Wireless Video Quality Assessment: A Study of Subjective Scores and Objective Algorithms

Anush Krishna Moorthy, Kalpana Seshadrinathan, *Member, IEEE,* Rajiv Soundararajan, *Student Member, IEEE,* and Alan Conrad Bovik, *Fellow, IEEE*

***Abstract*—Evaluating the perceptual quality of video is of tremendous importance in the design and optimization of wireless video processing and transmission systems. In an endeavor to emulate human perception of quality, various objective video quality assessment (VQA) algorithms have been developed. However, the only subjective video quality database that exists on which these algorithms can be tested is dated and does not accurately reflect distortions introduced by present generation encoders and/or wireless channels. In order to evaluate the performance of VQA algorithms for the specific task of H.264 advanced video coding compressed video transmission over wireless networks, we conducted a subjective study involving 160 distorted videos. Various leading full reference VQA algorithms were tested for their correlation with human perception. The data from the paper has been made available to the research community, so that further research on new VQA algorithms and on the general area of VQA may be carried out.**

***Index Terms*—H.264 compression, image processing, quality assessment, subjective quality assessment, video quality, wireless.**

I. Introduction

WITH AN INCREASING demand for entertainment and with the ever-improving technology to fuel this demand, the pervasiveness of digital video in everyday life cannot be debated. From entertainment on the move—hand-held phones spewing out videos—to entertainment at home, digital videos are everywhere. Moreover, wireless systems are rapidly replacing present-day wire-line systems, and new-generation encoders with tremendously improved compression efficiency are being standardized. In such an environment, a digital video passes through numerous processing stages before it finally reaches the end-user. The original video sequence at the transmitter end is passed through an encoder which compresses and restructures the video sequence, which is then passed over a channel. At the receiver end, a decoder decompresses the sequence into a format visible to the end-user. Throughout this process distortions are introduced in the video stream which can produce visually annoying artifacts at the end-user. The encoder, the channel, the decoder, and the display can introduce distortions in the video sequence. Encoder errors may include blocking artifacts, blurring, discrete cosine transform, basis image effect, color bleeding, ringing, and so on [2] due to restrictions on bit-rate and errors in the motion estimation process. The channel, being inherently noisy, can corrupt the video in many ways.

Given that the ultimate receivers of wireless videos are usually human observers, human subjective opinion is the ultimate arbiter of video quality. Thus, evaluation of the perceived quality of degraded video requires selecting a large-enough sample of the human populace and asking each of them to rate the quality of the video on some scale. The value of this score pooled across the human subjects constitutes a score which is representative of the perceived quality of that video. Such an estimation of quality is known as a *subjective assessment* and studies of this type are time-consuming and cumbersome. Alternatively, one may design algorithms that seek to predict the quality of distorted videos in agreement with human subjectivity. Indeed, in the recent past, a variety of effective indices that accurately predict the perceptual quality of images [3] and videos [4]–[8] have been developed. Algorithmic assessment of quality is referred to as *objective quality assessment*.

Evaluation of the effectiveness of objective quality indices for accurately emulating human perception of quality is important, as these algorithms may be used in significant and widely deployed commercial applications. In [9], an extensive subjective quality evaluation of the leading still image quality assessment indices was performed, and their suitability for predicting perceived visual quality was evaluated. Two algorithms—the multiscale structural similarity index (MS-SSIM) [10] and visual information fidelity (VIF) [11] index were demonstrated to correlate significantly higher with human perception than other algorithms. The only publicly available video database containing subjective and objective quality evaluation is the Video Quality Experts Group (VQEG) FRTV Phase I [12], where ten leading video quality assessment (VQA) algorithms were compared and their correlation with human opinion studied. It was found that all the metrics were statistically indistinguishable from peak signal-to-noise ratio (PSNR) [12].

The video database from the VQEG is dated—the report was published in 2000, and was made specifically for TV and hence contains interlaced videos. The presence of interlaced videos complicates the prediction of quality, since the de-

Manuscript received February 4, 2009; revised May 25, 2009 and July 31, 2009. First version published January 29, 2010; current version published April 2, 2010. This paper was recommended by Associate Editor, T. Nguyen.

A. K. Moorthy and A. C. Bovik are with the Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78712-1084 USA (e-mail: anushmoorthy@mail.utexas.edu; bovik@ece.utexas.edu).

K. Seshadrinathan is with Intel Corporation, Chandler, AZ 85226 USA.

R. Soundararajan is with the Department of Electrical and Computer Engineering, University of Texas, Austin, TX 78712 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2010.2041829

1051-8215/$26.00 © 2010 IEEE

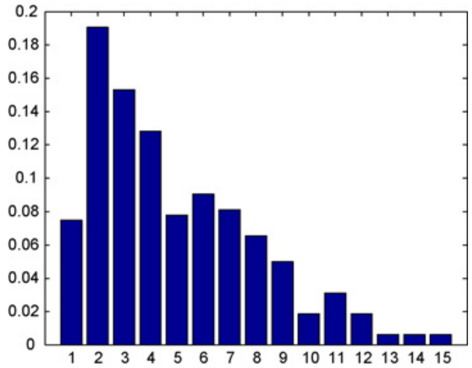


Fig. 1. Histogram (normalized) of differential mean opinion scores from the entire VQEG dataset [12]. Notice how the distribution of scores is highly skewed demonstrating poor perceptual separation.

interlacing algorithm can introduce further distortion before computation of algorithm scores. Further, the VQEG paper included distortions only from old generation encoders such as the H.263 [13] and MPEG-2 [14], which exhibit different error patterns compared with present generation encoders like the H.264 advanced video coding/MPEG-4 Part 10 (referred to as H.264 henceforth) [15]. Finally, the VQEG Phase I database of distorted videos suffers from problems with poor perceptual separation. Both humans and algorithms have difficulty in producing consistent judgments that distinguish many of the videos, lowering the correlations between humans and algorithms and the statistical confidence of the results. For example, in Fig. 1, we plot a histogram of all subjective scores from the VQEG Phase-I dataset. It is clear that the range of quality that the dataset spans is highly skewed.

To address this need, we have conducted a large-scale human and algorithm study using H.264 compressed videos and simulated wireless transmission errors as distortions. An effort has been made to include a wide variety of distortion types having good perceptual separations. For wireless applications H.264 is being widely included in relevant technologies as the Digital Video Broadcasting—Handheld [16], [17] and Mediaflo [17] broadcast standards. As another example, the World Airline Entertainment Association (WAEA) has standardized the H.264 encoder for delivery of wireless video entertainment [18], for on-board video presentations.

This paper is aimed at studying the effectiveness of video quality assessment algorithms in predicting human perception of quality. Our paper is organized as follows. Section II describes the original and distorted videos used, the human study, and the various objective quality assessment algorithms evaluated. Section III describes the performance of the algorithms in terms of their correlation with human opinion, and also evaluates the statistical significance of the obtained results. Finally, we conclude this paper in Section IV.

II. Details of the Subjective Study

A. Source Sequences

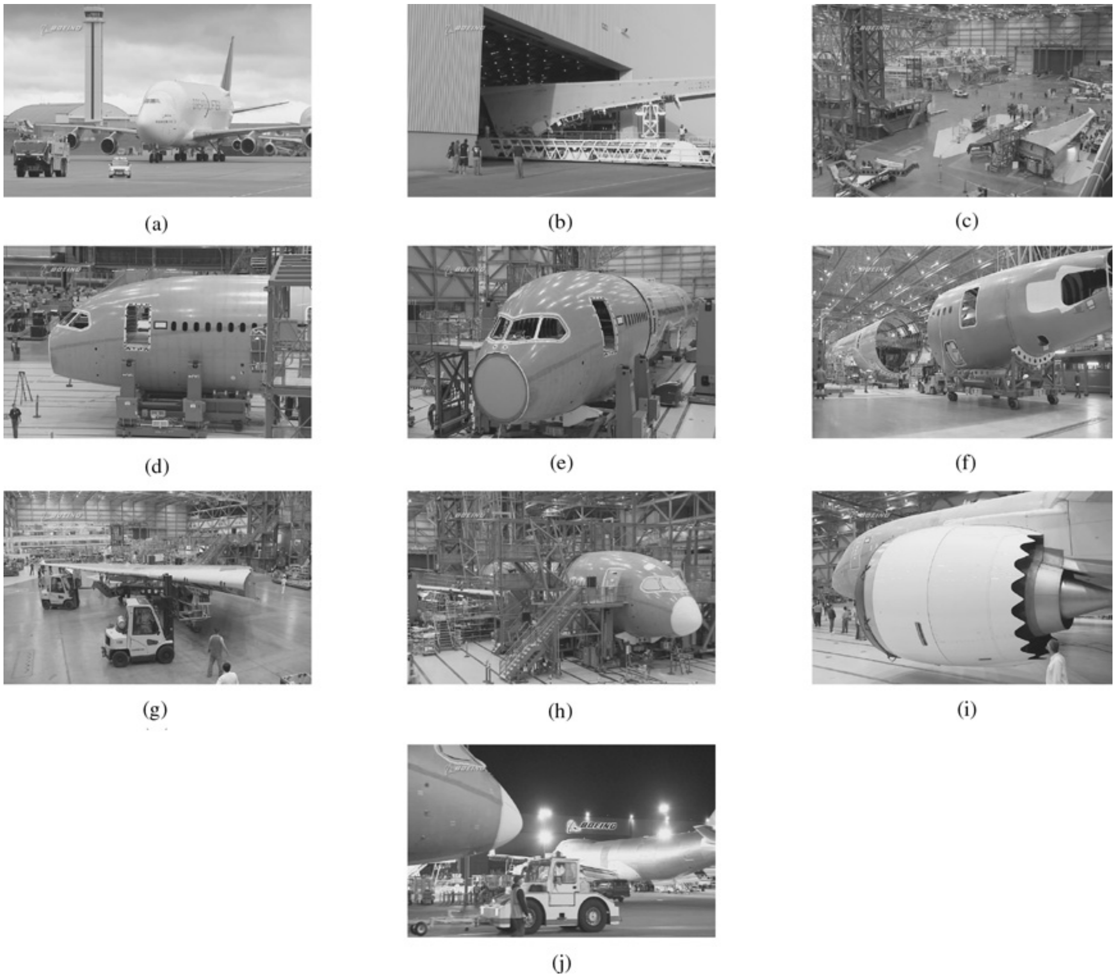
The source videos are in RAW uncompressed progressive scan YUV420 format with a resolution of 768 × 480 and a frame rate of 30 frames per second (f/s). They were provided by Boeing. From a large collection, the chosen videos were those which incorporated a diverse range of interesting motions, objects, and people. Some of the videos are night-sequences. Many of the videos chosen contain scene cuts—in order to include as much of the space of videos as possible. There are ten source sequences, each ten seconds long and hence containing 300 frames. The various videos are as described below:

1) *vid_a:* shows a plane driving up to the camera, with two vehicles flanking it on either side. The flanking vehicles have their blinkers on. Almost zero camera motion;
2) *vid_b:* still camera as object moves toward the left of screen with human motion at the bottom left;
3) *vid_c:* camera pans to the left inside a hangar on a scene with little motion;
4) *vid_d:* camera moves to the right, covering a side of a still plane in a hangar;
5) *vid_e:* camera moves up, covering the front of a still plane in a hangar;
6) *vid_f:* camera still as back-half of a plane moves toward a stationary front half. Scene cuts. Still camera focuses at the point of joining of the two halves;
7) *vid_g:* still camera captures an object moving in a curved path. Scene cuts. Camera moves toward the left covering a scene with little motion;
8) *vid_h:* camera zooms out of still scene with human motion at bottom left, then moves downward;
9) *vid_i:* camera slowly moves, covering the right engine of a plane. Scene cuts. Camera moves upward covering the engine and the plane;
10) *vid_j:* Night sequence. Camera zooms out as object and humans move toward the right of the screen.

Fig. 2 shows frames from the various video sequences.

In this paper, we did not use the raw YUV videos as the pristine videos, but instead converted the videos first into H.264 compressed videos, which are visually lossless (i.e., having a PSNR > 40 dB). Since the user is never likely to see the pristine YUV videos, such a visually lossless H.264 serves as a good reference for assessing the quality of videos degraded over the channel as well as due to compression. Our reasons for using visually lossless, low-compression H.264 videos as the elements of the reference test set are two-fold. First, the overall compressed test set is enormously smaller in size than the original raw video dataset. While this was not an advantage for conducting the human or algorithmic studies, it is a major advantage in allowing others to conduct the studies. As we intend to make all the videos and human scores freely available, the use of visually lossless reference videos is highly convenient for delivering the video set electronically. Second, the visually lossless reference videos have available quality motion vectors which can be used by others (as well as ourselves) to develop VQA algorithms that use motion. By making available quality motion vectors, we make it possible for developers to focus their efforts on other aspects of VQA algorithm development. To date, very few VQA algorithms use motion information.

In order to create perceptually lossless videos, the following parameters for H.264 compression are used:



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j)

Fig. 2. Example frames of the videos used. (a)–(j) correspond to videos a–j mentioned in the description.

1) quantization parameters (Q_p, Q_i) = 18;
2) I-frame period = 14.

Although the I-frame period does not influence the perceived quality, we code at a period of 14 frames in order to reduce the time complexity of the encoding process. We also note that with the quantization parameters set as above, the average PSNR is greater than 45 dB, exceeding the 40 dB level.

B. Test Sequences

We created a set of 160 distorted sequences (4 bit-rates × 4 packet-loss rates = 16 distorted videos *per* reference sequence). The details are as follows.

1) *H.264 Compression:* We used the JM reference software (Version 13.1) [19], [20] made available by the Joint Video Team (JVT) for H.264 encoding. The reference videos were encoded using different bitrates: 500 kb/s, 1 Mb/s, 1.5 Mb/s, 2 Mb/s; with number of slice groups = 3. The bit-rates chosen for encoding follow the WAEA recommendations [18] which recommends a minimum bit-rate of 1 Mb/s for transmission. Additionally, we simulated a bit-rate of 500 kb/s as well. Rate control is as described in [21] and implemented by the JM reference software.

All videos were created using the same value of the I-frame period (96). We also enabled rate-distortion optimization, and used real-time transport protocol as the output file mode. We used the baseline profile for encoding, and hence did not include B-frames. We aimed for wireless transmission of the videos and hence restricted the packet size to between 100 and 300 bytes [22]. A detailed explanation of how packet sizes were computed for the number of slice groups is shown in the appendix. We set the flexible macroblock ordering (FMO) mode as "dispersed" and used three slices per frame.

2) *The Wireless Channel:* We used the software provided by International Telecommunication Union (ITU) [23] documented in [24] to simulate wireless channel errors of packet-loss. The software allows for six different error patterns and hence for six different bit-error rates of 9.3×10^{-3}, 2.9×10^{-3}, 5.1×10^{-4}, 1.7×10^{-4}, 5.0×10^{-4}, and 2.0×10^{-4}. The bit-error patterns used are captured from different real or emulated mobile radio channels. For the packet sizes we simulated, these bit-error rates correspond on an average to packet-loss rates are around 0.4%, 0.5%, 1.7–2%, 2%, 5%, and 17–18%. We assumed that a packet containing an erroneous bit is an erroneous packet [22]. Under this assumption, we base all further discussion on packet-loss rates.

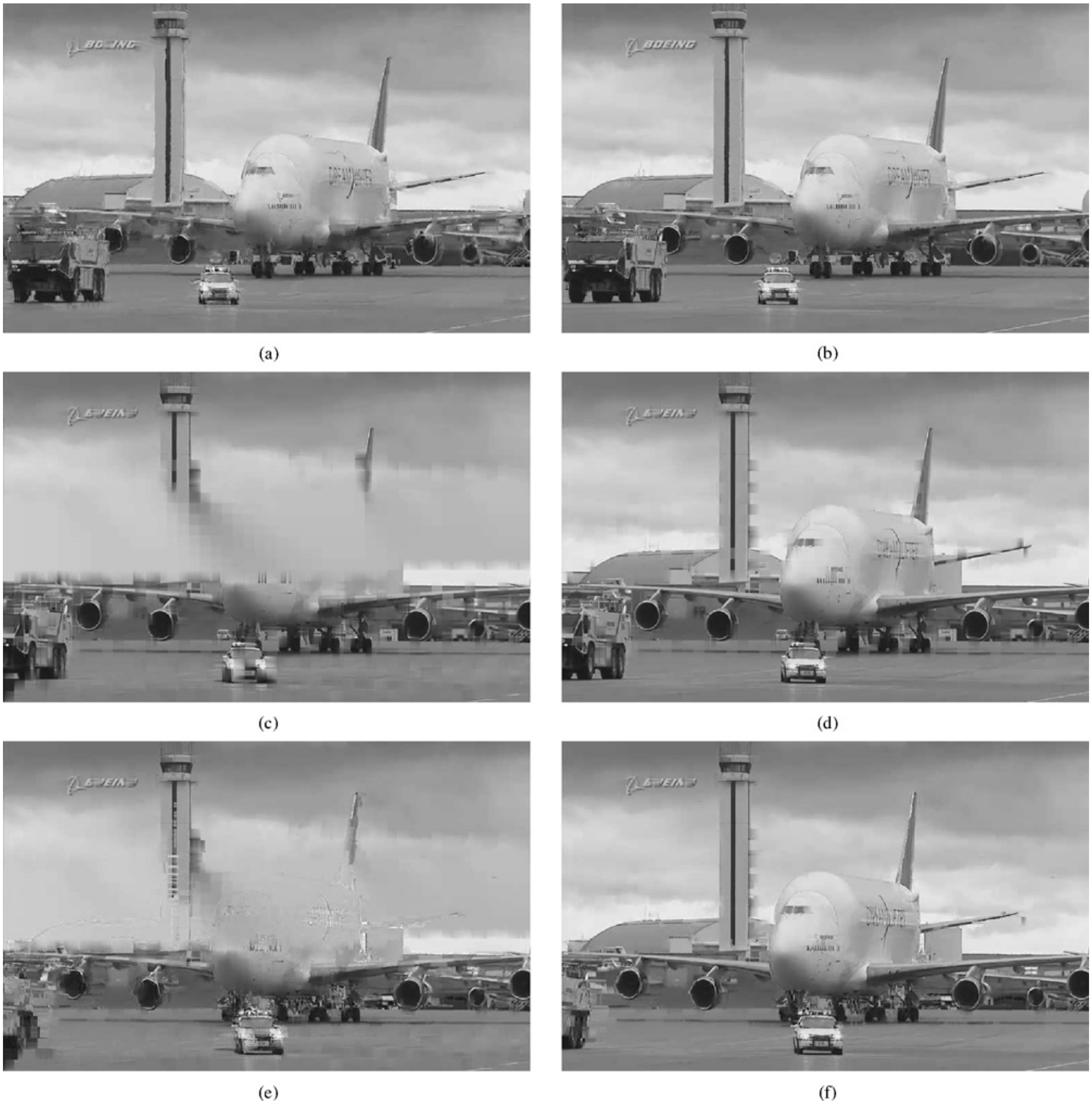

(a) (b) (c) (d) (e) (f)

Fig. 3. Distortions induced in video compressed at 1 Mb/s. (a), (c) and (e) are frames from video passed through a channel with packet loss rate = 17%. (b), (d), and (f) correspond to frames from video passed through a channel with packet-loss rate = 5%. (a) and (b) correspond to frame number 40 (P-frame). (c) and (d) correspond to frame number 97 (I-frame). (e) and (f) correspond to frame 140 (P-frame)—error propagation due to lost packets from I-frame is visible here.

It is clear that the packet loss rates can be divided into four groups instead of the six, since there are two pairs of loss-rates that are quite similar (0.4%, 0.5% and 1.7–2%, 2%). Hence, the distorted videos were created using the simulated channel such that packet loss rates of 0.5%, 2%, 5%, and 17% were achieved.

In order to see how these different parameters affect visual quality, Fig. 3 shows different frames from a video with 5% packet loss rate and 17% packet loss rate (compression rate = 1 Mb/s). Frames corresponding to a random P-frame, an I-frame, and another P-frame following the I-frame are shown in order to visualize the distortions. Loss of information from the I-frame propagates through the P-frames and this is visible in the figure.

At the decoder-end, the JM reference software was used to decode the compressed video stream. The error concealment procedures undertaken by the reference software in accordance with the H.264 standard can be found in [25]. Briefly, at the decoder, all correctly received slices are first decoded and concealment is initiated for "lost" MBs. The processing starts at the edge of the frame and moves inward column-by-column. For lost INTRA frames, a weighted spatial averaging is undertaken. For INTER coded frames, a strategy based on (guessed) motion vectors is utilized for motion compensation and hence for concealment [25].

3) *Comments on Selected Parameters:* There are two major contributions that we wish to make with this paper. One of them is the creation of a publicly available video quality assessment database that can be utilized by researchers as a test-bed for algorithm design and performance evaluation. By evaluating popular algorithms we also provide an objective way of assessing the performance of the video quality assessment algorithms. The other contribution is toward application of VQA algorithms in a practical system. Our goal is to provide the users of these VQA algorithms with an objective comparison of popular algorithms in terms of not only their correlation with human perception, but also the trade-off between performance and computational complexity. It is clear that the parameters for H.264 compression could be modified. Since it was not our goal to assess the ability of the H.264 encoder, we fix certain parameters. The algorithms are evaluated for this set of parameters. It is not unreasonable to believe that the performance of VQA algorithms will not be severely affected by H.264 parameters. Hence, algorithms that perform well on this dataset should ideally perform equally well in a general scenario.

C. Test Methodology

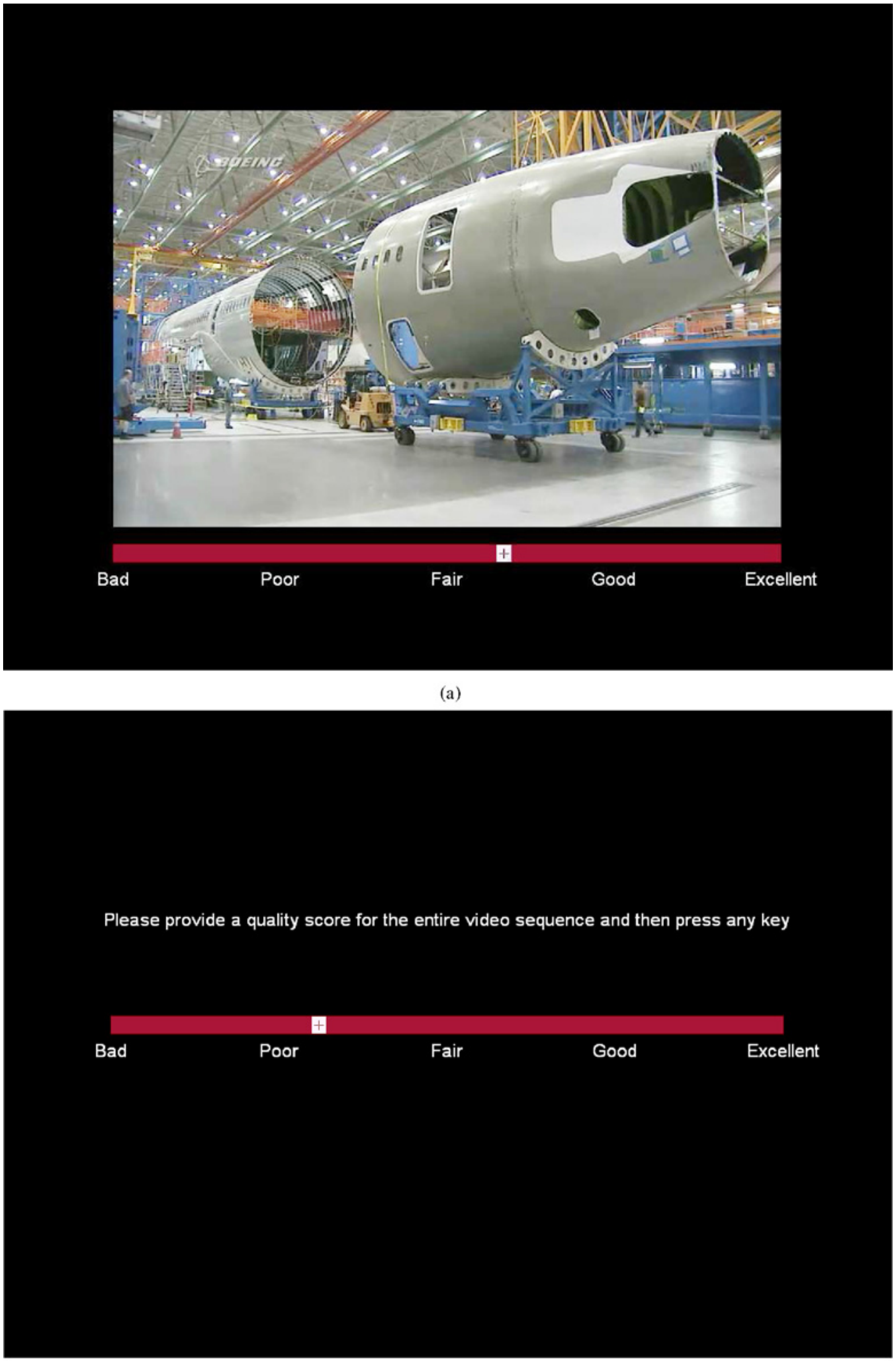
1) *Design:* The study conducted was a single stimulus continuous quality evaluation (or SSCQE) as detailed in [26]. The only difference in this paper was the use of a "hidden-reference." In recent literature (see, e.g., [27]), this model is used in order to "equalize" scores. Specifically, in the set of videos that the subject is shown, the original reference videos are displayed as well. The subject is unaware of its presence or its location in the displayed video set. The score that the subject gives this reference is representative of the supposed bias that the subject carries, and when the scores for the distorted videos are subtracted from this bias, a compensation is achieved, giving us the difference score for that distorted video sequence.

Single stimulus (SS) methods offer advantages relative to double stimulus (DS) methods. In the SS studies the viewer is shown only one video at a time. In DS studies, two videos are shown simultaneously on a split-screen environment (simultaneous double stimulus for continuous quality evaluation), which tends to distract the subject [27], or two sequences are shown one after the other, twice [double stimulus continuous quality-scale method (DSCQS)], thereby increasing the length of the study. The case for SSCQE has been made before in [27]; apart from the arguments provided there, we were also concerned about the time requirements that a DS study would need. For example, the DSCQS described in [26] would require slightly more than twice/four (there are two types of DSCQS, see [26]) times the amount of time as against a single-stimulus approach. This would mean that the number of sessions (see below) would increase by approximately a factor of 2/4, assuming that the 30 min/session limit is not violated (see below). This increased number of sessions could then lead to debates about how to best combine data from different sessions. For example, in [9], a re-alignment study was conducted in order to align scores from different sessions. The memory effects associated with using a SS approach [26] were debated in [27]; since we use videos spanning a duration of 10 s, the memory effects are unlikely to influence the perceived quality [27]. The use of single stimulus thus reduces the time consumed by the study as well as provides a more accurate description of the quality of a sequence. We use a continuous scale for evaluation of quality—i.e., the user is not limited to only discrete scores, but is allowed to provide an score that he feels is appropriate between the lowest and highest ranges on the scale. The use of such a continuous scale, we believe, is superior to the ITU-R absolute category scale that uses a 5-category quality scale adopted by the VQEG studies [12], due to the expanded range of scores that a continuous scale can provide.

2) *Display:* The user interface was developed on a PC running Windows XP, on MATLAB, using the XGL Toolbox for MATLAB which was developed at The University of Texas at Austin [28]. The XGL Toolbox allows precise presentation of psychophysical stimuli to human observers. It is obvious that any errors in displaying the videos, such as latencies, must be avoided when conducting such a study, since these artifacts affect the perceived quality of a video. In order that display issues do not factor into the quality score provided by a subject, all the distorted videos were first loaded into the memory completely before their presentation. The XGL toolbox interfaces with the ATI Radeon X600 graphics card in the PC and utilizes its ability to play out the YUV videos.

A cathode ray tube (CRT) monitor was used to display the videos. Again, the debate between perceiveability of errors on different monitors is discussed in [29] and [30]. The relevance of the findings in [29] is questionable since the monitor sizes for the CRT and liquid crystal display (LCD) were not the same. However, there has been evidence that effects such as motion blur are amplified on an LCD screen [31]. The reproduction of colors on a CRT versus those on an LCD is another point of debate [32]. Although most of the algorithms that we test (see below) do not use color information, we decided to use the CRT for the purposes of this paper. The monitor was calibrated using the Monaco Optix XR Pro device. The same monitor was used for the entire course of the paper. The monitor refresh rate was set at 60 Hz, and each frame of the 30 Hz video was displayed for two monitor refresh cycles. The screen was set at a resolution of 1024 × 768 pixels and the videos were displayed at their native resolution; the remaining areas of the display were black.

3) *Subjects, Training and Testing:* The subjective study was conducted over a course of two weeks at the University of Texas at Austin (UT). The subject pool consisted majorly of under-graduate students from UT. The subjects were a mix of males and females, with a male majority. No monetary compensation for participating in the study was offered. The average subject age was between 22 and 28 years and the subjects were inexperienced with video quality assessments and perception of quality. Though no vision test was performed, a verbal confirmation of soundness of (corrected) vision from the subject was taken to be sufficient.



(a)

Please provide a quality score for the entire video sequence and then press any key

(b)

Fig. 4. Study setup. (a) Video is shown at the center of the screen and a bar at the bottom is provided to rate the videos as a function of time. The pointer on the bar is controlled by using the mouse. (b) At the end of the presentation, a similar bar is shown on the screen so that the subject may rate the entire video. This score is used for further processing.

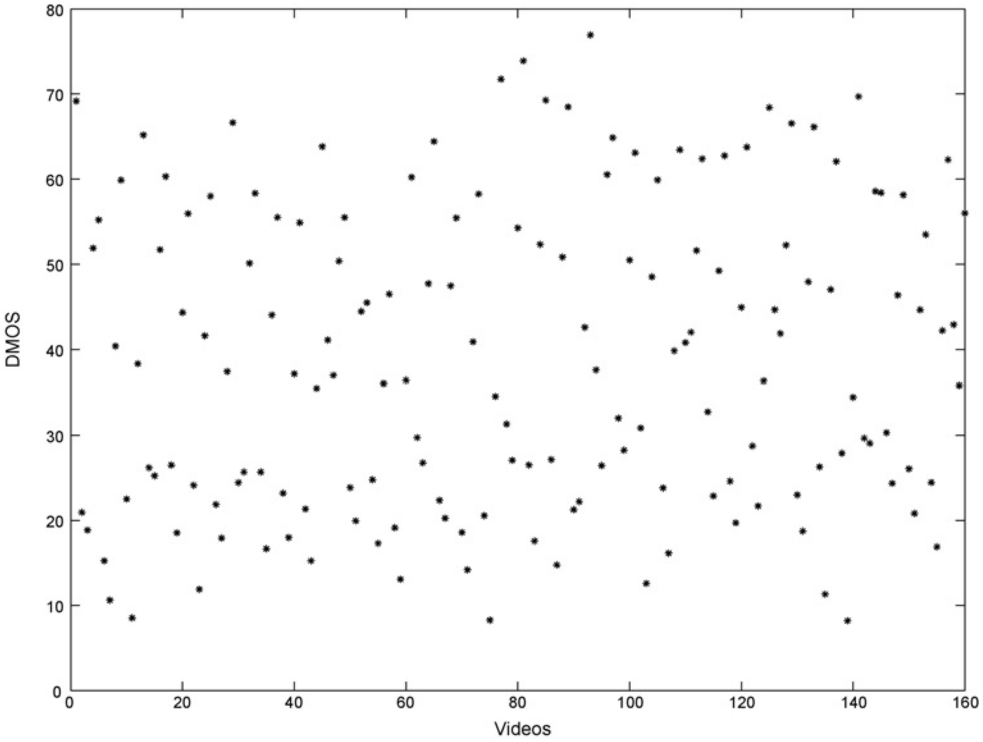


Fig. 5. DMOS scores for all video sequences. Notice how the entire DMOS range seems to be covered. This is indicative of good perceptual separation and a large range of qualities seen in the study.

The study was conducted over two sessions, each lasting less than half an hour, as per recommendations in [26] in order to minimize subject fatigue. An informal after-study feedback conducted indicated that the length of the study was appropriate, and that the subjects did not experience any uneasiness or fatigue during the course of the study. The subjects were briefed as per the recommendations in [26]. Each session consisted of 90 videos each (80 distorted + 10 reference), with a short training set of sequences shown before the actual session. The training videos shown were different from the ones used for the actual study and were selected to span the range of quality that the subject was bound to see in the study. The training sessions consisted of six and three training sequences, respectively.

The study consisted of the set of sequences shown in random order. The order was randomized for each subject as well as for each session. Care was taken to ensure that two consecutive sequences did not belong to the same reference, to minimize memory effects [26]. As mentioned above, each session consisted of the reference sequence (also in random order) without the subjects' knowledge of its presence.

The sequences were shown at the center of the CRT monitor with a bar at the bottom of the screen, calibrated—"Bad," "Poor," "Fair," "Good," and "Excellent," equally spaced across the scale. Although the scale was continuous, the calibrations served to guide the subject. A screen indicating that the video was ready to be played was shown, and the video was played when the user pressed any key on the keyboard. The rating bar was controlled using a mouse. The subjects were asked to rate the videos continuously, i.e., as a function of time; at the end of the sequence a similar bar was shown at the center of the screen, where the subject was asked to rate the quality of the video sequence. Once the score was entered, the subject was not allowed to go back and change the score. The quality rating was converted into a score between 0 and 100. A sample screen shot of the setup is seen in Fig. 4.

Although scores of videos as a function of time were collected, the following analysis is based on the cumulative scores of the video sequence, as is the norm. The collected continuous data will be used in the future to better understand the decision-making process of the human. This will require accounting for the latency of human response to changes in visual quality.

D. Processing of the Scores

A total of 31 subjects participated in the study. The score that each subject assigned to a distorted sequence in a session was subtracted from the score that the subject assigned to the reference sequence in that session, thus forming a difference score. A subject rejection algorithm was run as per the recommendations of the ITU [26], which rejected one subject. We used the double stimulus continuous quality evaluation (DSCQE) subject rejection technique, since the SSCQE with a hidden reference corresponds in principle to the DSCQE technique. A detailed explanation of the method is included in the appendix for completeness.

The scores from the remaining subjects were then averaged to form a differential mean opinion score (DMOS) for each sequence. The DMOS score is representative of the perceived

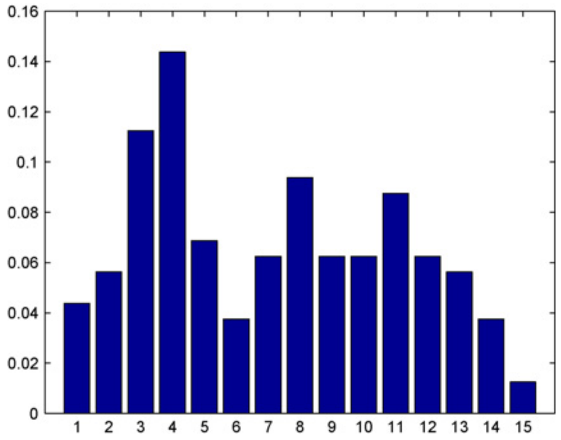


Fig. 6. Histogram (normalized) of differential mean opinion scores from our wireless video quality study. Notice how the distribution of scores is uniform compared to that from the VQEG—Fig. 1.

quality of the video. Specifically, let s_{ijk} denote the score assigned by subject i to video j in session k and let N_{ik} be the number of test videos seen by subject i in session k. The difference scores d_{ijk} are computed as

$$d_{ijk} = s_{ijk} - s_{ij_{ref}k}.$$

The DMOS (after subject rejection) is then

$$\mathrm{DMOS}_j = \sum_i \sum_k d_{ijk}.$$

Although the VQEG FRTV phase-I study [12] used the DMOS scores for further processing, an alternative is the use of Z-scores [33]. The Z-score for a sequence *per session* is calculated as

$$\mu_{ik} = \frac{1}{N_{ik}} \sum_{j=1}^{N_{ik}} d_{ijk}$$

$$\sigma_{ik} = \sqrt{\frac{1}{N_{ik}-1} \sum_{j=1}^{N_{ik}} (d_{ijk} - \mu_{ik})^2}$$

$$z_{ijk} = \frac{d_{ik} - \mu ik}{\sigma_{ik}}.$$

We found in subsequent analysis that the performance of video quality assessment algorithms did not change much, regardless of which human measure (DMOS, Z-scores) were used as descriptors of perceived video quality. Hence, in all further analyses, we use DMOS scores. The DMOS scores are plotted for each of the distorted video sequences in Fig. 5, which demonstrates that the videos shown span the entire range of visual quality and that they exhibit good perceptual separation. Further, in Fig. 6 we plot a histogram of scores from our database. Notice how the scores are uniformly distributed as compared to those from the VQEG studies (Fig. 1).

E. Video Quality Assessment Algorithms

1) *PSNR:* The peak signal-to-noise ratio, used even today for image/video quality assessment, is a measure of the mean-square-error between the two signals being compared. For video-sequences, the PSNR is calculated for each frame then averaged across frames (Y component only).

2) *Frame-SS-SSIM:* The single-scale structural similarity index (SS-SSIM) [34], designed for still images, is based on the principle that image "structure" is perceptually important. It is defined as a product of a structure term, an intensity term, and a contrast term. The SS-SSIM index value was calculated on each frame, then averaged across all frames (Y component only). The software implementation used is available at [35].

3) *Frame-MS-SSIM:* The multiscale SSIM index [10] corrects the viewing-distance dependence of SS-SSIM and accounts for the multiscale nature of both natural images and human visual system. The MS-SSIM index performs better (relative to human opinion) than the SS-SSIM index on images. Here, the MS-SSIM was calculated on every frame and then averaged across all frames (Y component only). The software implementation used was obtained from the authors.

4) *VQM:* Video quality metric [7], proposed by Pinson and Wolf, was the top performer in the VQEG phase II video quality study, and is an American National Standards Institute and International Organization for Standardization standard. The VQM index was designed for videos and the inputs are raw YUV—original and distorted. The software implementation used is available at [36].

5) *VSNR:* The visual signal-to-noise ratio [37] is a wavelet domain image quality metric proposed by researchers at Cornell. Since this is designed as an image quality assessment metric, the VSNR was applied on each frame, then averaged across all frames (Y component only). The software implementation used is available at [38].

6) *Speed-Weighted SSIM:* Since the regular frame-SS-SSIM does not incorporate any temporal weighting, a recent algorithm [5] which accounts for motion was also evaluated. A weighting scheme is assigned to the frame-SS-SSIM values on each frame and then the scores are averaged across all frames (Y component only). This temporally-weighted frame-SS-SSIM is referred to as "speed-weighted SSIM" (SW-SSIM) henceforth. The software implementation used was obtained from the authors.

7) *P-SS-SSIM:* Human beings tend to perceive poor regions in an image/video with greater severity than an objective algorithm that pools scores from each region with equal weight (simple mean across the scores) [39]. The P-SS-SSIM index changes the pooling strategy by weighting the order statistics of the scores, and correlates better with the human perception of quality than the SS-SSIM. This was applied on the Y-component only and on a frame-by-frame basis, since this was developed for images.

8) *Video VIF:* The video VIF [8] is an information-theoretic approach to video quality assessment—an extension of the VIF for images [11]. This algorithm evaluates the quality as a ratio of mutual-informations between quantities in the wavelet domain. VIF uses natural-scene statistics to model the image and the distortions, and performs well on

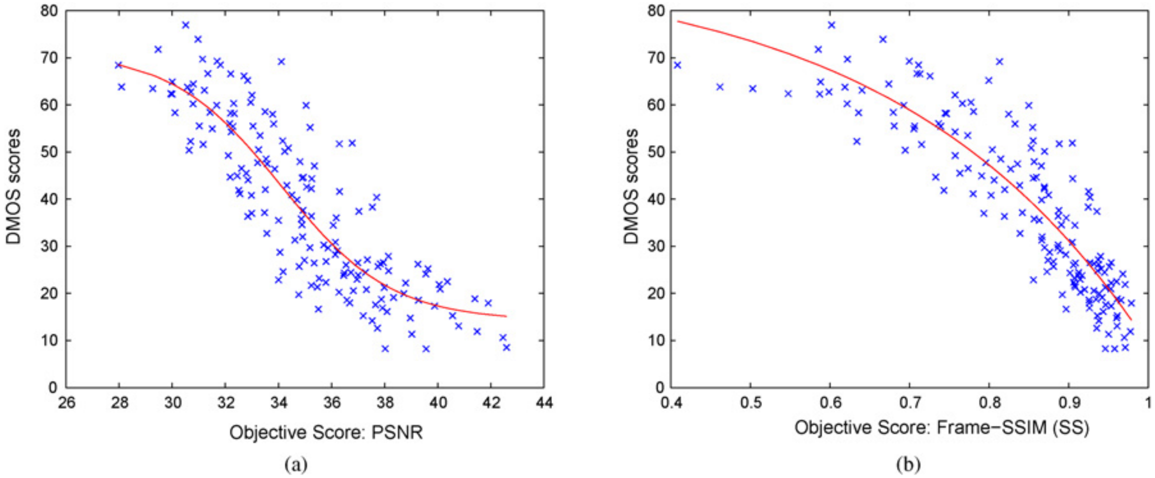


Fig. 7. Scatter plots. (a) PSNR. (b) Frame-SS-SSIM.

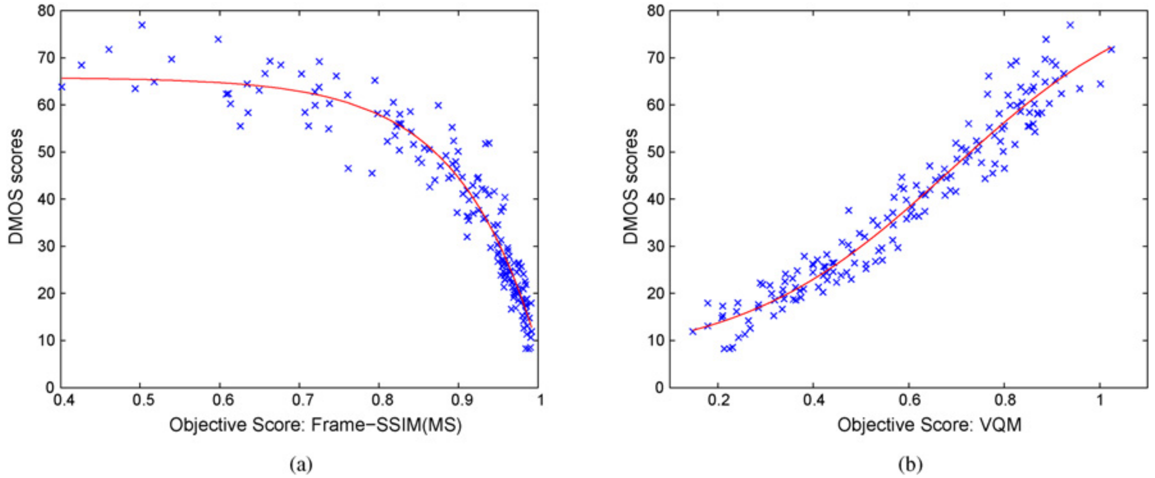


Fig. 8. Scatter plots. (a) Frame-MS-SSIM. (b) VQM.

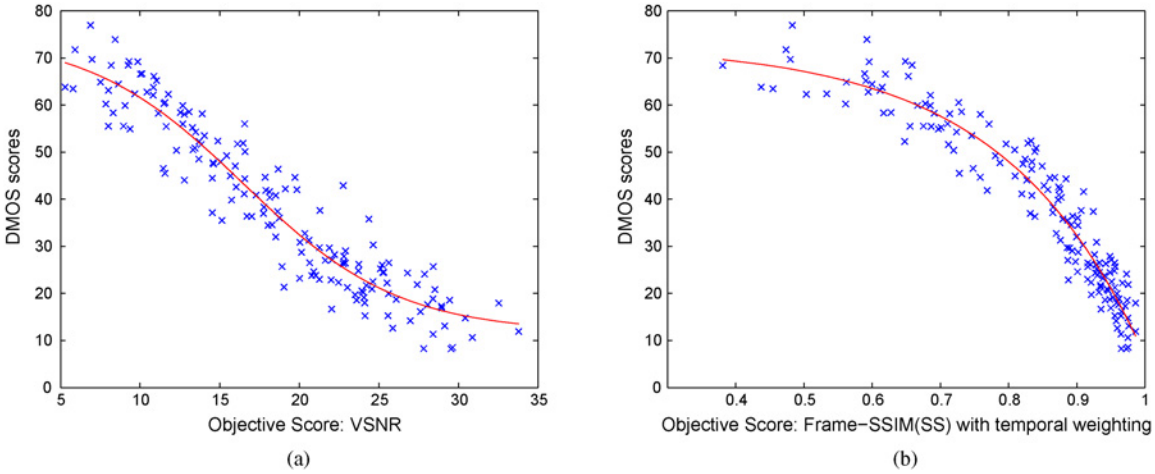


Fig. 9. Scatter plots. (a) VSNR. (b) SW-SSIM.

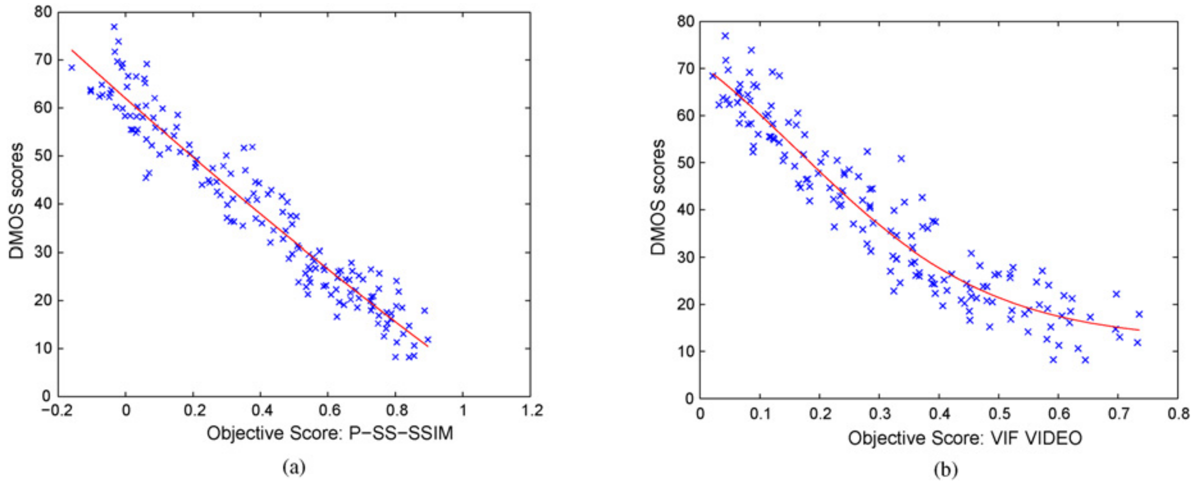


Fig. 10. Scatter plots. (a) P-SS-SSIM. (b) Video VIF.

TABLE I

SROCC: Objective Quality Assessment Algorithms: Distortion

VQA Algorithm	Mild Loss	Average Loss	High Loss	Severe Loss
PSNR	0.6987	0.6501	0.4764	0.4465
Frame-SS-SSIM	0.7270	0.7133	0.6236	0.4604
Frame-MS-SSIM	0.8574	0.8176	0.6989	0.6268
VQM	0.8914	0.7906	0.8598	0.5726
VSNR	0.7174	0.4977	0.5987	0.5629
SW-SSIM	0.8310	0.7287	0.8471	0.7486
P-SS-SSIM	0.8550	0.7388	0.7606	0.5182
Video VIF	0.7959	0.6385	0.7501	0.5775

Mild loss = 0.5%, average loss = 2%, high loss = 5%, severe loss = 17% packet-loss rate.

TABLE II

SROCC: Objective Quality Assessment Algorithms: Compression

VQA Algorithm	0.5 Mb/s	1 Mb/s	1.5 Mb/s	2 Mb/s
PSNR	0.8546	0.8248	0.8570	0.8400
Frame-SS-SSIM	0.8619	0.8681	0.8959	0.8752
Frame-MS-SSIM	0.9567	0.9460	0.9724	0.9480
VQM	0.9561	0.9602	0.9666	0.9565
VSNR	0.9480	0.9148	0.9477	0.9340
SW-SSIM	0.9533	0.9477	0.9477	0.9426
P-SS-SSIM	0.9610	0.9556	0.9655	0.9418
Video VIF	0.9236	0.9094	0.9084	0.9598

the VQEG database. The software implementation used was obtained from the authors.

III. Results

The scatter plots for various algorithms along with the "best-fit" regressed curve (see below) are seen in Figs. 7–10.

A. *Performance Metrics*

The objective metrics were evaluated based on:

1) prediction accuracy;
2) prediction monotonicity.

TABLE III

Performance of Various Objective Quality Assessment Algorithms Over All Distortion Types

VQA Algorithm	SROCC	CC	RMSE
PSNR	0.8615	0.8639	8.8997
Frame-SS-SSIM	0.8967	0.8875	8.1448
Frame-MS-SSIM	0.9608	0.9588	5.0196
VQM	0.9721	0.9711	4.2172
VSNR	0.9418	0.9484	5.6028
SW-SSIM	0.9599	0.9617	4.8450
P-SS-SSIM	0.9628	0.9637	4.7180
Video VIF	0.9470	0.9524	5.3854

The results for each of the algorithms is seen in Table III. The metrics used for evaluation are—Spearman rank ordered correlation coefficient (SROCC), the linear correlation coefficient (CC)—after non-linear regression and the root mean square error (RMSE)—after non-linear regression as prescribed in [12]. We used a 4-parameter logistic function [12], constrained to be monotonic to transform the objective score

$$Quality(x) = \frac{\beta_1 - \beta_2}{1 + exp\left(-\frac{x-\beta_3}{|\beta_4|}\right)} + \beta_2.$$

This logistic function was recommended by the VQEG [12] and has been widely used in evaluating the performance of algorithms that were tested on the VQEG Phase I dataset [12]. Further, a similar logistic was used for evaluating the performance of image quality assessment algorithms [9]. After such a transformation, we calculated the Linear (Pearson's) correlation coefficient, and the root-mean-squared error, between the transformed score and the DMOS scores.

The CC and the RMSE measure the *prediction accuracy*, the SROCC measures the *prediction monotonicity*. We report the value of SROCC, CC, RMSE for all data in Table III. We also report the SROCC across two sets, where the videos are grouped in terms of the compression rate (Table II) and packet-loss rate (Table I).

TABLE IV
STATISTICAL SIGNIFICANCE ANALYSIS

	PSNR	Frame-SS-SSIM	Frame-MS-SSIM	VQM	VSNR	SW-SSIM	P-SS-SSIM	Video VIF
PSNR	–	–	0	0	0	0	0	0
Frame-SS-SSIM	–	–	0	0	0	0	0	0
Frame-MS-SSIM	1	1	–	0	–	–	–	1
VQM	1	1	1	–	1	1	–	1
VSNR	1	1	–	0	–	0	0	–
SW-SSIM	1	1	–	0	1	–	–	1
P-SS-SSIM	1	1	–	–	1	1	1	1
Video VIF	1	1	0	0	–	0	0	–

A "1" indicates that the metric in that row is statistically better than the metric in the column; a "0" indicates that it is statistically worse and a "–" indicates that the scores are statistically indistinguishable.

B. Statistical Significance and Hypothesis Testing

Similar to the approach in [9], we perform a statistical significance analysis based on an assumption of Gaussianity of the residuals between the VQA algorithm scores (after non-linear regression) and the DMOS for each video sequence. We used the Kolmogorov–Smirnov test to evaluate Gaussianity [40] on the normalized scores. In our analysis, we found that we could not reject the null hypothesis (the scores have a standard normal distribution) at the 5% level for any metric and hence our assumption of Gaussianity is valid for all metrics. We used the F-statistic [41] for comparing the variance of sets of samples. The test was performed for the dataset taken as a whole.

The null hypothesis is that the residuals from one VQA algorithm come from the same distribution and are statistically indistinguishable with 95% confidence from the resiudals from another VQA algorithm. The alternative hypothesis is that the sample variance of one VQA algorithm is greater than the other. Table IV shows results from the statistical significance analysis. A "1" indicates that the metric in that row is statistically better than the metric in the column; a "0" indicates that it is statistically worse and a "–" indicates that we could not reject the null hypothesis at the 5% level, and hence the scores are statistically indistinguishable.

The correlations exhibited by the various algorithms are higher than those seen in the VQEG studies [12], and this can be attributed to the uniformity of the content as well as the uniformity of the distortion. However, some observations with regard to the metrics can be made under the assumption that the inter-metric performance will remain identical for a non-uniform dataset. SW-SSIM, VQM, and MS-SSIM seem to perform the best across distortion types, while VQM, MS-SSIM, and P-SSIM seem to do well across compression rates. Overall, VQM, MS-SSIM, and P-SSIM perform the best amongst the algorithms. The statistical analysis leads to the conclusion that at 95% confidence level, except frame-SS-SSIM, all other algorithms are statistically better than PSNR.

C. Complexity Versus Performance

Even though there exists a host of VQA algorithms, PSNR is still used as an indicator of quality. This stems from the fact that the computation of PSNR is easy to implement and real-time estimates for PSNR may be made available. In order to assist researchers interested in deploying these algorithms practically, we perform a rough complexity analysis of the proposed algorithms. Fig. 11 shows a plot of the trade-off between complexity and performance. Complexity increases along the horizontal axis and performance increases along the vertical axis. Algorithms which are statistically indistinguishable are grouped together. Note that the figure should act as a guide in choosing algorithms for applications rather than as an absolute measure of performance versus complexity.

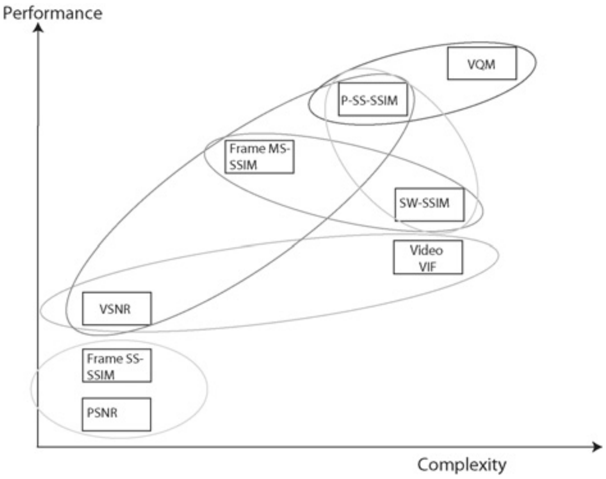


Fig. 11. Performance versus complexity tradeoff. Algorithms which are statistically indistinguishable in performance are circled together. Figure not to scale.

IV. CONCLUSION

A subjective study to assess the perceived quality of H.264 compressed video sequences transmitted over a wireless channel was performed. Based on the results from the study, various leading objective quality assessment algorithms were evaluated using popular metrics, to gauge their correlation with human perception. The ten reference sequences as well as the 160 distorted sequences have been made available to the research community in order to further research on perceptual video quality assessment.

APPENDIX A
SELECTION OF H.264 ENCODER PARAMETERS

In this section, we explain how the packet sizes are fixed for a given number of slice groups. Consider a video to be encoded at 1 Mb/s. We fix the number of slice groups at $SG = 3$. The video dimensions are 768×480 and the frame rate is 30 Hz. We encode packets such that each packet contains one slice. In order to fix the packet size between 100 and 300 bytes, we need that the number of slices per frame SF be

$$\frac{1/8 \times 10^6}{300 \times 30} \leq SF \leq \frac{\frac{1}{8} \times 10^6}{100 \times 30}.$$

Hence

$$13.8889 \leq SF \leq 41.6667.$$

We select $SF = 18$. With $SF = 18$, the packet size is $\frac{\frac{1}{8}\times 10^6}{18\times 30} =$ 231.48 bytes. In order for $SF = 18$ to make sense, we need that the number of macroblocks per slice MBS be an integer. We can verify this as: $MBS = \frac{(768\times 480)}{18\times 16\times 16} = 80$, where each macroblock is an element of size 16×16. Hence, we encode this video at a rate of 1 Mb/s with three slice groups and 18 slices/frame with 80 macroblocks per slice to achieve a packet-length of 231.48 bytes.

Appendix B
Subject Rejection Procedure

The subject rejection procedure we follow is the one prescribed by the ITU for the DSCQE [26]. For each presentation, we compute the mean μ_{jkr}, standard deviation σ_{jkr}, and kurtosis β_{jkr} where kurtosis is the ratio of the fourth moment to the square of the second moment. For each observer i we find the parameters P_i, Q_i as follows:

```
for $j, k, r = 1, 1, 1$ to $J, K, R$
if $2 \leq \beta_{jkr}, \leq 4$ then
    if $u_{ijkr} \geq \mu_{jkr} + 2\sigma_{jkr}$, then $P_i = P_i + 1$
    if $u_{ijkr} \leq \mu_{jkr} - 2\sigma_{jkr}$, then $Q_i = Q_i + 1$
else
    if $u_{ijkr} \geq \mu_{jkr} + \sqrt{20}\sigma_{jkr}$, then $P_i = P_i + 1$
    if $u_{ijkr} \leq \mu_{jkr} - \sqrt{20}\sigma_{jkr}$, then $Q_i = Q_i + 1$.
```

We then compute the ratios $\psi = \frac{P_i+Q_i}{J\times K\times R}$ and $\gamma = \left|\frac{P_i-Q_i}{P_i+Q_i}\right|$. If $\psi > 0.05$ and $\gamma < 0.3$ then reject subject i.

In the above equations, J = number of test conditions including the reference, K = number of test videos, R = number of repetitions. u_{ijkr} is the score assigned by subject i in condition j for the video k and for repetition r.

References

[1] A. K. Moorthy and A. C. Bovik. (2009, Sep. 27). *Live Wireless Video Quality Assessment Database* [Online]. Available: http://live.ece.utexas.edu/research/quality/live_wireless_video.html

[2] M. Yuen and H. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Process.*, vol. 70, no. 3, pp. 247–278, 1998.

[3] A. C. Bovik and Z. Wang, *Modern Image Quality Assessment*. New York: Morgan and Claypool, 2006.

[4] K. Seshadrinathan and A. C. Bovik, "A structural similarity metric for video based on motion models," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Apr. 2007, pp. 869–872.

[5] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Am.*, vol. 24, no. 12, pp. B61–B69, Dec. 2007.

[6] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Process. Image Commun.*, vol. 19, no. 2, pp. 121–132, Feb. 2004.

[7] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcasting*, vol. 50, no. 3, pp. 312–313, Sep. 2004.

[8] H. R. Sheikh and A. C. Bovik, "A visual information fidelity approach to video quality assessment," in *Proc. 1st Int. Workshop Video Process. Quality Metrics Consumer Electron.*, Jan. 2005, pp. 23–25.

[9] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[10] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. IEEE Asilomar Conf. Signals Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[11] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[12] Video Quality Experts Group (VQEG), *Final Report from the Video Quality Experts Group on the Validation of Objective Quality Metrics for Video Quality Assessment*, 2000 [Online]. Available: http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI

[13] K. Rijkse, "H. 263: Video coding for low-bit-rate communication," *IEEE Commun. Mag.*, vol. 34, no. 12, pp. 42–45, Dec. 1996.

[14] *Generic Coding of Moving Pictures and Associated Audio Information Part 2: Video*, ITU-T and ISO/IEC JTC 1, ITU-T Rec. H.262 and ISO/IEC 13818-2 MPEG-2, 1994.

[15] *Advanced Video Coding*, ITU-T Rec. H.264 and ISO/IEC 14496-10, 2003.

[16] *302 304 V1. 1.1, Digital Video Broadcasting (DVB): Transmission System for Handheld Terminals (DVB-H)*, ETSI Standard, Dec. 2004.

[17] B. Furht and S. Ahson, *Handbook of Mobile Broadcasting: DVB-H, DMB, ISDB-T, and Mediaflo*. Auerbach Publications, 2008.

[18] *Digital Content Delivery Methodology for Airline In-Flight Entertainment Systems*. World Airline Entertainment Association.

[19] *H.264/AVC Software Coordination* [Online]. Available: http://iphome.hhi.de/suehring/tml/

[20] (2007). *H.264/mpeg-4 AVC Reference Software Manual* [Online]. Available: http://iphome.hhi.de/suehring/tml/JM(JVT-X072).pdf

[21] *Joint Model Reference Encoding Methods and Decoding Concealment Methods; Section 2.6: Rate Control*, document JVT-I049.doc, JVT, 2003.

[22] T. Stockhammer, M. Hannuksela, and T. Wiegand, "H.264/AVC in wireless environments," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 657–673, Jul. 2003.

[23] *Common Test Conditions for RTP/IP Over 3GPP/3GPP2* [Online]. Available: http://ftp3.itu.ch/av-arch/videosite/0109San/VCEG-N80 software.zip

[24] *Common Test Conditions for RTP/IP Over 3GPP/3GPP2*, document VCEG-M77.doc, ITU-T SG16, 2001.

[25] Y.-K. Wang, M. M. Hannuksela, V. Varsa, A. Hourunranta, and M. Gabbouj, "The error concealment feature in the H.26L test model," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2002, pp. 729–732.

[26] *BT-500-11: Methodology for the Subjective Assessment of the Quality of Television Pictures*. International Telecommunication Union.

[27] M. H. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," *SPIE Proc.*, vol. 5150, no. 3, pp. 573–582, 2003.

[28] J. Perry, *The XGL Toolbox*, 2008 [Online]. Available: http://128.83.207.86/jsp/software/xgltoolbox-1.0.5.zip

[29] M. Pinson and S. Wolf, "The impact of monitor resolution and type on subjective video quality testing," Nat. Telecommun. Inform. Administration (NTIA), Washington D.C., NTIA Tech. Mem. TM-04-412, 2004.

[30] S. Tourancheau, P. L. Callet, and D. Barba, "Impact of the resolution on the difference of perceptual video quality between CRT and LCD," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 3. 2007, pp. 441–444.

[31] H. Pan, X. F. Feng, and S. Daly, "LCD motion blur modeling and analysis," in *Proc. IEEE Int. Conf. Image Process.*, 2005, pp. 21–24.

[32] G. Sharma, "LCDs versus CRTs: Color-calibration and Gamut considerations," *Proc. IEEE*, vol. 90, no. 4, pp. 605–622, Apr. 2002.

[33] A. M. van Dijk, J. B. Martens, and A. B. Watson, "Quality assessment of coded images using numerical category scaling," *Proc. SPIE*, vol. 2451, pp. 90–101, 1995.

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

[35] *The Structural Similarity Index* [Online]. Available: http://live.ece.utexas.edu/research/Quality/index.htm

[36] *Video Quality Metric* [Online]. Available: http://www.its.bldrdoc.gov/n3/video/VQM software.php

[37] D. M. Chandler and S. S. Hemami, "VSNR: A wavelet-based visual signal-to-noise ratio for natural images," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2284–2298, Sep. 2007.

[38] *Visual Signal to Noise Ratio* [Online]. Available: http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr_matlab_source.zip

[39] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.

[40] A. Stuart and J. K. Ord, *The Advanced Theory of Statistics*. New York: Wiley, 1977.

[41] D. C. Montgomery and G. C. Runger, *Applied Statistics and Probability for Engineers*. Wiley-Interscience, 1999.

Anush Krishna Moorthy received the B.E. degree in electronics and telecommunication with a Silver Medal from the University of Pune, Pune, India, in June 2007, and received the M.S. degree in electrical engineering from the University of Texas, Austin, in 2009.

He joined the Laboratory for Image and Video Engineering (LIVE), University of Texas, Austin, in 2007. He is currently the Assistant Director with LIVE, Department of Electrical and Computer Engineering, University of Texas. His research interests include image and video quality assessment, image and video compression, and computational vision.

Kalpana Seshadrinathan (S'03–M'09) received the B.Tech. degree from the University of Kerala, Thiruvananthapuram, Kerala, India, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from the University of Texas, Austin, in 2004 and 2008, respectively.

She was an Assistant Director with the Laboratory for Image and Video Engineering (LIVE), University of Texas, from 2005 to 2008. She is currently a System Engineer with Intel Corporation, Chandler, AZ. Her current research interests include image and video quality assessment, computational aspects of human vision, motion estimation and its applications, and statistical modeling of images and video.

Dr. Seshadrinathan is a recipient of the 2003 Texas Telecommunications Engineering Consortium Graduate Fellowship, and the 2007 Graduate Student Professional Development Award from the University of Texas.

Rajiv Soundararajan (S'08) received the B.E. (Hons) degree in electrical and electronics engineering from the Birla Institute of Technology and Science, Pilani, Rajasthan, India, in 2006, and the M.S. degree in electrical engineering from the University of Texas, Austin, in 2008, where he is currently pursuing the Ph.D. degree.

His current research interests include statistical signal processing and information theory with applications to image and video compression, and quality assessment.

Alan Conrad Bovik (F'96) was born in Kirkwood, MO, on June 25, 1958. He received the B.S. degree in computer engineering in 1980, and the M.S. and Ph.D. degrees in electrical and computer engineering in 1982 and 1984, respectively, all from the University of Illinois, Urbana-Champaign.

He is currently the Curry/Cullen Trust Endowed Chair Professor with the University of Texas, Austin, where is he also the Director of the Laboratory for Image and Video Engineering (LIVE), Department of Electrical and Computer Engineering. He has published over 500 technical articles in these areas and holds two U.S. patents. He is the author of *The Handbook of Image and Video Processing* (Academic Press, 2005), *Modern Image Quality Assessment* (Morgan and Claypool, 2006), *The Essential Guide to Image Processing* (Academic Press, 2009), and *The Essential Guide to Video Processing* (Academic Press, 2009). His current research interests include image and video processing, computational vision, digital microscopy, and modeling of biological visual perception.

Dr. Bovik has received a number of major awards from the IEEE Signal Processing Society, including the Education Award in 2008, the Technical Achievement Award in 2005, the Distinguished Lecturer Award in 2000, and the Meritorious Service Award in 1998. He is also a recipient of the Hocott Award for Distinguished Engineering Research at the University of Texas. He received the Distinguished Alumni Award from the University of Illinois, Champaign-Urbana in 2008, the IEEE Third Millennium Medal in 2000, and two journal paper awards from the International Pattern Recognition Society, in 1988 and 1993. He is a Fellow of the Optical Society of America and a Fellow of the Society of Photo-Optical and Instrumentation Engineers. He has been involved in numerous professional society activities, including the Board of Governors of the IEEE Signal Processing Society from 1996 to 1998, being the Editor-in-Chief of the IEEE TRANSACTIONS ON IMAGE PROCESSING from 1996 to 2002, the Editorial Board on the PROCEEDINGS OF THE IEEE from 1998 to 2004, being the Series Editor for *Image, Video, and Multimedia Processing* (Morgan and Claypool, 2003-present), and becoming the Founding General Chairman of the 1st IEEE International Conference on Image Processing, held in Austin, TX, in November, 1994. He is a registered Professional Engineer in the State of Texas and is a frequent Consultant to legal, industrial, and academic institutions.