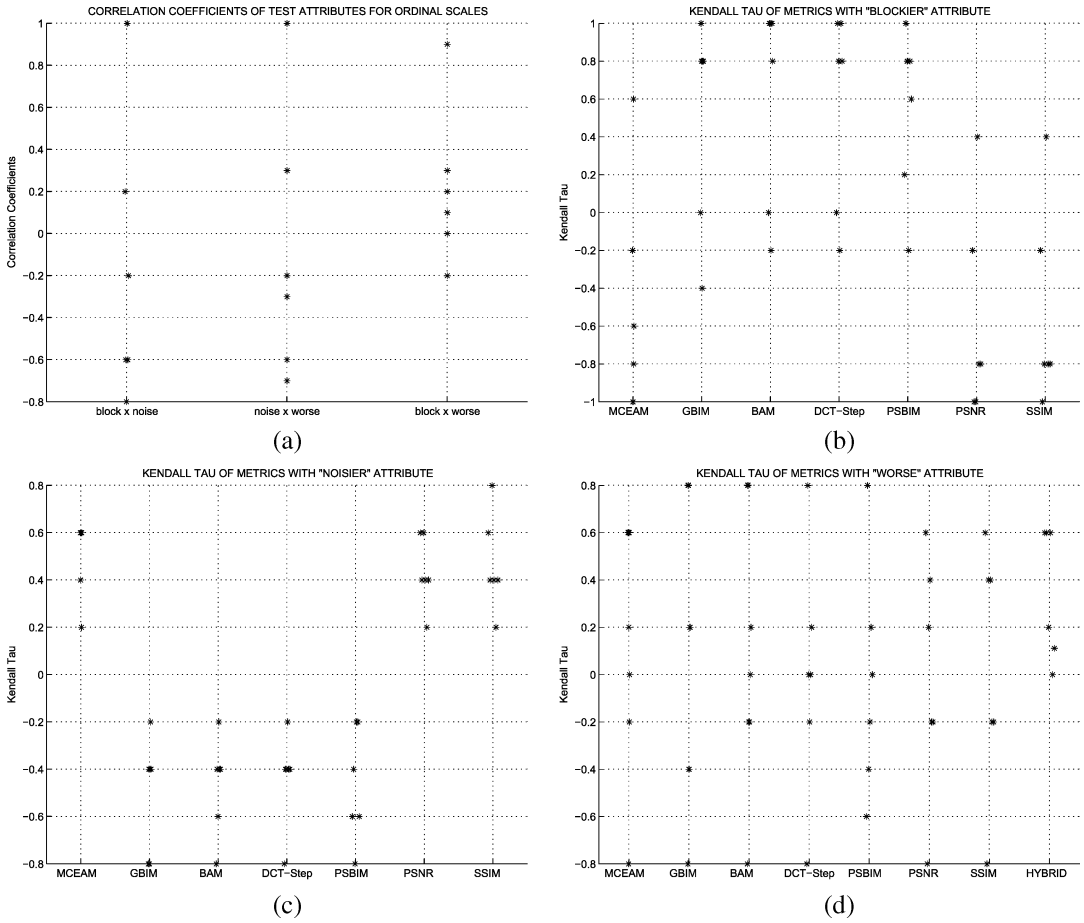


Fig. 2. Correlation coefficients and Kendall tau between d ordinal scale subjective test results and metrics. Cross correlation of the attribute (say "worse") with the output of the metric for the following distances d from the past INTRA frame: $d = \{0, 1, 6, 10, 14\}$. (a) Coefficients between test attributes, (b) "noise,"(c) "blockier," and (d) "worse."

C. Evaluation of Metrics With the d Intervals

Recall that MCEAM increases with d while GBIM decreases with d. Both increase for QP, and decrease for edge-attenuating filtering. The remaining three blocking metrics perform similarly to GBIM.

The spread of the correlation coefficients of the investigated metrics with the test attributes for the d intervals are shown in Fig. 3. Six values are plotted, similarly to the ordinal scale case, since two sequences and three QP values are investigated. In Fig. 3(b), the "block" component is shown to have low correlation with the MCEAM metric, but high correlation with the four blocking metrics. There is no statistical difference in the performance of the blocking metrics. Conversely, in Fig. 3(c) the MCEAM metric shows high correlation with the "noise" component throughout the parameter space. The blocking metrics on the other hand have low correlation with the "noise" component. Intuitively, these findings are to be expected. Moving our attention to the two FR metrics, we notice that they have similar correlation coefficients to those of MCEAM. They are more correlated with "noise" than "block." Fig. 3(b) and (c) shows that MCEAM behaves similarly to PSNR and SSIM. We show the correlation of the metrics with the "worse" component in Fig. 3(d). We observe that correlation coefficients of MCEAM, PSNR, and SSIM with the "worse" attribute are slightly more similar compared to the ones for the blocking metrics. It is encouraging that our NR MCEAM metric has similar performance with FR metrics.

D. Conclusion

Expectation E was shown to be valid. Furthermore, the rest of the expectations, besides B, are based on widely observed video encoder behavior and easy to verify. The subjective test showed that visual impairment is a combination of both blockiness and MCEA. The objective evaluation pointed to a slight edge of GBIM over the rest of the blocking metrics. The subjective evaluation that followed showed that the statistical difference among the blocking metrics is small. As a result, we are inclined to select GBIM as the best-performing blocking metric of this evaluation. Still, the other three blocking metrics exhibit very good performance, marginally lower than GBIM. Neither MCEAM nor GBIM can independently exhibit high correlation with the "worse" attribute throughout the test parameter space (QP and image sequence). Some combination of both, however, should be able to.

E. Hybrid Metric Combining MCEAM and GBIM

A simple hybrid metric is now described. Instead of being comprehensive, it serves to demonstrate that combining

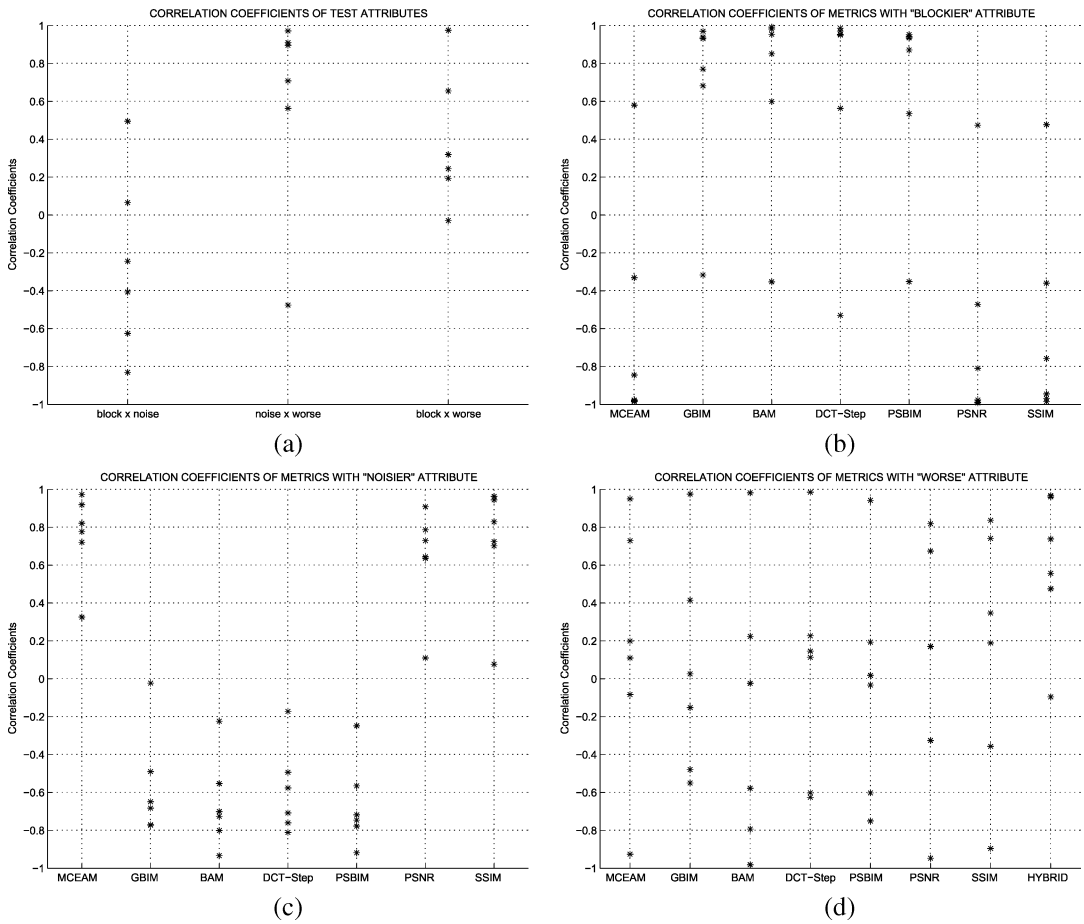


Fig. 3. Correlation coefficients between d interval scale subjective test results and metrics. Cross correlation of the attribute (say "worse") with the output of the metric for the following distances d from the past INTRA frame: $d = \{0, 1, 6, 10, 14\}$. (a) Coefficients between test attributes, (b) "blockier," (c) "noisier," and (d) "worse."

blocking and MCEAM yields better performance. This metric combines linearly the output of the GBIM and MCEAM metrics. The correlation coefficients of the metrics with the "worse" attribute (shown in Fig. 3) are used to select weighting coefficients for the linear combination. The conclusions of the previous paragraph motivate us to assign different weights according to the quantization level (which can be easily retrieved from the compressed bitstream). The weights were obtained through a thorough exploration of the QP parameter space. The same weighting coefficients are employed for both evaluated sequences. We write the hybrid metric as

$$\mathrm{COMP} = w_n \times \mathrm{MCEAM} + w_b \times \mathrm{GBIM}. \qquad (11)$$

The coefficients (w_n, w_b) are set to: (0.95, 0.05) for $QP = 10$, (0.15, 0.85) for $QP = 20$, and (0.50, 0.50) for $QP = 30$. We recalculate the correlation coefficients for the "worse" component and both sequences. The spread of the coefficients for the hybrid metric is shown in Fig. 3(d), as well as in Fig. 2(d). Comparing the above coefficients with those for MCEAM and GBIM we observe that the new metric achieves good performance, which would have been impossible by using either one of the two metrics alone.

VIII. Conclusion

The contributions of this work are summarized in the next few paragraphs in the following order: a) expectations, b) subjective test, c) metrics comparison, d) motion-compensated edge artifacts, and e) the MCEAM metric. We conclude with some thoughts on future work.

Objective quality metrics are traditionally evaluated with the help of subjective tests. However, subjective tests are often constrained with respect to the examined parameters and their conclusions are not easily applicable to cases other than the original experiment. Furthermore, they are costly in terms of time and resources. Consequently one should use them sparingly. We proposed to first conduct an objective evaluation to identify good quality metrics. Only if good quality metrics are identified should one initiate a subjective evaluation. The efficiency of this two-tier (objective-subjective) evaluation scheme depends on the criteria used in the objective evaluation. We hence designed a novel systematic objective framework to evaluate objective video quality metrics. Expectations derived through common sense were used to compile a collection of ordering criteria that a good video quality metric should satisfy. The expectations characterize the response of a metric for varying coding parameters such as QP and d, among others. The importance of this work is that this evaluation framework successfully identified good metrics without the help of subjective tests.

A subjective test was designed and conducted to validate our ordinal scale objective evaluation framework and further investigate the performance of the metrics. The data set was gathered through pairwise comparisons. Interval scale information was then extracted from the same data set using Case V of Thurstone's Law of comparative judgment. The interval scales proved an invaluable tool, both for evaluating the metrics, as well as exploring the relationships between the different types of visual impairment.

We employed the systematic expectations framework to evaluate several existing blocking, blurring, and similarity metrics. The systematic evaluation of metrics showed that most suffer from weaknesses. These include: a) lack of HVS modeling, b) the assumption that the encoding process preserves inner-block structure and pixel slopes, c) the assumption that blocking artifacts are located along the blocking grid of the DCT transform, d) the assumption on the periodicity of the blocking grid that breaks down when P frames are encountered, and e) the discarding of useful information; i.e., taking into account only the DC coefficients of an 8×8 DCT block. In terms of individual blocking metrics we found that GBIM, PSBIM, DCT-Step, and BAM performed the best. These were further evaluated with a comprehensive subjective test and GBIM was found to be slightly better than the other three. PSNR and SSIM were identified as good similarity metrics. After further subjective evaluation, PSNR and SSIM were found to be practically equivalent. Of the three blurring metrics, edge-blur performed well.

A new component of visual impairment was discussed and defined, which, while being ubiquitous in modern block-based video codecs, had not been investigated before. Called motion-compensated edge artifact, it is a direct consequence of motion-compensated prediction. The subjective experiments uncovered correlations between overall quality, blockiness, and the motion-compensated edge artifact. We found that visual impairment due to compression is affected by both blocking and MCEA. In fact, for several cases, it becomes the dominant visual impairment.

The frequent occurrence of motion-compensated edge artifacts leads us to the development of a novel measurement framework based on calculating and estimating DCT energies in the current and previous frame blocks. Both the ordinal and the interval scale evaluation showed the accuracy of our metric and the potential of measurement frameworks based on DCT coefficient energies. Furthermore, a trivial linear combination of the MCEAM and GBIM metrics outperformed each of the metrics comprising it.

Future work will explore an overall quality metric that incorporates both components: noisiness and blockiness. In addition, the MCEAM metric could be used during encoding to keep the MCEA artifacts under control. Intra blocks could be selectively sent or the loop filter could be selectively applied (to avoid excessive blurring) when the accumulated MCEA becomes too large.

REFERENCES

[1] A. Leontaris and A. R. Reibman, "Comparison of blocking and blurring metrics for video compression," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Mar. 2005, vol. 2, pp. 585–588.

[2] A. Leontaris, P. C. Cosman, and A. R. Reibman, "Measuring the added high frequency energy in compressed video," in *Proc. IEEE Int. Conf. Image Processing*, Sep. 2005, vol. 2, pp. 498–501.

[3] VQEG [Online]. Available: http://www.vqeg.org/, Final Report From the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment, Mar. 2000

[4] B. Girod, "What's wrong with mean-squared error?," in *Digital Images and Human Vision*. Cambridge, MA: MIT Press, 1993.

[5] J. McCarthy, M. A. Sasse, and D. Miras, "Sharp or smooth? Comparing the effects of quantization vs. frame rate for streamed video," presented at the CHI, Apr. 2004.

[6] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *EURASIP J. Signal Process.*, vol. 70, pp. 247–278, Oct. 1998.

[7] S. A. Karunasekera and N. G. Kingsbury, "A distortion measure for image artifacts based on human visual sensitivity," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, Apr. 1994, vol. 4, pp. 117–120.

[8] M. C. Q. Farias, S. K. Mitra, and J. M. Foley, "Perceptual contributions of blocky, blurry and noisy artifacts to overall annoyance," in *Proc. IEEE ICME*, 2003, vol. 1, pp. 529–532.

[9] C. J. van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatio-temporal model of human video system," in *Proc. SPIE EI*, 1996, vol. 2668, pp. 451–461.

[10] M. A. Masry and S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Signal Process. Image Commun.*, vol. 19, pp. 133–146, Feb. 2004.

[11] Z. Yu, H. R. Wu, S. Winkler, and T. Chen, "Vision-model-based impairment metric to evaluate blocking artifacts in digital video," *Proc. IEEE*, vol. 90, no. 1, pp. 154–169, Jan. 2002.

[12] K. T. Tan and M. Ghanbari, "A multi-metric objective picture-quality measurement model for MPEG video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 7, pp. 1208–1213, Oct. 2000.

[13] A. M. Eskicioglu, "Quality measurement for monochrome compressed images in the past 25 years," in *Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing*, 2000, vol. 4, pp. 1907–1910.

[14] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.

[15] P. Cuenca, L. Orozco-Barbosa, A. Garrido, and F. Quiles, "Study of video quality metrics for MPEG-2 based video communications," in *Proc. Pacific Rim Conf. Communications, Computers and Signal Processing*, 1999, pp. 280–283.

[16] A. Mayache, T. Eude, and H. Cherifi, "A comparison of image quality models and metrics based on human visual sensitivity," in *Proc. IEEE Int. Conf. Image Processing*, 1998, vol. 3, pp. 409–413.

[17] S. Winkler, A. Sharma, and D. McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications," in *Proc. Int. Symp. Wireless Personal Multimedia Communications*, Sep. 2001, pp. 547–552.

[18] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "An objective video quality assessment system based on human perception," in *Proc. SPIE Visual Processing and Digital Display*, 1993, vol. 1913, pp. 15–26.

[19] T. Vlachos, "Detection of blocking artifacts in compressed video," *IEE Electron. Lett.*, vol. 36, no. 13, pp. 1106–1108, Jun. 2000.

[20] Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. IEEE Int. Conf. Image Processing*, 2000, vol. 3, pp. 981–984.

[21] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric for video coding," *IEEE Signal Process. Lett.*, vol. 4, no. 11, pp. 317–320, Nov. 1997.

[22] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[23] R. W. Chan and P. B. Goldsmith, "A psychovisually-based image quality evaluator for JPEG images," in *Proc. IEEE Int. Conf. Systems, Man, Cybernetics*, 2000, pp. 1541–1546.

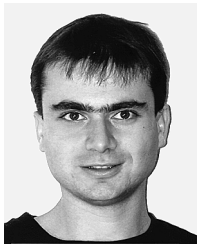
[24] J. Yang, H. Choi, and T. Kim, "Noise estimation for blocking artifacts reduction in DCT coded images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 7, pp. 1116–1134, Oct. 2000.

[25] S. Minami and A. Zakhor, "An optimization approach for removing blocking effects in transform coding," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 5, no. 2, pp. 74–82, Apr. 1995.

[26] W. Gao, C. Mermer, and Y. Kim, "A de-blocking algorithm and a blockiness metric for highly compressed images," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 12, no. 12, pp. 1150–1159, Dec. 2002.

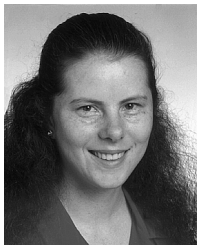
[27] S. Liu and A. C. Bovik, "Efficient DCT-domain blind measurement and reduction of blocking artifacts," *IEEE Trans. Circuits, Syst., Video Technol.*, vol. 12, no. 12, pp. 1139–1149, Dec. 2002.

[28] S. Suthaharan, "Perceptual quality metric for digital video coding," *Electron. Lett.*, vol. 39, no. 5, pp. 431–433, Mar. 2003.
[29] X. Marichal, W.-Y. Ma, and H. J. Zhang, "Blur determination in the compressed domain using DCT information," in *Proc. IEEE Int. Conf. Image Processing*, Oct. 1999, pp. 386–390.
[30] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. IEEE Int. Conf. Image Processing*, 2002, vol. 3, pp. 57–60.
[31] J. Caviedes and S. Gurbuz, "No-reference sharpness metric based on local edge kurtosis," in *Proc. IEEE Int. Conf. Image Processing*, 2002, vol. 3, pp. 53–56.
[32] L. L. Thurstone, "The method of paired comparisons for social values," *J. Abnormal Social Psych.*, vol. 21, pp. 384–400, 1927.
[33] ——, "A law of comparative judgment," *Psych. Rev.*, vol. 34, pp. 273–286, 1927.
[34] T. C. Brown and T. C. Daniel, "Scaling of ratings: Concepts and methods," Research Paper RM-293. USDA Forest Service, Rocky Mountain Forest and Range Experiment Station, 1990, pp. 1–24.
[35] W. S. Torgerson, *Theory and Methods of Scaling*. New York: Wiley, 1958.
[36] J. C. Nunnally, *Psychometric Theory*, 2nd ed. New York: McGraw-Hill, 1978.
[37] D. A. Silverstein and J. E. Farrell, "Quantifying perceptual image quality," in *Proc. IS&T Image Processing, Image Quality, Image Capture, Systems Conf.*, May 1998, pp. 242–246.
[38] A. Leontaris, V. Chellappa, and P. C. Cosman, "Optimal mode selection for a pulsed-quality dual frame video coder," *IEEE Signal Process. Lett.*, vol. 11, no. 12, pp. 952–955, Dec. 2004.
[39] MPlayer 1.0-pre4 Software [Online]. Available: http://www.mplayerhq.hu/
[40] M. G. Kendall, *Rank Correlation Methods*. New York: Hafner, 1955.
[41] G. E. Noether, Why Kendall Tau? [Online]. Available: http://rsscse.org.uk/ts/bts/noether/text.html
[42] S. Arndt, C. Turvey, and N. C. Andreasen, "Correlating and predicting psychiatric symptom ratings: Spearman's r versus Kendall's τ correlation," *J. Psych. Res.*, vol. 33, no. 2, pp. 97–104, Mar. 1999.



Athanasios Leontaris (S'97–M'06) received the Diploma degree in electrical and computer engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2000, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at San Diego (UCSD), La Jolla, in 2002 and 2006, respectively.

He was a summer intern at AT&T Labs—Research, New Jersey and at NTT Network Innovation Labs, Japan in 2004 and 2005, respectively. Currently, he is a Senior Research Engineer at Dolby Laboratories, Inc., Burbank, CA. His research interests include image and video compression, video communication, multimedia processing, and image quality modeling.



Pamela C. Cosman (S'88–M'93–SM'00) received the B.S. degree (with honors) in electrical engineering from the California Institute of Technology, Pasadena, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1989 and 1993, respectively.

She was an NSF postdoctoral fellow at Stanford University and a Visiting Professor at the University of Minnesota during 1993–1995. In 1995, she joined the faculty of the Department of Electrical and Computer Engineering, University of California, San Diego, where she is currently a Professor and Director of the Center for Wireless Communications. Her research interests are in the areas of image and video compression and processing.

Dr. Cosman is the recipient of the ECE Departmental Graduate Teaching Award (1996), a Career Award from the National Science Foundation (1996–1999), and a Powell Faculty Fellowship (1997–1998). She was a Guest Editor for the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS June 2000 Special Issue on Error-Resilient Image and Video Eoding, and was the Technical Program Chair of the 1998 Information Theory Workshop, San Diego. She was an Associate Editor of the IEEE COMMUNICATIONS LETTERS (1998–2001), and an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (2001–2005). She was a Senior Editor (2003–2005), and is now the Editor-in-Chief, of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. She is a member of Tau Beta Pi and Sigma Xi.

Amy R. Reibman (SM'01–F'05) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Duke University, Durham, NC, in 1983, 1984, and 1987, respectively.

From 1988 to 1991, she was an Assistant Professor in the Department of Electrical Engineering, Princeton University, Princeton, NJ. In 1991, she joined AT&T Bell Laboratories, and became a Distinguished Member of Technical Staff in 1995. She is currently a Technical Consultant in the Dependable Distributed Computing and Communication Research Department, AT&T Laboratories, Florham Park, NJ. Her research interests include video compression systems for transport over packet and wireless networks and video quality metrics.

Dr. Reibman was elected an IEEE Fellow in 2005 for her contributions to video transport over networks. In 1998, she won the IEEE Communications Society Leonard G. Abraham Prize Paper Award. She was the Technical Co-Chair of the IEEE International Conference on Image Processing in 2002, the Technical Co-Chair for the First IEEE Workshop on Multimedia Signal Processing in 1997, and the Technical Chair for the Sixth International Workshop on Packet Video in 1994.